IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 20, 2020, accepted September 29, 2020, date of publication October 16, 2020,
date of current version October 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3031763

A Speech Emotion Recognition Model Based on Multi-Level Local Binary and Local Ternary Patterns

YEŞİM ÜLGEN SÖNMEZ[1] **AND ASAF VAROL**[2]
[1]Department of Software Engineering, Faculty of Technology, Fırat University, 23119 Elâzığ, Turkey
[2]Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Maltepe University, 34857 Istanbul, Turkey
Corresponding authors: Yeşim Ülgen Sönmez (phdyus@gmail.com) and Asaf Varol (asafvarol@maltepe.edu.tr)

ABSTRACT Interpreting a speech signal is quite challenging because it consists of different frequencies and features that vary according to emotions. Although different algorithms are being developed in the speech emotion recognition (SER) domain, the success rates vary according to the spoken languages, emotions, and databases. In this study, a new lightweight effective SER method has been developed that has low computational complexity. This method, called 1BTPDN, is applied on RAVDESS, EMO-DB, SAVEE, and EMOVO databases. First, low-pass filter coefficients are obtained by applying a one-dimensional discrete wavelet transform on the raw audio data. The features are extracted by applying textural analysis methods, a one-dimensional local binary pattern, and a one-dimensional local ternary pattern to each filter. Using neighborhood component analysis, the most dominant 1024 features are selected from 7680 features while the other features are discarded. These 1024 features are selected as the input of the classifier which is a third-degree polynomial kernel-based support vector machine. The success rates of the 1BTPDN reached 95.16%, 89.16%, 76.67%, and 74.31% in the RAVDESS, EMO-DB, SAVEE, and EMOVO databases, respectively. The recognition rates are higher compared to many textural, acoustic, and deep learning state-of-the-art SER methods.

INDEX TERMS Discrete wavelet transform, local binary pattern, local ternary pattern, neighborhood component analysis, speech emotion recognition.

I. INTRODUCTION

Speech processing methods are used in the domain of human-computer interaction (HCI) such as security applications, computer education applications, vehicle card systems, automatic translation systems, call center applications, psychosis monitoring and diagnosis of neuropsychological disorders, voice message sorting, telecommunication, assistive technologies, and audio mining [1]. It is also used in digital forensics, games, robots, and the legal evaluation of an individual's psychological integrity [2]. Although speech processing has been researched in different research areas, it is mainly studied in digital signal processing [1], [3].

The human voice has biometric features; thus, there is a big difference among utterances of the same sentence by different speakers. The voice may have information on human's thoughts, emotions, age, gender, physiology, and health, while also containing sounds from the recording environment. There are also differences between the speech signals of the same speaker recorded at different times [1]. Therefore, one of the challenging research areas in speech processing is the detection and understanding of emotion, called speech emotion recognition (SER) [1]. Emotion is a cognitive mechanism of the human brain, that contains a whole set of opinions, feelings, and behavior of the speaker [4], [5].

SER algorithms have advanced with novelties in feature extraction, selection, and classification stages. In these stages, discriminative and robust features that correctly contain information about the speaker's emotional state are one of the major challenges in the SER domain [6]. The other main difficulties to be addressed are listed as follows: the differences in the generation of the databases; the number of emotions in the databases; the variability of speech features with the speaker's language; the length of speech; noise in the speech signal; variability of features with age and/or gender, and the

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

shape of the vocal tract. The main aim in this field is the development of a system overcoming these difficulties.

To contribute to the literature in this area, in this research, a novel SER method called 1BTPDN is proposed, after an extensive literature review in the feature extraction portion of the SER domain. In the 1BTPDN method, a one-dimensional local binary pattern (1D-LBP) and one-dimensional local ternary pattern (1D-LTP) are applied to the raw speech data. After that, a one-dimensional discrete wavelet transform (1D-DWT) is applied. Then, the most discriminative features are selected by using Neighborhood Component Analysis (NCA) weights. Finally, the third-degree polynomial kernel (cubic) based SVM classifier is used. The essential contributions of the proposed 1BTPDN method are listed below:

1. This work particularly enhanced the feature extraction stage of the SER process. We obtained local and textural (histogram-based) features using 1D-LBP and 1D-LTP. The aim is to achieve global optimum features from local features. This framework has encouraging accuracy results which are comparable to many deep learning and traditional machine learning SER methods.

2. This novel framework applied to raw audio data, eliminating the workload of preprocessing such as windowing and framing. Noises in the speech signals are removed by applying 1D-DWT. Feature extraction and noise reduction are performed together by using 1D-DWT with nine levels. 1D-DWT forms a sequential structure.

3. The proposed 1BTPDN method is a new lightweight handcrafted SER method because the 1D-LBP and 1D-LTP have low computational complexity. The presented method consists of three main stages. These stages are feature extraction, feature selection, and classification. Also, it is a feed-forward method. Therefore, there is no need to set a large number of parameters like deep learning. In the feature extraction, binary pattern, ternary pattern, and DWT are used together. By using these methods, a multileveled feature generation method is presented. The time complexity of this method is calculated as $O(n\log n)$. In feature selection, NCA is used. It is a distance-based feature selector. $O(nd)$ (d is the dimension) is NCA complexity. In the classification stage, SVM is utilized as a classifier. The complexity of the SVM is $O(n\text{^}3)$.

4. The best results were reached by using SVM Cubic. In order to see the effect of the features, the traditional machine learning algorithm is used instead of deep learning. This indicates that classical classifiers can be very successful when robust and distinctive features are selected. 1BTPDN was tested over four different benchmark databases such as RAVDESS, EMO-DB, SAVEE, and EMOVO. The accuracy results for RAVDESS-Song, EMO-DB, SAVEE, and EMOVO databases are 95.16%, 89.16%, 76.67%, and 74.31%, respectively.

The rest of the study is divided into nine sections as follows: a literature review of SER feature engineering methods, the feature selection techniques, and the classification models are explained in Section II, Section III, and Section IV, respectively. The background of our proposed method (LBP, LTP, and DWT) is described in Section V. The framework and algorithms are defined in Section VI. The experimental setup, results, and discussion of the novel SER method are shared in Section VII, Section VIII, and Section IX, respectively. The paper is completed with a Conclusion and Future Work directions in Section X.

II. ACOUSTIC FEATURES IN SER LITERATURE

In SER methods, feature extraction is the stage of emphasizing the most accurate, dominant, and distinctive features that imitate the characteristics of the speech signal [6]. The feature optimization stage is executed by reducing and selecting the extracted features with minimal information loss [6].

The features of the speech signal are categorized such as local (segmental or short-term) speech features that are obtained from the speech frames and global (supra-segmental or long-term) speech features which are computed from the statistics of complete utterance [7]. These features and related studies are listed as follows:

- Time-Based Features

They are zero-crossing rate (ZCR) [8] and amplitude-based features, such as amplitude descriptor, log attack time, attach, delay, sustain, release envelop, short-time energy (STE) [9], shimmer [10], rhythm-based features [8], [11], volume, and temporal centroid [3].

- Frequency-Based Features

A speech is an acoustic analog signal that constantly varies on time. In Fourier analysis, the time and period of the frequency components are created [12], [13]. Autoregressive model analysis or Fourier Transform (FT) function transforms the time-domain speech signal into a frequency-domain speech signal [3].The frequency domain features are linear predictive coding (LPC) coefficients [11], [14] linear spectral frequency, peak frequency, short time FT –STFT-based, chroma related, envelope modulation spectrum-based, and tonality-based features [3].

Tonality-based features are fundamental frequency [11], pitch histogram, harmonicity, jitter [10], and harmonic-to-noise ratio (HNR) [3]. Spectrum-shape-based features are entropy, bandwidth, the crest factor, centroid, skewness, and kurtosis of the spectrum [3], [11]. Spectral features are time dependent features, i.e., the energy of each frequency component changes depending on time [15].

- Cepstrum Based Features

A cepstrum is the inverse FT of the spectrum logarithm, so the cepstrum of the cepstral features are cepstral-based spectral features [3]. Power cepstrum is more convenient than the amplitude, power, and phase cepstrum types for speech processing [3]. These features obtained with different algorithms are called cepstrum coefficients and abbreviated as CC. They are Mel-frequency CC (MFCC) [16], [17] gammatone CC (GTCC) [18], linear prediction CC (LPCC), perceptual linear prediction (PLP) CC, relative-spectral PLP features, Greenwood function CC (GFCC), log-frequency power cepstrum

(LFPC), and linear frequency CC (LFCC) [16], [18]. The cepstral coefficients are achieved by calculating the logarithm of energy [16], [18].

- Wavelet Transform Based Features

Wavelet Transform (WT) sub-divides the time-domain entire speech signal into signals with both high frequencies and low frequencies, in the time-frequency domain [3]. The WT is a spectral analysis technique that produces coefficients (detail and approach) to extract the frequency features of the local domain and time domain [19].

- Deep Features

Deep learning is a powerful technique for obtaining high-level features called high-level descriptors (HLDs) from low-level information [3]. Local features are spectral low-level descriptors (LLDs) (Mel Filter Bank, MFCCs), energy and voice LLDs (loudness, F0, jitter, shimmer); global features involve functions extracted from the LLDs, such as maximum, minimum, mean, standard deviation, duration, regression coefficients [20]. The deep learning methods are deep neural networks (DNN), deep stacked auto-encoder (SAE), convolutional neural network (CNN) [21], [22], long-short term memory network (LSTM) [21], recurrent neural network (RNN) and other similar methods [23].

- Texture Based Features

Methods that extract textural features are co-occurrence matrices, Gabor filters (GF), histogram of gradients (HOG), scale invariant feature transform, local binary pattern (LBP), and ternary pattern (LTP) [3]. These methods measure the statistical information of visual signs.

As a result of the literature review, the most used acoustic properties in SER are prosodic features, spectral features, vocal qualities, and nonlinear Teager energy operator (TEO), over the past decade. Prosodic or supra-segmental features are statistical measurements related to pitch, energy, duration, timing, and articulation [15]. They reflect the intensity and intonation of the sound in different emotions. They are paralinguistic features that are extracted from large units of speech such as syllables, words, phrases, and sentences. They are also known, therefore, as long-term features [6]. Spectral features are divided into spectral and cepstral features; MFCC, LPCC, LFPC, GFCC, and formants are used for the SER. The characteristics of the vocal tract are well symbolized in the frequency domain. Spectral features that well simulate the vocal tract [7], [15]. The quality of the voice depends on the physical structure of the vocal tract and the glottal waveforms [8]. It includes HNR, jitter, shimmer, normalized amplitude quotient, quasi open quotient, parabolic spectral parameter, and maxima dispersion quotient [7], [15]. Teager energy operator (TEO) detects the energy of speech for stressed emotions such as anger and sadness [24]. Stressful conditions cause nonlinear airflow during speech production. TEO based features are the energy variation of the gap between lips and glottis [24], [25].

III. FEATURE SELECTION IN THE SER LITERATURE

Feature optimization is performed to obtain the most descriptive and distinctive features on the feature set after the feature extraction stage. Normalization, dimension reduction, or feature selection are used to increase the speed and accuracy of the classifier, as well as to reduce its workload and run time [26], [27]. Normalization is used to eliminate unit differences between the feature values [28]. Sequential forward selection, principal components analysis (PCA), and fast correlation-based filter are outstanding feature selection methods in the SER [27]–[29].

Linear dimensionality reduction is a linear projection of high-dimension to low-dimension [30]. Locality preserving projection, locality sensitive discriminant analysis, neighborhood preserving embedding, principal components analysis (PCA), linear discriminant analysis (LDA), factor analysis, independent components analysis, relevant components analysis, and neighborhood components analysis (NCA) are linear dimension reduction approaches [27]–[29].

PCA is an unsupervised statistical feature selection method using orthogonal transformation. This approach optimizes variance loss and protects the global structure [27], [28]. LDA is a supervised feature selection method that aims to find vectors in the space where classes are best decomposed; it fails to discover the local geometric structure, but it is successful for global geometric structures [27]–[29]. Fisher selection is a statistical method that uses the standard deviation values for each feature [28], [29]. NCA is a supervised algorithm used for classification, dimension reduction, and mainly, distance metric learning. Using the nearest neighbor technique, NCA tries to find a space in which the neighborhoods of points on the same label are tighter than the points on different labels. NCA's components are not orthogonal, and NCA optimizes system accuracy [31].

IV. CLASSIFICATION IN THE SER LITERATURE

Selecting the correct feature from emotional speech data and the appropriate classifier improves the SER performance [15]. The machine learning (ML) algorithm have input data categorized that defines their classes to predict the class of new input data. ML involves two processes, training and testing. The conventional classifiers are naive Bayes [6], decision trees (DT) [6], artificial neural networks (ANN) [8], hidden Markov model (HMM) [16], Gaussian mixture model (GMM) [24], extreme Learning machine (ELM) [25], k-nearest neighbor (KNN) [25], incomplete sparse least square regression [32], and support vector machines (SVM) [33]. K-means algorithm is an unsupervised clustering method; other classifiers may be used with in ensemble methods [34]. The sample size of emotional speech should be high in GMM and HMM algorithms in the training process [19]. GMM is an effective algorithm for global features extracted from training data. The ANN algorithm is limited in increasing robustness and accuracy but it is useful for nonlinear features (Kernel function) [19]. In smaller samples of training data, ANN gives better results than GMM

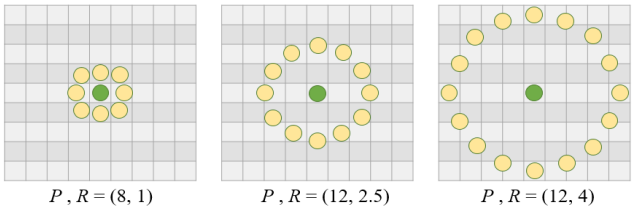

FIGURE 1. Different circular neighborhood sets of LBP function on images.

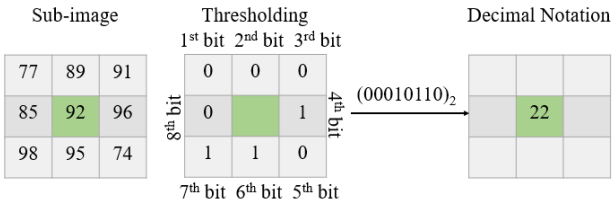

FIGURE 2. Comparing the center pixel to neighbor pixels and generating the LBP binary value.

and HMM [8], [32]. SVM is a statistical and supervised ML algorithm; it decomposes linear and nonlinear features of the input data; the features are transformed to a high dimensional vector space by a kernel function [33]. KNN is a distance-based, supervised ML algorithm [32]. When these algorithms are examined, SVM performs the best, followed by GMM, HMM, KNN, and ELM with decreasing success rates [32]. Multi-layer perceptron is also a supervised neural network learning model used in the SER [35]. Deep-learning classifiers widely used in the SER domain as classification techniques are RNN [23], CNN [21], [22], DNN [36], LSTM network [8], auto encoders, multitask learning, transfer learning, and attention mechanism [6], [15].

V. BACKGROUND

The accuracy of the SER studies depends on the extracted features, feature set size, the classifier used, and the speech databases. It has been decided to take advantage of the textural features that are not used much in the SER applications. The textural features have been obtained from raw audio data by applying LBP and LTP functions. LBP [37] is used in audio scene classification, depression analysis, and emotion detection [3]. LBP and LTP are used in two dimensional (2D) images for texture segmentation and feature detection in image processing. They have discriminative power. They also have computational and programming simplicity, which make them utilizable for realtime applications [38].

A. ONE DIMENSIONAL LOCAL BINARY PATTERN

The basic LBP function takes into account the eight neighborhoods of a pixel, whereas the description includes all circular neighborhoods [38]. The notation (P, R) in Fig.1, represents the pixel neighborhoods P, sampling points on a circle of radius R [38], [39].

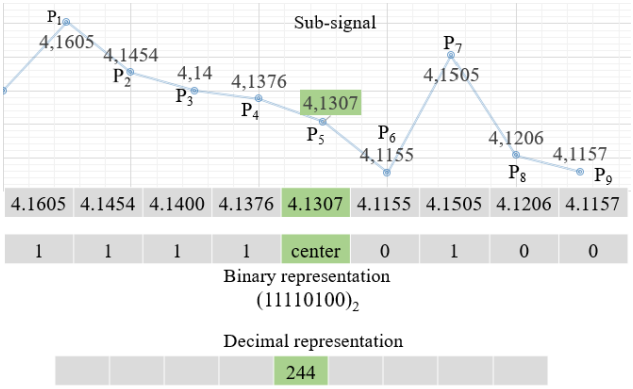

FIGURE 3. The generation of a binary-base number by comparing the central value with the neighboring values and calculation of decimal base number by 1D-LBP function.

P-bit binary number shown in Fig. 2, is generated by comparing the number of the central pixel with the numbers of its neighbor pixels, in a 2D image with 3 × 3 dimensions and this binary number is computed by (1) [37]–[39].

$$LBP_{P,R} = \sum_{i=0}^{P-1} S(g_i - g_c(R))x2^i \tag{1}$$

where g_c is the central pixel number, g_i is the number of its ith neighbor; P is the number of circular neighbors; R *is* the radius which is the distance (Euclidean, Hamming); and $S(x)$ is a conditional function, presented in (2) [38].

$$S(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{2}$$

The LBP method was developed by making a one-dimensional LBP (1D-LBP) for feature extraction from raw audio signals. LBP-based speech processing is functional for signal segmentation and voice activity detection. 1D-LBP function has low computational complexity as shown in Fig. 3. [38], [40].

The steps of 1D-LBP function are listed as follows: loading signal, dividing signal into nine overlapping blocks, and setting the 5th block of the nine blocks as the central number of this nine blocks; extracting the binary features of the block using (3); then converting the bits to a decimal number using (4) [38]–[41].

$$b_i = \begin{cases} 1, & p_i - p_c \geq 0 \\ 0, & p_i - p_c < 0, \end{cases} \quad i \in N, N = \{0, 1, 2, \ldots 7\}, 5 \notin N \tag{3}$$

where b_i is ith bit number, p_i is the ith signal value of the nonoverlapping block, p_c is the center, and N is the index number as shown in Fig. 3. [38]–[41].

$$LBP\ value = \sum_{i=0}^{7} b_i x 2^{7-i} \tag{4}$$

where the value is the decimal value of the LBP function which describes the 1D-LBP function that returns 256 features [38]–[41].

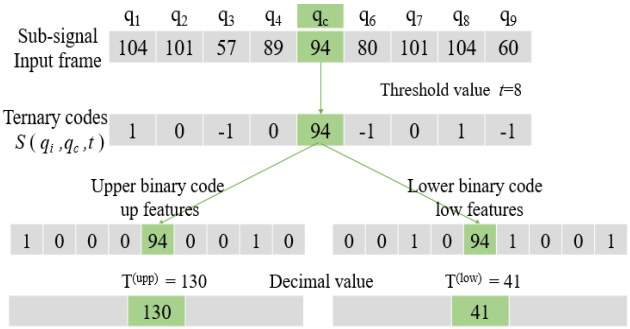

FIGURE 4. The basic steps of 1D-LTP function. Generating ternary codes from the sub-signal then generating upper and lower codes. Finally calculating decimal values.

B. ONE DIMENSIONAL LOCAL TERNARY PATTERN

LTP is a statistical approach for textural analysis in image processing [42]. It encodes the density differences between central and neighboring pixels as -1, 0, and 1 [43]. One dimensional LTP (1D-LTP) has been improved from LTP for speech signal frames; it compares each value with its neighbors according to a threshold value. To compute the ternary pattern, a three-valued function S is given by (5) [42]–[44].

$$S\left(q_i, q_c, t\right) = \begin{cases} +1, & q_c - t > q_i \\ 0, & q_c - t \leq q_i \; and \; q_c + t \geq q_i \\ -1 & q_c + t < q_i \end{cases} \tag{5}$$

where $S\left(q_i, q_c, t\right)$ is ith value of the signal frame q_i is the ith block value of the nine blocks, q_c is the center –the 5th block, and t is the threshold value [42], [43]. The magnitude differences between $(q_c \pm t)$ and q_i are calculated. $S\left(q_i, q_c, t\right)$ is divided into upper features in the nine S_{upp} blocks and lower features in the nine S_{low} blocks by (6) and (7) [42]–[44]. In Fig. 4, the threshold t is 8, but threshold changes depending on the purpose of the application [43].

In S_{upp}, $+1$ values take the value one, while all the other values $(-1, 0)$ take the value zero. Similarly, in S_{low}, -1 values take the value one, while all other values $(+1, 0)$ take the value zero.

$$S_{upp}\left(q_i, q_c, t\right) = \begin{cases} 1, & S\left(q_i, q_c, t\right) = +1 \\ 0, & otherwise \end{cases} \tag{6}$$

$$S_{low}\left(q_i, q_c, t\right) = \begin{cases} 1, & S\left(q_i, q_c, t\right) = -1 \\ 0, & otherwise \end{cases} \tag{7}$$

1D-LTP extracts 512 features from a signal, 256 features from the upper blocks and 256 features from the lower blocks. These features are computed and encoded through their decimal values T^{upp} and T^{low} by (8) and (9) [42]–[44].

$$T^{upp} = \sum_{i=0}^{7} S_{upp}\left(q_i, q_c, t\right) x 2^{7-i} \tag{8}$$

$$T^{low} = \sum_{i=0}^{7} S_{low}\left(q_i, q_c, t\right) x 2^{7-i} \tag{9}$$

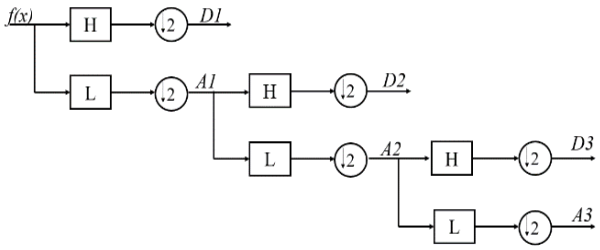

FIGURE 5. Multi levels of a signal applied DWT. The process of multiresolution of the $f(x)$ signal.

C. ONE DIMENSIONAL DISCRETE WAVELET TRANSFORM

In wavelet analysis, the signal is generated from only one function called the main or the parent wavelet, when it decomposes, wavelet coefficients are formed [45]. The main property of the wavelets is time–frequency localization, i.e., the WT has a changing window size, it produces wide time windows at low frequencies and narrow time windows at high frequencies [46]. WT runs on multiscale, but the classic FT runs on only one scale, such as time scale or frequency scale [45].

Discrete wavelet transform (DWT) is a spectral analysis technique used for analyzing non-stationary signals, and provides time-frequency representation of the signals. DWT decomposes a signal into a set of sub-bands through consecutive high-pass and low-pass filtering of the time domain signal. The down sampled signals through first filters are called first level approximation, A1 and detail coefficients, D1 as shown in Fig. 5. Then, approximation and detail coefficients of next level are obtained by using the approximation coefficient of the previous level. The number of decomposition levels is determined depending on the dominant frequency components of the signals [45], [46].

The DWT uses two functions; the wavelet function, set on the high-pass filter and the scaling function, set on the low-pass filter [46]. DWT uses parameters such as coiflet, symmlet, haar, and daubechies according to the input signal.

VI. METHODOLOGY

We propose a novel SER method called speech emotion recognition model based on multi-level local binary pattern and local ternary pattern, which has been abbreviated as 1BTPDN. The methodology of the 1BTPDN can be explained step by step as follows.

1) First, audio data have been properly labeled in four different databases, then labeled raw audio data are uploaded to Matlab.

2) 768 features are obtained by applying the 1D-LBP function and applying the 1D-LTP function to the raw audio data. In the implementation of the 1D-LTP, the threshold value is taken as half of the standard deviation of the signal.

The 1D-LBP function of the 1BTPDN method is shown in Algorithm 1. The 1D-LTP function of the 1BTPDN method is shown in Algorithm 2.

3) 1D-DWT (symlets 4 filter) with nine levels is applied to the raw data and low-pass filter coefficients, such as L1, L2,

Algorithm 1 1*D-LBP* Function

Input: Raw audio signal x.
Output: 1*D-LBP* 256-feature set.

1: **function** *1D-LBP* (x, 5)
2: Assign the h as the histogram of the sub-signals.
3: **for** $i = 1$ **to** length (x) – 8 **do**
4: $block = x\ (i : i+ 8)$
5: $k = 0$; $u\ (i) = 0$; $l\ (i) = 0$
6: **for** $j = 1$ **to** 9 **do**
7: **if** $j \neq 5$
8: **if** $block\ (\ j\ ) - block\ (5) > 0$
9: $l\ (i) = l\ (i) + 2\ \hat{}k$
10: **end if**
11: $k = k+ 1$
12: **end if**
13: **end for** ***j***
14: $h\ (l\ (i) + 1) = h\ (l\ (i) + 1) + 1$
15: **end for*i***

L3……L9 are obtained. The main purpose of it to generate low, high and moderate level features.

4) 768 features are extracted from each filter using 1D-LBP and 1D-LTP. The 768 features are extracted from a total of ten levels with nine levels of wavelets and one main function. These features are concatenated to generate a 7680-feature set.

5) Min–max normalization [44], [47] is applied before feature reduction and selection.

6) Weights of the 7680 features are calculated using NCA. The feature with the highest weight is the most discriminative feature and the feature with the lowest weight is a redundant feature. The most discriminative 1024 features are selected by using the NCA weights. Other features are eliminated. NCA is a distance-based and weight-based feature selection method which guarantees good results. It calculates features using distance parameters like KNN [31], [48]. NCA generates nonnegative weights for all features [44], it uses a gradient-based optimizer for the best weights. NCA is a nonconvex optimization, such as a k-means algorithm, so it is difficult to train [31]. It can't select features automatically. That is, different solutions are obtained each time the NCA runs, and it is recommended to run it more than once to get the best solution [31], [44].

7) The selected 1024 most discriminative features are utilized as the input of the classifier. The most successful results are obtained in 1024 features.

The feature extraction-selection algorithm of the 1BTPDN is shown in Algorithm 3.

8) The third-degree polynomial kernel-(cubic) based SVM is used as the classifier. The coding is one-to-all and the kernel scale mode is automatically set. The SVM classifier that has variable kernel functions, is one of the widely used optimization-based classifiers. The kernel functions increase

Algorithm 2 1*D-LTP* Function

Input: Raw audio signal x.
Output: 1*D-LTP* 512-feature set.

1: **function** *1D-LTP* (x, 5)
2: Assign *threshold* as half of the standard deviation of the signal.
3: Assign the *h1* and *h*2 as the histograms of the sub-signals.
4: **for** $i = 1$ **to** length (x)− 8 **do**
5: $block = x\ (i : i+ 8)$;
6: $k = 0$; $u\ (i) = 0$; $l\ (i) = 0$;
7: **for** $j = 1$ **to** 9 **do**
8: **if** $j \neq 5$
9: **if** $block\ (j) - block\ (5) < -threshold$
10: $l\ (i) = l\ (i) + 2\ \hat{}k$;
11: **else if** $block\ (j) - block\ (5) > threshold$
12: $u\ (i) = u\ (i) + 2\ \hat{}k$;
13: **end**
14: $k = k+1$;
15: **end if**
16: **end for*j***
17: $h1\ (l\ (i) + 1) = h1\ (l\ (i) + 1) + 1$;
18: $h2\ (u\ (i) + 1) = h2\ (u\ (i) + 1) + 1$;
19: **end for*i***
20: Obtain 512 features with *h1* and *h2*.

Algorithm 3 Feature Extraction

Input: Speech audio (*sp*)
Output: The selected features (*sf*) with size

1: **for** $i = 1$ **to** 10 **do**
2: $features\ ((i-1) * 768 + 1 : i * 768) = [1D - LBP\ (sp)\ 1D - LTP(sp)]$
3: $[L, H] = 1D - DWT(sp, sym4)$
4: $sp = L$;
5: **end for*i***
6: Apply Min-Max normalization.
7: Apply NCA to features.
8: Select 1024 most discriminative features as *sf*.

the sample space to a higher dimension so that data can be separated linearly and also the kernels reduce the computational complexity by increasing the dimension [6], [33], [49]. The nonlinear classifier, SVM calculates the boundary of the decision precisely for speaker-independent and small-sized sample applications [6].

All steps used in the 1BTPDN method are in Algorithm 4. The framework of 1BTPDN is shown in Fig. 6.

VII. EXPERIMENTAL SETUP

The 1BTPDN method was tested on sentence based EMO-DB, EMOVO, SAVEE, and song based RAVDESS datasets to obtain comparable results.

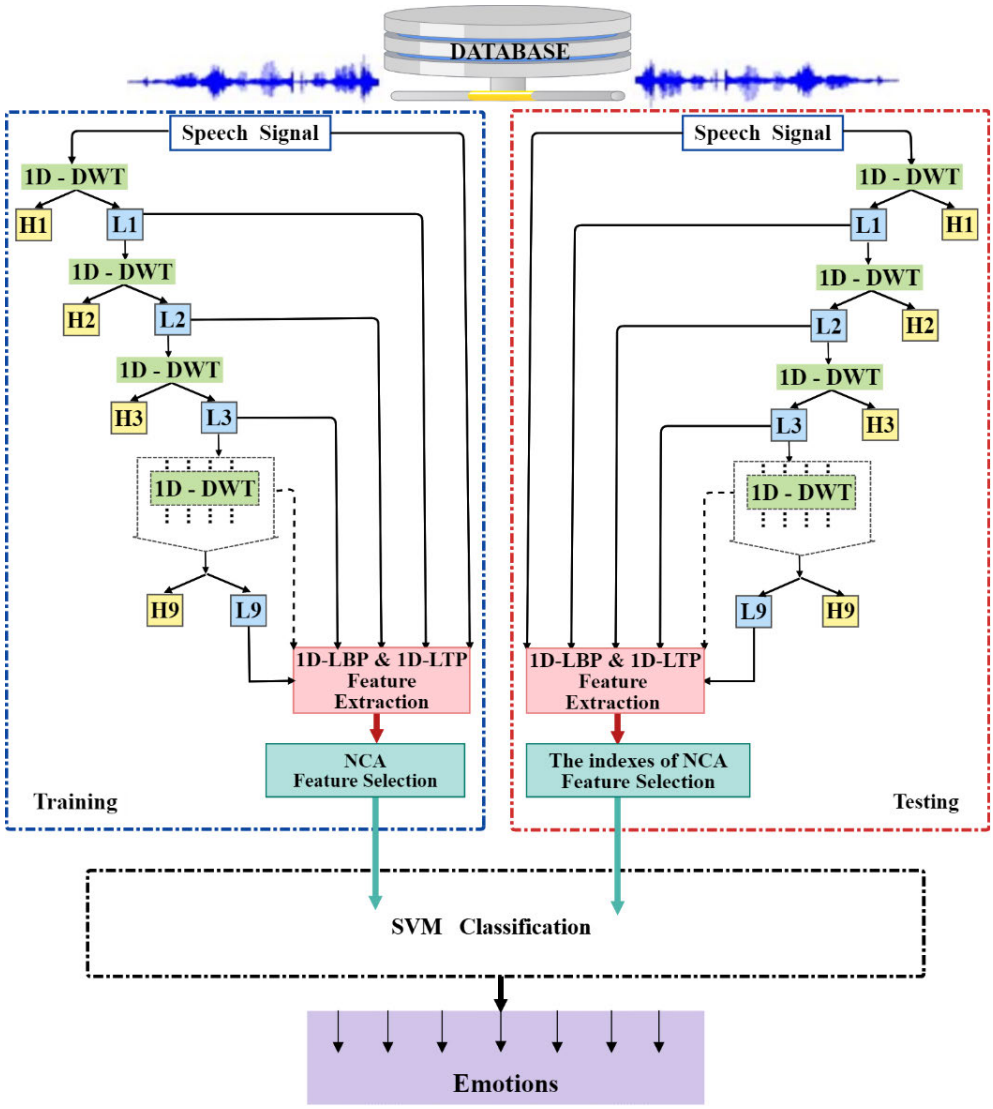

FIGURE 6. The framework that shows all steps of the 1BTPDN method.

A. EMOTIONAL SPEECH DATASETS

We selected widely used databases such as EMO-DB, EMOVO, SAVEE, and RAVDESS-Song in which emotions, chosen according to Ekman's discrete theory. In Table 1, the number of emotions, class numbers, and the emotion numbers in each class are presented. SER results can be easily compared in the literature using acted databases [32]; thus, these type of databases are chosen for this study.

In the EMO-DB Berlin emotional speech database, five female and five male actors recorded ten sentences for each of the seven emotions. 535 German speech audio data were recorded in different emotional states. These audio files in wav formats contain speech signals sampled at 16 KHz rate, 16-bit size, 256 kbps bit rate, and are approximately 2–3 seconds long [6], [50].

EMOVO corpus is an Italian emotional speech database that involves 14 Italian sentences voiced by three female and three male actors. It consists of 588 speech audio data in seven different emotional states. These audio files in stereo wav formats contain speech signals sampled at 48 KHz rate, 16-bit size, and 1.536 kbps bit rate [2], [51].

The Surrey audio-visual expressed emotion database (SAVEE) that involves 15 English sentences voiced by four male actors. It consists of 480 speech data in seven different emotional states. These audio files are in stereo wav formats and contain speech signals sampled at 44.1 KHz rate, 16-bit size, and 705 kbps bit rate [6], [7].

Ryerson Audio-Visual Database of Emotional Speech and Song Database (RAVDESS) [52] consists of two North American English sentences voiced by 24 (12 male, 12 female) actors, each sentence is voiced twice with normal and strong intensities. These audio files, in wav formats, contain signals sampled at 48 KHz rate, 16-bit size, and 768 kbps bit rate. The song part of RAVDESS includes 1,012 data in six different emotional states [52].

Algorithm 4 The Main Algorithm of 1BTPDN

Input: Speech data (*sp*)
Output: The selected features (*sf*) with size

1: Upload *sp* to the Matlab software.
2: Apply 1D-LBP and 1D-LTP to the raw *s*p to obtain 768 features.
3: Apply 1D-DWT (symlets 4) with nine levels to *sp* and obtain L1, L2, L3.....L9.
4: Generate 7680-feature set by concatenating features extracted from a total of ten levels.
5: Apply Min-Max Normalization.
6: Calculate the weights of the 7680 features using NCA.
7: Select the most discriminative 1024 features as *sf*.
8: Classify the *sf* features using SVM.

TABLE 1. Distribution of audio data by emotions in the datasets.

Emotions	EMO-DB	EMOVO	SAVEE	RAVDESS Song
Anger	127	84	60	184
Boredom	81	-	-	-
Calm	-	-	-	184
Disgust	46	84	60	-
Fear	69	84	60	184
Happiness	71	84	60	184
Neutral	79	84	120	92
Sadness	62	84	60	184
Surprise	-	84	60	-
Total	535	588	480	1012

B. EVALUATION METHOD AND PARAMETERS

We used 10-fold-cross validation method for the evaluation method since the data size of the datasets is different. We applied 10-fold-cross validation to obtain robust features. Matlab performs 10-fold-cross validation on all samples of the datasets. The samples in the database are divided into two, i.e., the training data and the testing data [53]. The test set is used to evaluate the classifier recognition ability [1]. The Matlab classification learner toolbox has automated cross-validation capability. It divides the observations automatically.

C. PERFORMANCE ANALYSIS PARAMETERS

We used statistical parameters to analyze the achievement of the 1BTPDN method. We calculated the accuracy rate, recall rate, precision rate, and F1-measure for evaluating the performance of the 1BTPDN method. The accuracy rate specifies the general recognition rate. Since the accuracy result alone will not be sufficient, other parameters such as precision, recall, F1 evaluation criteria are needed. The F1-measure is defined as the harmonic mean of the precision rate and the recall rate [42], [43].

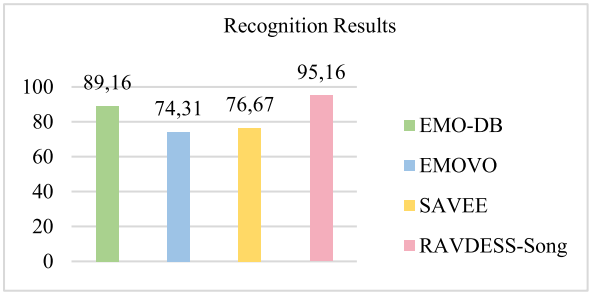

FIGURE 7. The accuracy results of the 1BTPDN method.

TABLE 2. The EMO-DB confusion matrix of 1BTPDN.

EMO-DB		Estimated Class: Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Real Class	Anger	121	0	1	1	4	0	0
	Boredom	0	67	0	3	0	9	2
	Disgust	2	0	39	3	2	0	0
	Fear	1	0	0	63	3	1	1
	Happiness	10	0	0	4	56	1	0
	Neutral	0	6	0	1	0	72	0
	Sadness	0	1	0	1	0	1	59

VIII. RESULTS

The system recognition rates are presented in Fig. 7. When the experimental results are examined, the success rates are 95.16% on RAVDESS, 89.16% on EMO-DB, 76.67% on SAVEE, and 74.31% on EMOVO. The highest recognition results in the databases with unbalanced distributions of emotions are obtained in the RAVDESS and EMO-DB. The lowest rate is obtained in the EMOVO database.

The results of the 1BTPDN method gives better results than the other texture analysis and the acoustic analysis methods. The feature extraction is carried out on raw audio data using LBP and LTP algorithms. As a result of the experimental studies, we observed the superiority of the LBP–LTP combination, applied without any preprocessing of the data. The reason for the highest recognition rate in RAVDESS-Song that it has less classes and a greater number of data. Also, songs reflect the emotions much better than speech sentences.

The confusion matrice of EMO-DB is shown in Table 2. Boredom is only present in the EMO-DB dataset.

The confusion matrice of EMOVO is given in Table 3.

The confusion matrice of SAVEE is shown in Table 4.

RAVDESS-Song are given in Table 5. Calm is only present in the RAVDESS-Song database.

Experimental results according to emotions are presented in Tables 6, 7, 8, and 9 for EMODB, EMOVO, SAVEE, and RAVDESS-Song, respectively.

TABLE 3. The EMOVO confusion matrix of 1BTPDN.

EMOVO		Estimated Class						
		Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Real Class	Anger	71	3	0	8	0	0	2
	Disgust	1	60	5	3	6	1	8
	Fear	1	3	62	6	3	5	4
	Happiness	8	4	7	38	11	2	14
	Neutral	0	2	0	4	76	2	0
	Sadness	0	0	5	2	2	75	0
	Surprise	0	8	11	9	1	0	55

TABLE 4. The SAVEE confusion matrix of 1BTPDN.

SAVEE		Estimated Class						
		Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Real Class	Anger	47	0	3	5	4	0	1
	Disgust	6	38	2	3	8	1	2
	Fear	4	0	36	5	1	4	10
	Happiness	6	0	2	43	1	1	7
	Neutral	1	1	0	0	115	3	0
	Sadness	1	0	1	0	11	47	0
	Surprise	2	5	5	5	0	1	42

TABLE 5. The RAVDESS-Song confusion matrix of 1BTPDN.

RAVDESS Song		Estimated Class					
		Anger	Calm	Fear	Happiness	Neutral	Sadness
Real Class	Anger	177	0	5	1	0	1
	Calm	0	179	1	4	0	0
	Fear	5	0	167	1	0	11
	Happiness	2	2	0	180	0	0
	Neutral	0	0	0	0	92	0
	Sadness	1	2	12	0	1	168

According to the EMO-DB experiments, the highest accuracy results of 95.27% are obtained for anger. The lowest recall rate, 78.87%, is achieved for happiness. Emotion-based success rates of texture analysis methods in [48] are the highest rate of 85.8% for anger and the lowest rate of 69.6% for disgust. Emotion-based success rates of the 1BTPDN are higher than success rates in [48].

According to the EMOVO experiments, the highest recall results of 90.47%, 89.28%, and 84.52% were obtained for neutral, sadness, and anger, respectively. The low accuracy results of 65.47% and 45.23% were achieved for surprise and happiness.

According to the SAVEE experiments, the highest recall results of 95.83%, 78.33%, and 78.33% are obtained for neutral, sadness, anger respectively, while the low recall results of 63.33% and 60% are achieved for disgust and fear. Emotion-based success rates of texture analysis methods in [48] are the highest rate of 85.8% for neutral and the lowest rate of 48.3% for fear. Emotion-based success rates of the 1BTPDN are higher than success rates in [48].

According to the RAVDESS-Song experiments, the highest recall results of 100%, 97.82%, and 97.28% are obtained for neutral, happiness, and calm, respectively, while the lowest accuracy results of being 91.30% and 90.76% that

TABLE 6. Experimental results of the 1BTPDN model for EMO-DB.

EMO-DB Emotions	Recall %	Precision %	F1-measure %
Anger	95.27	90.29	92.72
Boredom	82.71	90.54	86.45
Disgust	84.78	97.5	90.69
Fear	91.30	82.89	86.89
Happiness	78.87	86.15	82.35
Neutral	91.13	85.71	88.34
Sadness	95.16	95.16	95.16
Overall	**88.46**	**89.75**	**89.10**

TABLE 7. Experimental results of the 1BTPDN model for EMOVO.

EMOVO Emotions	Recall %	Precision %	F1-measure %
Anger	84.52	87.65	86.06
Disgust	71.42	75	73.17
Fear	73.80	68.88	71.26
Happiness	45.23	54.28	49.35
Neutral	90.47	76.76	83.06
Sadness	89.28	88.23	88.75
Surprise	65.47	66.26	65.86
Overall	**74.31**	**73.87**	**74.09**

TABLE 8. Experimental results of the 1BTPDN model for SAVEE.

SAVEE Emotions	Recall %	Precision %	F1-measure %
Anger	78.33	70.14	74.01
Disgust	63.33	86.36	73.07
Fear	60	73.46	66.05
Happiness	71.66	70.49	71.07
Neutral	95.83	82.14	88.46
Sadness	78.33	82.45	80.34
Surprise	70	67.74	68.85
Overall	**73.92**	**76.11**	**75.00**

TABLE 9. Experimental results of the 1BTPDN model for RAVDESS.

RAVDESS Emotions	Recall %	Precision %	F1-measure %
Anger	96.19	95.67	95.93
Calm	97.28	97.81	97.54
Fear	90.76	90.27	90,51
Happiness	97.82	96.77	97.29
Neutral	100	98.92	99.45
Sadness	91.30	93.33	92.30
Overall	**95.56**	**95.46**	**95.51**

TABLE 10. The accuracy results of 1BTPDN method using DT, LDA, KNN classifiers.

Classifiers	EMO-DB	EMOVO	RAVDESS Song	SAVEE
	Acc. %	Acc. %	Acc. %	Acc. %
DT	58.70	38.10	61.17	43.75
LDA	76.70	51.90	43.70	54.60
KNN	76.07	47.27	90.91	61.70
SVM	89.16	74.31	95.16	76.67

are achieved for sadness and fear. Since songs reflect emotions better than speech, the success rate is much higher in RAVDESS.

When the experimental results are examined, the unweighted average recall results of the 1BTPDN are 95.56% on RAVDESS, 88.46% on EMO-DB, 74.31% on EMOVO, and 73.92% on SAVEE. The accuracy and recall results are the same for EMOVO, because the number of data in classes are equal in EMOVO. The recall rates in EMO-DB and SAVEE are lower than the accuracy rates, while the RAVDESS-Song recall rate is slightly higher than the accuracy rate. Fig. 8 shows the emotion-based recognition rates of the 1BTPDN, graphically.

The 1BTPDN method has more successful recognition for anger, neutral, sadness and fear than other emotions.

We compared the classification using in our proposed method among other traditional classifiers such as decision trees, linear discriminant analysis, and k-nearest neighbor. It is shown in Table 10 that SVM Cubic emotion recognition rates are very successful according to the recognition rates obtained using DT, LDA, and KNN. The overall recognition performance is decreased with baseline methods. All classifiers were re-tested without feature selection but worse results and too long run time occurred.

IX. DISCUSSION

While the success rate of the methodology is promising, there are challenges and limitations in the 1BTPDN method:

1. An accurate choice of the emotional speech database (speech corpora) is the most important part of the SER.

TABLE 11. The SER methods as a benchmark comparison.

Ref.	Datasets	Input Data & Features	Classifier	Acc. (%)
[9]	EMO-DB	Raw data & ZCR, STE, MFCC	GMM SVM	76.31 81.57
[16]	EMO-DB	Raw data & MFCC	HMM	80
[21]	EMO-DB RAVDESS Speech	Spectrograms & Deep Features	RBFN CNN BILSTM	85.57 77.02
[22]	RAVDESS Speech	Spectrograms	DSCNN	79.5
[24]	EMO-DB	Raw data & LPC-TEO Raw data & LPC-Pitch	GMM GMM	88 78.7
[25]	SAVEE	Raw data & Weighted features	ELM PNN	73.81 73.45
[29]	EMO-DB SAVEE EMOVO	Raw data & Acoustic features	SVM	84.62 74.39 60.4
[34]	SAVEE	Raw data & Spectrogram MFCC, Pitch, Deep features	K-Means	80.41
[36]	SAVEE	Spectrogram & Deep features	DNN	59.7
[49]	EMO-DB SAVEE	Spectrogram & Raw data Textural & Acoustic features	SVM SVM	84.5 75.9
[54]	EMO-DB	Raw Data LPC & Deep features	LSM-SNN	82.35
[55]	EMO-DB	Raw data MFCC & Deep features	CNN LSTM	80
[56]	EMO-DB	Spectrogram & Deep features	CNN	84.3
[57]	SAVEE	Raw data MFCC & Pitch & Energy	SVM	82.8
[58]	EMO-DB SAVEE	Raw data Cepstral & wavelets & VQ based wavelets	RBFN	91.82 93.67
[59]	EMO-DB	Raw data & Spectrogram	LSTM-CNN	86.73 95.89
[60]	EMO-DB	Spectrogram Deep frequency features	CNN	92.01
1BTPDN method	EMO-DB SAVEE EMOVO RAVDESS Song	Raw Data Textural local features	SVM	89.16 76.67 74.31 95.16

Low-quality or inaccurate data can lead to false predictions for the SER [6]. For a successful SER, a context-rich emotional speech database is preferred. We used

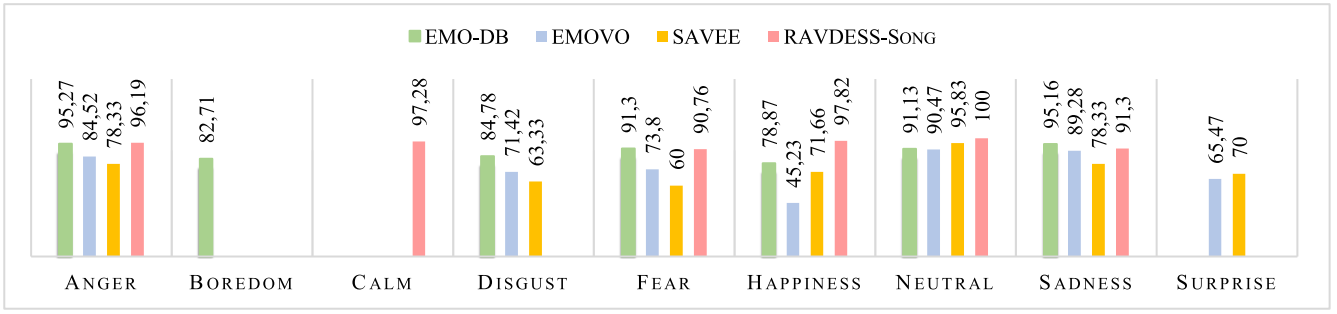

FIGURE 8. The success rates of the 1BTPDN method for each emotion.

acted datasets. This method can be test on large datasets and real-time audio data in future studies.

2. Databases are in many different languages such as English, German, Italian, Chinese, Russian, and Spanish. The selected language is another vital decision for the success of the SER results. In this study, English, German and Italian languages were used. Databases in other languages can be added in future studies.
3. The inequality of the number of data in the classes is another limitation.

Despite the unbalanced distribution in the databases, successful results have been obtained for sentence based EMO-DB, EMOVO, SAVEE, and song based RAVDESS according to the previous studies in the literature, including deep learning algorithms.

Table 11 is a comparison table showing many studies performed on EMO-DB, EMOVO, SAVEE, and RAVDESS.

The SER results can be easily compared in the literature using acted databases [32], so we selected acted databases for our proposed method. With the 1BTPDN method, RAVDESS results are highly accurate as the songs reflect the emotions better. EMOVO which contains Italian sentences has the lowest success. The ML methods have not reached high classification accuracy for the EMOVO [29], [49]. Therefore it is one of the hardest corpora for ML based speech emotion recognition. We can list the comparison results with the studies using handcrafted features and traditional classifiers as follows.

The accuracy results of [9] and [16] for EMO-DB are 7.59% and 9.16% lower than the accuracy results of the 1BTPDN method. The accuracy result of [24] for EMO-DB is 1.16% and 10.46% lower than the accuracy result of 1BTPDN method. The accuracy results of [57] for SAVEE are 6.13% upper than the accuracy result of 1BTPDN method. Comparing 1BTPDN and [29], the success rate increased by 13.91% for EMOVO, 4.54% for EMO-DB and 2.28% for SAVEE. The reason for the high increase in the EMOVO is that the EMOVO has uniform emotion distribution in the classes. Reference [49] gives less successful accuracy rates than the 1BTPDN method.

Reference [58] which uses neural network, [59] and [60] which use deep learning algorithms have more successful results than 1BTPDN method.

X. CONCLUSION AND FUTURE WORK

We observed new or less-tried algorithms for feature extraction engineering after a thorough examination of the literature. As a result, a novel text-independent and speaker-independent, SER method, called 1BTPDN has been developed with a lightweight method that solves a nonpolynomial problem by extracting handcrafted features. The 1BTPDN model, based on texture analysis, the combination of 1D-LBP and 1D-LTP, reduces the loss of important information and increases the accuracy rate. Obtaining successful results without any preprocessing of the raw data such as windowing, framing, and without any preliminary definitions, MFCC and LPCC also show the superiority of this model. Another important point is the combination of LBP, LTP, DWT, and NCA, not only the texture analysis methods. The combination of NCA and DWT extracts the 1024 most distinctive features from the 7680 features. The experimental results indicate that textural features can be as successful as linguistic, prosodic, and spectral features. The computational complexity of the 1BTPDN is as low as $O(nlogn)$. There is no need to set many parameters such as deep learning algorithms. The 1BTPDN method can be used in psychiatric clinics and courts. In future studies, more effective methods can be developed by using datasets involving real-time audio data. Moreover, cross-subject and cross-sentence validations can be conducted to evaluate the generalization for new and same subjects.

REFERENCES

[1] Y. Sönmez and A. Varol, "In-depth analysis of speech production, auditory system, emotion theories and emotion recognition," in *Proc. 8th ISDFS*, Beirut, Lebanon, 2020, pp. 1–8.

[2] A. Mencattini, E. Martinelli, G. Costantini, M. Todisco, B. Basile, M. Bozzali, and C. Di Natale, "Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure," *Knowl.-Based Syst.*, vol. 63, pp. 68–81, Jun. 2014.

[3] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Appl. Acoust.*, vol. 158, Jan. 2020, Art. no. 107020.

[4] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," *Electron. Notes Theor. Comput. Sci.*, vol. 343, pp. 35–55, May 2019.

[5] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, Jul. 2020.

[6] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, Jan. 2020.

[7] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies," *ATSIP*, vol. 3, pp. 1–18, Nov. 2014.

[8] X. Zhang, Y. Sun, and S. Duan, "Progress in speech emotion recognition," in *Proc. TENCON*, Macao, China, 2015, pp. 1–6.

[9] S. Lukose and S. S. Upadhya, "Music player based on emotion recognition of voice signals," in *Proc. ICICICT*, Kannur, India, 2017, pp. 1751–1754.

[10] J. Villalba, A. Ortega, A. Miguel, and E. Lleida, "Analysis of speech quality measures for the task of estimating the reliability of speaker verification decisions," *Speech Commun.*, vol. 78, pp. 42–61, Apr. 2016.

[11] G. Liu, W. He, and B. Jin, "Feature fusion of speech emotion recognition based on deep learning," in *Proc. IC-NIDC*, Guiyang, China, 2018, pp. 193–197.

[12] K. R. Anne, S. Kuchibhotla, and D. Vankayalapati, "Acoustics modeling for emotion recognition," in *SpringerBriefs in Speech Technology*. New York, NY, USA: Springer, 2015. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-15530-2

[13] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC," in *Proc. WiSPNET*, Chennai, India, 2017, pp. 2257–2260.

[14] N. Dave, "Feature extraction methods LPC, PLP and MFCC in speech recognition," *Int. J. Adv. Res. Eng. Technol.*, vol. 1, no. 6, pp. 1–4, 2013.

[15] S. Basu, J. Chakraborty, and M. Aftabuddin, "Affect detection from speech using deep convolutional neural network architecture," in *Proc. 14th IEEE INDICON*, Roorkee, India, Dec. 2017, pp. 1–5.

[16] S. Sahoo and A. Routray, "MFCC feature with optimized frequency range: An essential step for emotion recognition," in *Proc. ICSMB*, Kharagpur, India, 2016, pp. 162–165.

[17] Y. Sönmez and A. Varol, "New trends in speech emotion recognition," in *Proc. ISDFS*, Barcelos, Portugal, 2019, pp. 1–7.

[18] D. Torres-Boza, M. C. Oveneke, F. Wang, D. Jiang, W. Verhelst, and H. Sahli, "Hierarchical sparse coding framework for speech emotion recognition," *Speech Commun.*, vol. 99, pp. 80–89, May 2018.

[19] K. Wang, Z. Zhu, J. Zhang, and L. Chen, "Speech emotion recognition of chinese elderly people," *Web Intell.*, vol. 16, no. 3, pp. 149–157, Sep. 2018.

[20] P.-W. Hsiao and C.-P. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *Proc. IEEE ICASSP*, Apr. 2018, pp. 2526–2530.

[21] M. Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.

[22] M. Mustaqeem and S. Kwon, "CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, Dec. 2019.

[23] H. Zhao, N. Ye, and R. Wang, "A survey on automatic emotion recognition using audio big data and deep learning architectures," in *Proc. 4th BigDataSecurity*, Omaha, NE, USA, May 2018, pp. 139–142.

[24] S. R. Bandela and T. K. Kumar, "Emotion recognition of stressed speech using Teager energy and linear prediction features," in *Proc. IEEE 18th ICALT*, Mumbai, India, Jul. 2018, pp. 422–425.

[25] C. K. Yogesh, H. Muthusamy, R. Yuvaraj, R. Ngadiran, A. H. Adom, S. Yaacob, and K. Polat, "Bispectral features and mean shift clustering for stress and emotion recognition from natural speech," *Comput. Electr. Eng.*, vol. 62, pp. 676–691, Aug. 2017.

[26] E. Alexandre, L. Cuadra, M. Rosa, and F. Lopez-Ferreras, "Feature selection for sound classification in hearing aids through restricted search driven by genetic algorithms," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 8, pp. 2249–2256, Nov. 2007.

[27] J. Rong, G. Li, and Y.-P.-P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," *Inf. Process. Manage.*, vol. 45, no. 3, pp. 315–328, May 2009.

[28] T. Özseven and M. Düğenci, "SPeech ACoustic (SPAC): A novel tool for speech feature extraction and classification," *Appl. Acoust.*, vol. 136, pp. 1–8, Jul. 2018.

[29] T. Özseven, "A novel feature selection method for speech emotion recognition," *Appl. Acoust.*, vol. 146, pp. 320–326, Mar. 2019.

[30] E. Yıldız and Y. Sevim, "Comparison of linear dimensionality reduction methods on classification methods," in *Proc. Nat. Conf. Electr., Electron. Biomed. Eng. (ELECO)*, Bursa, Turkey, 2016, pp. 161–164.

[31] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. NIPS*, Vancouver, BC, Canada, 2004, pp. 513–520.

[32] Y. B. Singh and S. Goel, "Survey on human emotion recognition: Speech database, features and classification," in *Proc. ICACCCN*, Greater Noida, India, Oct. 2018, pp. 298–301.

[33] A. Rajasekhar and M. K. Hota, "A study of speech, speaker and emotion recognition using mel frequency cepstrum coefficients and support vector machines," in *Proc. ICCSP*, Chennai, India, Apr. 2018, pp. 0114–0118.

[34] N. Hajarolasvadi and H. Demirel, "3D CNN-based speech emotion recognition using K-means clustering and spectrograms," *Entropy*, vol. 21, no. 5, p. 479, May 2019.

[35] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, "A Review on Emotion Recognition using Speech," in *Proc. ICICCT*, Coimbatore, India, Mar. 2017, pp. 109–114.

[36] H. M. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using deep neural networks," in *Proc. 9th ICSPCS*, Cairns, QLD, Australia, 2015, pp. 1–5.

[37] T. Ojala, M. Pietikäinen, and D. A. Harwood, "Comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.

[38] Y. Kaya, M. Uyar, R. Tekin, and S. Yıldırım, "1D-local binary pattern based feature extraction for classification of epileptic EEG signals," *Appl. Math. Comput.*, vol. 243, pp. 209–219, Sep. 2014.

[39] S. Sahoo and A. Routray, "Emotion recognition from audio-visual data using rule based decision level fusion," in *Proc. IEEE Students' Technol. Symp. (TechSym)*, Kharagpur, India, Sep. 2016, pp. 7–12.

[40] N. Chatlani and J. J. Soraghan, "Local binary patterns for 1-D signal processing," in *Proc. 18th EUSIPCO*, Aalborg, Denmark, 2010, pp. 95–99.

[41] T. Tuncer, S. Dogan, and F. Ertam, "Automatic voice based disease detection method using one dimensional local binary pattern feature extraction network," *Appl. Acoust.*, vol. 155, pp. 500–506, Dec. 2019.

[42] Y. Kaya and Ö. F. Ertuğrul, "A stable feature extraction method in classification epileptic EEG signals," *Australas. Phys. Eng. Sci. Med.*, vol. 41, no. 3, pp. 721–730, Sep. 2018.

[43] S. M. Adnan, A. Irtaza, S. Aziz, M. O. Ullah, A. Javed, and M. T. Mahmood, "Fall detection through acoustic local ternary patterns," *Appl. Acoust.*, vol. 140, pp. 296–300, Nov. 2018.

[44] T. Tuncer, S. Dogan, F. Özyurt, S. B. Belhaouari, and H. Bensmail, "Novel multi center and threshold ternary pattern based method for disease detection method using voice," *IEEE Access*, vol. 8, pp. 84532–84540, 2020.

[45] A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," *Expert Syst. Appl.*, vol. 32, no. 4, pp. 1084–1093, May 2007.

[46] U. Orhan, M. Hekim, and M. Özer, "EEG signals classification using the K-means clustering and a multilayer perceptron neural network model," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13475–13481, Sep. 2011.

[47] D. Yuan and C. D. Elvidge, "Comparison of relative radiometric normalization techniques," in *Proc. ISPRS*, vol. 51, no. 3, pp. 117–126, 1996.

[48] C. Qin, S. Song, G. Huang, and L. Zhu, "Unsupervised neighborhood component analysis for clustering," *Neurocomputing*, vol. 168, pp. 609–617, Nov. 2015.

[49] T. Özseven, "Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition," *Appl. Acoust.*, vol. 142, pp. 70–77, Dec. 2018.

[50] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.

[51] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: An Italian emotional speech database," in *Proc. LREC*, Reykjavik, Iceland, 2014, pp. 3501–3504.

[52] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[53] A. Sharma and D. V. Anderson, "Deep emotion recognition using prosodic and spectral feature extraction and classification based on cross validation and bootstrap," in *Proc. IEEE SP/SPEW*, Salt Lake City, UT, USA, Jan. 2015, pp. 421–425.

[54] R. Lotfidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," in *Proc. IEEE ICASSP*, New Orleans, FL, USA, Mar. 2017, pp. 5135–5139.

[55] S. Basu, J. Chakraborty, and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," in *Proc. 2nd ICCES*, Coimbatore, India, Oct. 2017, pp. 333–336.

[56] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. Int. Conf. Platform Technol. Service (PlatCon)*, Busan, South Korea, 2017, pp. 1–5.
[57] S. Susan and A. Kaur, "Measuring the randomness of speech cues for emotion recognition," in *Proc. 10th Int. Conf. Contemp. Comput.*, Noida, India, Aug. 2017, pp. 1–6.
[58] H. K. Palo and M. N. Mohanty, "Wavelet based feature combination for recognition of emotions," *AIN Shams Eng. J.*, vol. 9, no. 4, pp. 1799–1806, Dec. 2018.
[59] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, Jan. 2019.
[60] T. Anvarjon, M. Mustaqeem, and S. Kwon, "Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features," *Sensors*, vol. 20, no. 18, p. 5212, Sep. 2020.

YEŞİM ÜLGEN SÖNMEZ received the B.S. degree from the Electronics and Computer Education Department, Faculty of Technical Education, Fırat University, Elâzığ, Turkey, in 2002, and the M.S. degree from the Electronics and Computer Education, Control Systems Department, Faculty of Technical Education, Fırat University, in 2008. She is currently pursuing the Ph.D. degree with the Software Engineering Department, Faculty of Technology, Fırat University. She started working as an Official Computer Teacher. Her research interests include digital forensics and speech emotion recognition.

ASAF VAROL received the B.S. degree (Hons.) from the Department of Mechanical Engineering, Fırat University, Turkey, the M.S. degree from the Institute for Nuclear Energy, Istanbul Technical University, in 1977, the second master's degree in public administration from Sam Houston State University (SHSU), USA, in 2017, and the Ph.D. degree from Karadeniz Technical University, Turkey, in 1983. He started working at the Institute for Nuclear Energy, Istanbul Technical University, where he gained a vital position as a Mechanical Engineer at ITU Research Reactor TRIGA MARK II. He studied Intensive German Language and attended the Machine Engineering Training Course that was supported by IAESTE, Kassel, Germany, from 1972 to 1973. He became a Research Assistant in the field of energy at Fırat University, in 1979. In 1990, he had academic studies on the field of computer systems at Indiana and Purdue universities, USA, using a scholarship of the YOK/World Bank. He was promoted to an Associate Professor position in the field of the energy education. In 1992, he had academic studies on computer aided education and design at Salford and Bradford universities, U.K. In 1995, he was invited to Oklahoma State University, USA, in order to have some studies on vocational and technical education and informatics. He earned the title of Professor in the field of computer systems education, in 1997, at Fırat University. He was also invited to Germany through the Institution of DAAD to work on a project in the field of robotics at Bremen University, in 1998. From 2003 to 2005, he taught different courses for one and half years at West Virginia University, USA. He was a Visiting Professor with Wilkes University, in 2011. He was promoted to the full time Professor position in the field of software engineering at Fırat University. He gave some courses at Sam Houston State University, in 2013 and from 2018 to 2019. He is currently the Founder of the Department of Digital Forensics Engineering, Fırat University. He established the International Joint Degree Program, which enables student exchange between Fırat University and SHSU. He has been promoted to a full time Professor position at Maltepe University, İstanbul, Turkey, since September 1, 2020, where he is currently working. He has published more than 300 articles, proceedings, books, and scientific research reports. His research interests include digital forensics, artificial intelligence, machine learning, distance education, robotics, and so on.