IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 25, 2019, accepted April 8, 2019, date of publication April 15, 2019, date of current version May 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2911314

A Pathological Multi-Vowels Recognition Algorithm Based on LSP Feature

TAO ZHANG[1,2], (Member IEEE), YAQIN WU[1], YANGYANG SHAO[1], MINGYANG SHI[1], YANZHANG GENG[1], AND GANJUN LIU[1]

[1]School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
[2]Texas Instruments DSP Joint Lab, Tianjin University, Tianjin 300072, China

Corresponding authors: Yanzhang Geng (gregory@tju.edu.cn) and Ganjun Liu (ganjun_liu@tju.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61350009.

ABSTRACT At present, pathological voice recognition is mainly based on the classification of pathological voice. However, almost all the researches are based on the single vowel \a\ samples, but few on multi-vowels. In addition, the current researches on multi-vowels recognition are mainly for normal voices, which are unsuitable for the speech recognition of normal and pathological multi-vowels simultaneously. This paper concentrates on developing an accurate and robust feature called enhanced-bark line spectrum pair (E-BLSP) to detect and classify normal and pathological multi-vowels. We explore the impact of E-BLSP feature on recognition performance and propose an effective method based on the combination of three features including E-BLSP for pathological and normal multi-vowels. In this paper, first LSP and difference of adjacent LSP (DAL) features of a vowel are extracted. Then LSP feature is warped at bark domain to get bark line spectrum pair (BLSP). In addition, then E-BLSP feature is calculated by adjusting BLSP using DAL feature. Finally, the adjusted E-BLSP feature and other two traditional features, including linear prediction cepstrum coefficient (LPCC) and mel-frequency cepstrum coefficients (MFCC) are applied to support vector machine (SVM) and deep neural network (DNN) classifiers to explore the classification performance of single feature and feature combinations for pathological and normal vowels /a/, /i/ and /u/. The results show that the highest achieved accuracies for DNN and SVM network are 98.6190% and 96.2693%, while the largest achieved area under curves (AUC) are 0.9925 and 0.9868, correspondingly with the combination of three features including LPCC, MFCC, and E-BLSP.

INDEX TERMS Pathological voice, multi-vowels, E-BLSP, feature combination, SVM, DNN.

I. INTRODUCTION

Individuals in the world are increasingly facing the risk of various pathological voice problems. Around 25% of the world population suffer from all kinds of voice problems because of the excessive vocalization caused by their own professions. Studies conducted give the fact that during the whole life of Americans, the spread of voice disorders for teachers and non-teachers account for 57.7% and 28.8%, respectively [1]. In addition, about 33% of teachers present with various voice disorders at some stage of their lives in the Riyadh area of Saudi Arabia [2]. According to the survey, 5% of U.S. children ages 3-17 have a speech disorder that lasted for a week or longer during the past 12 months in 2015 [3]. Besides, in Saudi Arabia, about 15% of patients visiting King Abdul Aziz University Hospital in Riyadh suffer from various voice problems [4]. Further, in the UK, around 2200 people are diagnosed with the laryngocarcinoma each year [5]. Therefore, it is very important to identify pathological voices and then repair them because voice is a means of communication with other people and the surrounding world.

At present, there are many medical methods for the diagnosis and treatment of voice diseases, such as voice microsurgical technique, vocal cord injection filling technique and laryngeal framework surgery. But as invasive examinations, the incompleteness of treatments will affect language expression and auditory perception of the patients [6]. As a consequence, working on the digital processing of speech signals has been found to provide a noninvasive analytical technique that is considered to be an effective assisting

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

tool to recognize pathological voices. In the research for speech recognition, multi-vowels recognition is the premise of speaker-independent continuous speech recognition with large vocabulary. At present, the research object of multi-vowels recognition is based on normal voices. The features commonly used in other studies are LPCC, MFCC, formants and so on. A Probabilistic Neural Network (PNN) neural network model was employed to identify these five vowels {a, e, i, o, u}, further, the effect of the smoothing parameters on the recognition of PNN was also studied in [7]. In view of the fact that the formant is one of the most important features in vowel recognition, the formants of three normal vowels /a/, /i/ and /u/ standardized by K-L transform were used for vowel recognition in [8]. After that, the reduced sets of MFCCs were adopted in Sharma and Das [9] to recognize vowels, which greatly improved the recognition performance for specific voices. Furthermore, the recognition of Kannada vowels with Euclidean distance by extracting LPC features of these vowels was performed in Unnibhavi and Jangamshetti [10], where the experimental accuracy reached 40%. In addition to the features mentioned above for vowel recognition, Relative Spectral Transform and Perceptual Linear Prediction (RASTA-PLP) [11], Amplitude Modulation Spectrogram (AMS) [12] features, and wavelet-based features [13] are often used in the field of normal speech recognition.

In the early researches of pathological voice recognition, most focused on the binary classification of pathological voice and normal voice. Considering most acoustic features have higher recognition performance for pathological vowel /a/ than other pathological vowels, the research object is generally pathological vowel /a/. There are different types of signal analyses that can be used to perform automatic voice pathology, such as long-term signal features. Jitter, Shimmer and seven other acoustic features which were extracted from residual estimator, to evaluate the laryngeal diseases in [14], where the accuracy of jitter achieved 54.8% for 21 pathologies. A joint time-frequency method applied to classify pathological voices in terms of continuous speech signals was proposed in [15]. According to this method, the speech signals were decomposed firstly by an adaptive time-frequency transform algorithm. Then the extracted features such as octave max, octave mean, energy ratio, length ratio, and frequency ratio could be obtained. Finally, statistical pattern classification techniques were adopted to classify the pathological voices. The experiments yielded a classification accuracy rate of 93.4% using Massachusetts Eye and Ear Infirmary (MEEI) database. Aiming at classifying pathological voice, a combined method of feature reduction methods including principal component analysis and linear discriminant analysis followed by SVM was mentioned in [16]. The experimental results denoted that linear discriminant analysis with SVM had the best recognition rate of 94.26%. Moreover, features based on regression such as MPEG-7 audio feature and Multidirectional Regression features were also used for pathological speech recognition with high accuracies [17], [18]. Aiming to increase the accuracy of the system developed for automatic detection of pathological voices, [19] evaluated the discrimination performance of 11 features extracted by nonlinear analysis, and SVM and GMM classifiers were combined to get the optimal accuracy of 98.23%±0.001 with MEEI database. Apart from MEEI database, Al Nasheri *et al.* [20] extracted the features including maximum peak values, corresponding lag values and the entropy, and then evaluated the contribution of different frequency bands to the classification process based on MEEI, SaarbrÜCken Voice Database (SVD) and Arabic Voice Pathology Database (AVPD). The highest obtained accuracies for classification were 99.54%, 99.53% and 96.02% for MEEI, SVD and AVPD, respectively, using SVM classifier. In recent years, MFCC features from MEEI database were extracted in [21], where the performances of three machine learning algorithms, namely, DNN, SVM and Gaussian mixture model, were evaluated based on a five fold cross-validation. The experimental results demonstrated that DNN obtained higher accuracy (99.32%) than the other two classifiers. In case of classification features for pathological voices, classification performance of glottic features on MEEI and SVD databases was studied in [22]. The experimental results showed that the optimal classification rates of SVD and MEEI databases were 99.27% and 93.66%, respectively. In addition to the glottic features, the Cepstral Peak Prominence Smoothed (CPPS) distribution was dealt with in Castellana *et al.* [23]. The result indicated that the CPPS features had a strong discrimination performance with AUC of 0.95.

In order to solve the problem that features and methods applied to the recognition of normal multi-vowels are unsuitable for normal and pathological multi-vowels, first, a new E-BLSP feature is proposed based on LSP feature which is the most important acoustic feature affecting the timbre of vowels with good quantization and interpolation characteristics in this paper. Then a new method based on the combination of three features including E-BLSP, LPCC and MFCC is proposed for recognition of normal and pathological multi-vowels. The proposed method achieves six classifications of normal vowels /a/, /i/, /u/ and pathological vowels /a/, /i/, /u/ with high recognition rate. The remainder of this paper is organized as follows. Section II describes the methodology of this paper. Section III presents some experimental results of our algorithm for pathological multi-vowels recognition. Finally the conclusion is given in Section IV.

II. METHODOLOGY

A pathological multi-vowels recognition algorithm based on LSP feature is proposed in this paper. Fig. 1 depicts a block diagram of the system used in this paper. This paper realizes the recognition of normal and pathological multi-vowels from the following three aspects. Firstly, the LSP and DAL features are extracted from normal and pathological multi-vowels signals based on the speech spectrum model, and then BLSP feature is obtained by Bark frequency warping considering

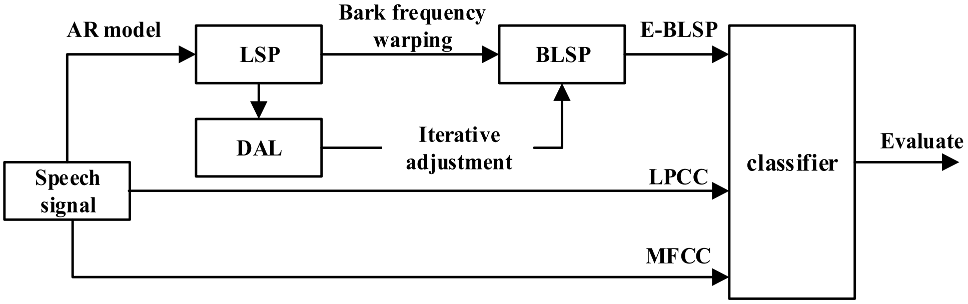


FIGURE 1. Structure of the proposed algorithm.

the non-uniform sensitivity characteristic that frequency resolution of human ears to low frequency is higher than high frequency. Secondly, in order to enhance the formant strength at spectrum peaks and improve the recognition rate of BLSP feature in the classification stage, bidirectional iterative adjustment of BLSP feature is carried out by using the DAL feature. At the same time, the BLSP feature is post-filtered by adjusting enhancement coefficients to get a new feature named E-BLSP. Thirdly, E-BLSP feature and other features such as LPCC, MFCC and delta MFCC features are applied to SVM or DNN classification networks. The discrimination capabilities of the proposed feature called E-BLSP are studied and compared with other features mentioned above along this paper. Thus, a method with the highest classification accuracy based on feature combination is proposed. Moreover, in order to evaluate the classification performance of different networks, DNN and SVM networks are used in the study. The experiment in this paper has been carried out with the following methodology.

A. THE MODEL OF SPEECH SPECTRUM

The concept of Line Spectrum Pair (LSP) was proposed by Itakura [24] in 1975. LSP feature is the frequency domain representation of linear prediction coefficients. It not only has better dynamic range and filtering stability, but also has better matching characteristics with formants. Position, width and amplitude of formants are main factors to determine vowel timbre. Generally speaking, it is enough to describe a vowel with only the first three formants.

Based on different speech spectrum analysis frameworks, computing methods of LSP feature are different. Among these frameworks, the Mel Generalized Cepstrum (MGC) analysis [25], [26] is one of the most effective methods. In this method, the model spectrum based on MGC can be varied continuously from all-pole to cepstral modeling by changing the values of the parameters. We assume that the speech spectrum system functions can be expressed as:

$$H_{MGC}(z) = \begin{cases} \left(1+\gamma \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\, z_{\partial}^{-m}\right)^{1/\gamma}, & 0<|\gamma|\leq 1 \\ \exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\, z_{\partial}^{-m}, & \gamma = 0 \end{cases} \qquad (1)$$

where $c_{\alpha,\gamma}(m)$ is MGC coefficient, α is Mel parameter and γ is generalized cepstrum parameter.

GMC model actually establishes a unified speech spectrum analysis framework. As shown in Table 1, with different values of (α, γ), common analytical models include Auto-Regressive (AR) model, Cepstrum (CEP) model, Mel Cepstrum (MCEP) model and Warped Linear Prediction (WLP) model.

According to GMC theory, for the same Mel parameter α, AR model can describe spectral peaks more effectively, and its orders should be consistent with the number of formants. CEP model corresponding to $\gamma = 0$ portrays valleys of the spectrum more effectively, but it is not as clear as the linear prediction in portraying spectral peaks. For the same generalized cepstrum parameter γ, with the increase of α in the range of $[0, 1)$, the low-frequency information of spectrum envelope increases. For the vowel signal studied in this paper, the vocal tract transfer function is actually an AR model without considering nasals and fricatives, so AR model is chosen to calculate the LSP feature. Considering that human ears have different resolutions for different frequency signals, LSP feature will be processed by frequency warping in Part B.

B. FREQUENCY WARPING

Based on AR model, the system function of the linear predictive inverse filter is assumed to be:

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} \qquad (2)$$

where p is the order of AR model and a_i is linear predictive coefficient. In this case, the amplitude of the linear predictive spectrum is defined by:

$$\begin{aligned} \left|H\left(e^{j\omega}\right)\right| &= \frac{1}{\left|A\left(e^{j\omega}\right)\right|} = \frac{2}{\left|P\left(e^{j\omega}\right)+Q\left(e^{j\omega}\right)\right|} \\ &= 2^{(1-P)/2}\left[\sin^2(\omega/2)\prod_{i=1}^{p}(\cos\omega-\cos\theta_i)^2 \right. \\ &\qquad \left. +\cos^2(\omega/2)\prod_{i=1}^{p}(\cos\omega-\cos\omega_i)^2\right]^{-1} \end{aligned} \qquad (3)$$

where $P(e^{j\omega})$ is the $(p+1)-order$ symmetric polynomial of $A(e^{j\omega})$, $Q(e^{j\omega})$ is the $(p+1)-order$ anti-symmetric polynomial of $A(e^{j\omega})$, $\cos\theta_i$ and $\cos\omega_i$ are expressions of LSP coefficients in cosine domain, θ_i and ω_i are Linear Spectrum

TABLE 1. Several examples of MGC model.

MGC coefficients	Spectral model	System function
$(\alpha,\gamma)=(0,-1)$	AR	$H(z)=1/\left(1-\sum_{m=0}^{M}c_{\alpha,\gamma}(m)z^{-m}\right)$
$(\alpha,\gamma)=(0,0)$	CEP	$H(z)=\exp\sum_{m=0}^{M}c_{\alpha,\gamma}(m)z^{-m}$
$(\alpha,\gamma)=((0,1),0)$	MCEP	$H(z)=\exp\sum_{m=0}^{M}c_{\alpha,\gamma}(m)\tilde{z}_{\alpha}^{-m}$
$(\alpha,\gamma)=((0,1),-1)$	WLP	$H(z)=1/\left(1-\sum_{m=0}^{M}c_{\alpha,\gamma}(m)\tilde{z}_{\alpha}^{-m}\right)$

Frequency (LSF) features corresponding to LSP coefficients. DAL feature is defined by:

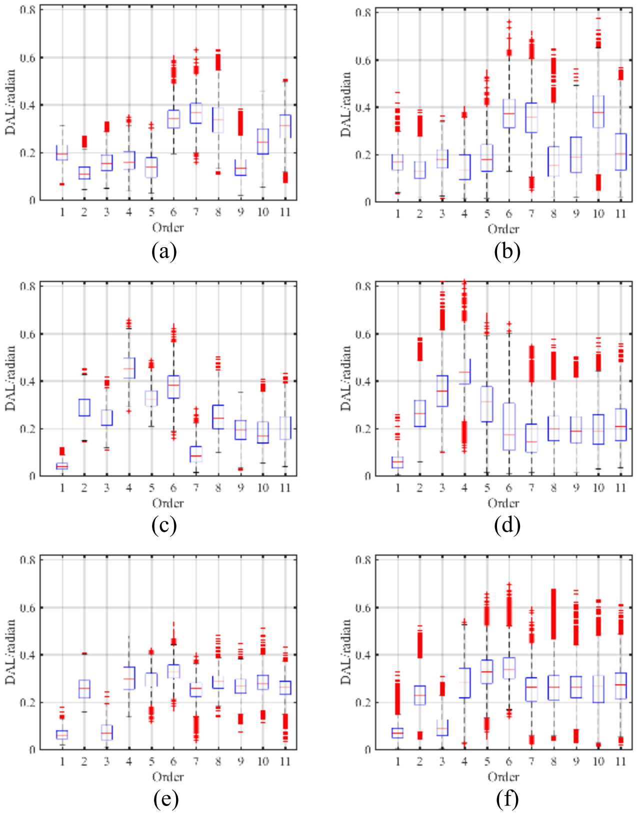
$$DAL_i = l_{i+1} - l_i \quad i = 1, 2, \ldots M(M < N) \tag{4}$$

where DAL_i is the $i-th$ order DAL feature, l_{i+1} is the $(i+1)-th$ order LSP feature.M is the maximum order of DAL features and N is the maximum order of LSP features.

If θ_i and ω_i are close to each other, they will have strong resonance characteristics when ω approaches these frequencies. Correspondingly, the peaks of the speech signal spectral envelope appear at these frequencies. Moreover, LSP features can reflect positions of formants, and DAL features can reflect intensity of formants. The closer LSP distribution is, the smaller DAL is, and the sharper the spectrum peak is. Therefore, different vowels can be distinguished by LSP and DAL features. Fig.2 depicts the boxplot distributions of 11-order DAL features for normal and pathological vowels /a/, /i/, /u/.

Figure 2 shows that for normal vowels/a/, /i/, /u/, the rectangular boxes of the first seven orders of DAL features are quite different. From the variation tendency of medians expressed as the red lines in rectangular boxes, distribution of the minimum value of the first seven orders of DAL for each vowel signal is different. Accordingly, the corresponding formants are also different, so the first seven orders of DAL features have better discrimination capabilities for the three vowels /a/, /i/ and /u/. Besides, the last four orders of DAL features of pathological vowel /a/ are completely different from those of normal vowel /a/, while the last four orders of DAL features of pathological vowel /i/ and pathological vowel /u/ have many overlapping parts, which make the discrimination capability worse. Considering low-order DAL features correspond to low frequency parts of vowel signals, the above statistical results show that the pathological changes of vocal cords have greater effect on high frequency parts of vowel signals than that on low frequency parts.

Generally speaking, the effect of laryngeal diseases on vocal cord vibrations is that noise first appears at high frequency parts of the spectrum. With the aggravation of disease, noise begins to appear gradually at lower frequency parts. Considering human ear is an auditory organ that can really distinguish different vowels, it is more reasonable to perform frequency warping based on human auditory perception.



(a) (b) (c) (d) (e) (f)

FIGURE 2. Comparison of DAL features of six vowel signals. (a) Normal vowel /a/. (b) Pathological vowel /a/. (c) Normal vowel /i/. (d) Pathological vowel /i/. (e) Normal vowel /u/. (f) Pathological vowel /u/.

It has been studied that frequency discrimination of human ears to speech signals mainly depends on the role of cochlear basement membrane, while the vibration of different Bark frequency bands stimulates different positions of the basement membrane. In the paper, the Bark transform scale is used to perform non-linear frequency warping for extracted LSP features considering that the discrimination of DAL features in low frequency parts is higher than that in high frequency parts, and Bark domain tends to truly reflect the sense of human ears to speech signals.

Distributions of the 12-order LSP features before and after frequency warping are shown in Fig. 3. Compared with Fig.3(a), Fig.3 (b) enlarges the low frequency parts of the

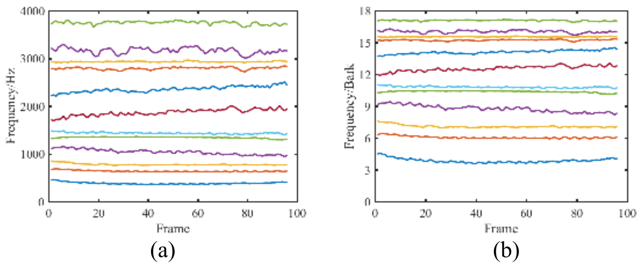


FIGURE 3. **Distributions of 12-order LSP and BLSP features. (a) Distribution of 12-order LSP feature. (b) Distribution of 12-order BLSP feature.**

signal, compresses the high frequency parts, and improves the discrimination between normal and pathological multi-vowels. Many scholars have proposed various approximate functions to express the Bark domain. The warping function of Bark domain used in this paper are as follows:

$$Bark = 26.81/\left(1 + (1960/f)\right) - 0.53 \tag{5}$$

C. ENHANCEMENT

From the analysis of Figure 2, for pathological vowels /a/, /i/and /u/, because the irregular vibration of vocal cords caused by pathological changes of vocal cords makes the voice signal infiltrate into noise components, distribution of the first seven orders of DAL features is more uniform than that of the normal vowels. To solve the problem of poor discrimination performance of BLSP features because of the formant shifting or even lost caused by the uniform distribution of DAL features of pathological voice, E-BLSP features with the retaining inter-dimensional information are calculated by enhancing and adjusting BLSP features using DAL in this paper. In this paper a method of bidirectional iterative is applied to adjust the j-th ($j = 2, \ldots, N-1$) order BLSP feature. After adjustment, the original BLSP features are directly updated to adjust the next order BLSP features. $\{b_1, b_2, \ldots b_N\}^N$ denote BLSP features of current frame and N denotes the total orders of BLSP features. The concrete iteration formulas can be expressed as follows:

$$b_i^{'} = b_{i-1} + c_{i-1} + \frac{c_{i-1}^2}{c_{i-1}^2 + c_i^2}\left[(b_{i+1} - b_{i-1}) - (c_i + c_{i-1})\right] \tag{6}$$

$$c_i = \eta\left(b_{i+1} - b_i\right), \eta < 1, i = 2, 3, \ldots, N-1 \tag{7}$$

1) FORWARD ITERATION

Adjusting the j-th order BLSP feature forward from $j = 2$ to $j = N - 1$;

2) BACKWARD ITERATION

Adjusting the j-th order BLSP feature backward from $j = N - 1$ to $j = 2$;

3) AVERAGING

Obtaining E-BLSP features by averaging the BLSP features of step 1) and step 2).

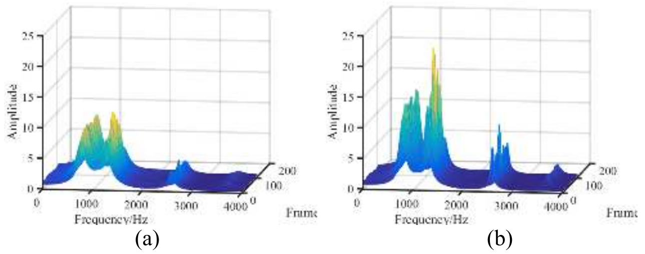


FIGURE 4. **Three-dimensional spectrum of BLSP and E-BLSP. (a) Three-dimensional spectrum of BLSP. (b) Three-dimensional spectrum of E-BLSP.**

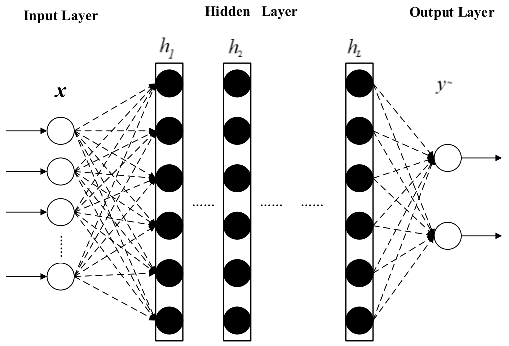


FIGURE 5. **Structure of the DNN network.**

In formula (7), η is the degree of formant enhancement, and the smaller η is, the more obvious the enhancement effect is. The value of η needs to be set reasonably in experiment. Fig. 4 shows three-dimensional spectrum of BLSP and E-BLSP of just one frame of the pathological voice signal. As shown in Fig. 4(b), the amplitude corresponding to the formant frequency is greatly increased, and the broadening effect is well suppressed, which greatly enhances the discrimination capability between normal and pathological multi-vowels.

D. CLASSIFICATION

1) DNN [27]

DNN model contains an input layer, an output layer, and multiple hidden layers, which aims to form a complex mapping function between the input and output vectors. It has been proved that DNN model can provide a satisfactory performance in speech recognition applications [28], [29]. As shown in Fig. 5, in DNN, the relationship between the input vector and the output vector of the first hidden layer can be expressed as follows:

$$h_1 = f\left(W_1 x\right) + b_1 \tag{8}$$

where W_1 and b_1 indicate the weight matrix and the bias vector, respectively. x indicates the input vector, h_1 indicates the output vector of the first hidden layer, and f indicates activation function. In this paper, the ReLU function with simple gradient calculation and no gradient disappearance is selected as activation function. In DNN, the relationship between current hidden layer and next hidden layer can be

TABLE 2. Statistics of normal and pathological samples.

	Normal /a/	Normal /i/	Normal /u/	Polyp /a/	Polyp /i/	Polyp /u/
Male: Female	93:87	93:87	93:87	104:76	104:76	104:76

written as:

$$h_{i+1} = f\left(W_{i+1}h_i + b_{i+1}\right), \quad i = 1, 2, \ldots L-1 \qquad (9)$$

where L is the number of hidden layer. In the last layer of recognition model, another function, $g(\cdot)$, is adopted on the output layer to get the output vector $y^{\sim}$.The calculation formula is shown as follows.

$$\tilde{y} = g\left(h_L\right) \qquad (10)$$

where h_L indicates the last hidden layer. For recognition tasks, Softmax function is used as $g(\cdot)$ to transform the output of neural network into a probability distribution, so as to optimize the recognition results. Finally, we can get the predicted labels of testing samples by the trained network and obtain the recognition rate by comparing the corresponding preset labels of testing samples.

2) SVM [30]

An SVM is one example of a classifier, which estimates decision surfaces directly rather than modeling a probability distribution across the training data. SVM has demonstrated good performance on several speech recognition problems [31], [32]. Practical recognition problems typically involve data that can only be separated using a nonlinear decision surface. In this case, the optimization of input data involves the use of kernel-based transformation. That is, the nonlinear mapping carried out by SVM can map the input data from low-dimensional space to high-dimensional space by kernel function. As for kernel function, Radial Basis Function (RBF) often delivers better performance because of good adaptability. Therefore, RBF is applied in this paper, with the definition as follows:

$$k\left(\|x - x_c\|\right) = \exp\left\{-\|x - x_c\| \wedge 2/\left(2 * \sigma\right) \wedge 2\right\} \qquad (11)$$

where x_c is the center of kernel function, σ is the width of kernel function.

In the classification problems, SVM can follow a procedure to find the separating hyperplane with the largest margin. There is a concern here, the classification performance of SVM depends on the solutions of the aperture of the kernel γ and a penalty parameter c [33], which can avoid the over-fitting and under-fitting. In this study, SVM based on RBF kernel function is used to maximize prediction accuracy with the optimal parameters c and γ obtained from a grid search method. Specifically, first of all, we get the input vectors of subject data sets and the corresponding class labels. Then, these input vectors are divided into K parts, one of which is reserved as the test set, and the remaining parts are considered as the training sets. And then recognition model can be well trained by the training sets. Finally, the trained model is used to predict labels of test sets. Thus, we can calculate the predictive accuracy by comparing the predictive labels of test sets obtained from the trained network and the preset labels of test sets. The predictive accuracy is defined as the percentage of subjects that are correctly classified.

III. EXPERIMENTS AND RESULTS

A. CORPUS OF SPEAKERS

Experiments in this paper are conducted on SVD voice disorders database established by the Institute of Phonetics at Saarland University, which is available for free download [34]. This database records separately sustained normal and different kinds of pathological vowels /a/, /i/, and /u/ under the case of different intonations (e.g. normal, low, high, low-high-low). All voices recorded in SVD database are sampled at 50 kHz with 16-bit resolution. In the experiment, polyp and normal multi-vowels recorded by different male and female speakers aged between 15 and 78 years are selected for experiments. Moreover, the multi-vowels /a/, /i/, and /u/ with different intonations mentioned above are resampled to 16 kHz for experiment. Table 2 shows the details related to the subjects who performed the experimental task.

B. EXPERIMENTAL SETUP

Prior to classification, firstly, six kinds of vowels (normal /a/, normal /i/, normal/u/, polyp /a/, polyp /i/, polyp /u/) are preprocessed. In this case, these speech signals are framed and windowed using 25ms Hamming window with a 50% frame shift. Then, the 12-order LSP feature based on AR model, the 12-order LPCC feature [10], the 24-order MFCC features (12 MFCC and 12 first-order delta) [9], 36-order RASTA-PLP features (12 RASTA-PLP cepstral feature, 12 first-order delta and 12 second-order delta) [11] and 15-order AMS feature [12] are calculated frame by frame. After that, LSP feature is warped at Bark frequency to obtain 12-order BLSP. Finally, the 11-order DAL features are used to iterate forward and backward for BLSP to enhance the formants. The enhancement coefficient is 0.4 in the experiment. Thus, for each class of the vowel samples, the feature vectors of different dimensions corresponding to different features mentioned above can be obtained to be used as the input vectors of classification network. There is a concern here, training data including feature vectors and corresponding class labels is given, so our experiment is supervised case.

TABLE 3. Classification Accuracies and AUCS under seven parameter sets using DNN network.

Parameter set	Polyp/a/	Polyp/i/	Polyp/u/	Normal/a/	Normal/i/	Normal/u/	The average accuracy	AUC
E-BLSP	97.2029	**97.6601**	96.8160	98.1048	96.8132	97.6367	97.3600	0.9894
LPCC	81.4887	82.8793	65.1372	72.8815	77.7505	65.1319	75.5444	0.8781
MFCC	96.7662	95.7857	94.8637	98.6997	97.4208	97.8368	96.7238	0.9734
LPCC+E-BLSP	97.5565	**97.8183**	97.2444	97.6597	98.6321	97.2845	97.7407	0.9908
MFCC+E-BLSP	97.8165	**98.4855**	95.8100	99.1983	98.0402	98.8844	97.9333	0.9915
MFCC+LPCC	95.8537	97.0944	94.0874	97.8552	98.7715	99.0741	96.8571	0.9745
E-BLSP +LPCC+MFCC	98.6688	**98.7836**	98.2207	98.3789	98.9846	98.9451	**98.6190**	**0.9925**

In the case of classification, 75% of feature vectors of each class of vowel are randomly selected as the training set and thus the rest as the testing set. This ensures that each class of the vowel samples is averagely distributed during the training and testing stages of classification network. In terms of SVM network, while searching the optimal penalty parameter and kernel parameter using K-fold cross validation and grid search methods, the value of K is 5. In terms of DNN network, for input layer of DNN network, we need to set the dimension information of input vectors and the total number of training samples according to different classification features. In addition, we also set the class labels of vowel samples, the weight matrix and the bias vector. For hidden layers of DNN network, we also evaluate the performance among different DNN structures (i.e. hidden layers and number of neurons), and the results show that the best performance is achieved when using 2 hidden layers with 100 neurons in each layer. For our experiments, first, E-BLSP, LPCC and MFCC features are input into SVM and DNN networks. And then we study the classification performances of seven parameter sets (three single features: E-BLSP LPCC, MFCC; four feature combinations: E-BLSP and LPCC, E-BLSP and MFCC, LPCC and MFCC, combination of E-BLSP, LPCC and MFCC). Finally, we compare the classification performances of the method based on the parameter set with high accuracy with other methods based on AMS and RASTA-PLP features which are often used in the field of normal speech recognition.

C. RESULTS

The experimental results for pathology classification are represented in different terms. These terms mainly include accuracy (i.e. the ratio between correctly detected samples and the total number of samples) and the area under Receiver Operating Characteristic (ROC) curve, called AUC. The ROC curves are used to represent graphically the performance of the proposed architecture. In addition, the AUC represents an estimation of the expected performance of system in a single scalar. The bigger the value of AUC is, the better the performance of classifier is. In order to ensure the accuracy and statistical significance of the experiments, each group of the experiments is performed five times and the average values of evaluation terms are taken as the final results.

Table 3 and Table 4 show the accuracy of each class, the average accuracy and AUC obtained independently for each parameter set using DNN network and SVM network, respectively. Table 5 shows the comparison of the proposed method in this paper and other methods based on AMS [12] and RASTA-PLP [11]. Fig. 6 (a) and Fig.6 (b) plot the ROC curves of all the methods mentioned in this paper under DNN and SVM networks, respectively.

As we can see from Table 3 and Table 4, the accuracy of each class, the average accuracy, and AUC vary from DNN network to SVM network for the same parameter set used in this paper. In case of each parameter set, the recognition performance of three normal vowels is higher than that of the corresponding pathological vowels, while the accuracies of vowels /a/ and /i/ are higher than that of vowel /u/ for both normal and pathological multi-vowels. We can also notice that the combination of parameter sets have higher accuracies and larger AUCs than the other single parameter sets. Moreover, the parameter set of three features combined together including E-BLSP, LPCC, MFCC features have the best classification performance in this paper. The highest average accuracies for DNN and SVM networks can reach 98.6190%

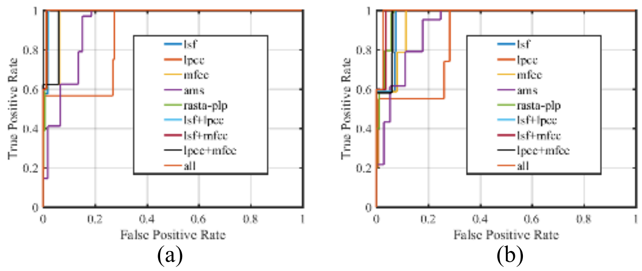
TABLE 4. Classification accuracies and AUCS under seven parameter sets using SVM network.

Parameter set	Polyp/a/	Polyp/i/	Polyp/u/	Normal/a/	Normal/i/	Normal/u/	The average accuracy	AUC
E-BLSP	93.9019	**94.3298**	89.1019	95.3462	94.3722	90.9207	92.9955	0.9761
LPCC	72.8947	79.2405	63.7037	74.9319	81.8182	68.6610	73.4222	0.8715
MFCC	92.1512	91.3295	82.5858	92.0228	91.6667	89.1566	89.7143	0.9568
LPCC+E-BLSP	92.1875	**95.2128**	91.3747	94.4591	95.4545	91.5301	93.3778	0.9720
MFCC+E-BLSP	97.7273	**97.9769**	90.7821	98.5673	97.1671	93.2749	95.9048	0.9822
MFCC+LPCC	88.6792	92.8375	83.8983	93.9577	97.5904	85.6734	90.3333	0.9714
E-BLSP +LPCC+MFCC	95.9569	**97.2452**	93.5574	98.8701	98.8701	93.3518	**96.2693**	**0.9868**

TABLE 5. Comparison of the proposed method in this paper and other methods based on AMS [12] and RASTA-PLP [11].

Network	Parameter set	Polyp/a/	Polyp/i/	Polyp/u/	Normal/a/	Normal/i/	Normal/u/	The average accuracy	AUC
DNN	E-BLSP +LPCC+ MFCC	98.6688	98.7836	98.2207	98.3789	98.9846	98.9451	**98.6190**	**0.9925**
	AMS [12]	81.3333	79.4667	64.2667	82.4000	92.2667	78.1333	79.6444	0.9227
	RASTA-PLP [11]	96.8000	92.8000	89.3333	96.8000	97.0667	93.3333	94.3556	0.9882
SVM	E-BLSP +LPCC+ MFCC	95.9569	97.2452	93.5574	98.8701	98.8701	93.3518	**96.2693**	**0.9868**
	AMS [12]	62.1333	68.2667	67.2000	76.0000	78.4000	78.1333	71.6889	0.9208
	RASTA-PLP [11]	92.0000	89.0667	87.4667	93.0667	90.6667	92.8000	90.8444	0.9761

and 96.2693%, while the largest values of AUCs can reach 0.9925 and 0.9868, respectively. From Table 3, in the case of using single features, it is clear that E-BLSP feature proposed in this paper is the best, and LPCC feature is the worst. The highest accuracies for E-BLSP, LPCC and MFCC are 97.3600%, 75.5444% and 96.7238%, and the largest values of AUCs are 0.9894, 0.8781 and 0.9734, respectively. Generally, the recognition performance of traditional features such as the MFCC feature for pathological vowel /a/ is higher than that of other pathological vowels. However, the recognition performance of E-BLSP feature for polyp vowel /i/ is higher than that of polyp vowel /a/. In case of combinations of two features, compared with the single feature cases, the recognition performance has been improved due to the introduction of E-BLSP feature. Besides, the best acquired accuracies for LPCC + E-BLSP, MFCC + E-BLSP and MFCC + LPCC are 97.7407%, 97.9333% and 96.8571%, and the largest values of AUCs are 0.9908, 0.9915 and 0.9745, respectively. Table 4 shows that SVM network does not provide better discrimination capability than DNN network for the same parameter set. Moreover, in all cases the best classification accuracies and AUCs are achieved using



(a) (b)

FIGURE 6. The ROC curves of all the methods mentioned in this paper under DNN and SVM networks. (a) The ROC curves of all the methods mentioned in this paper under DNN network. (b) The ROC curves of all the methods mentioned in this paper under SVM network.

combination of three features including E-BLSP, LPCC and MFCC. We can infer from obtained accuracies and AUCs mentioned in Table 4 that the contribution of individual feature is less compared with the combination of features in all cases. Again, similar to the results of DNN network, E-BLSP feature has more contributions in the classification of different normal and pathological multi-vowels than the other two features.

Table 5 shows the comparison of the proposed method in this paper and other methods based on AMS [12] and RASTA-PLP [11] commonly used for normal speech recognition. As we can see from Table 5, the accuracy and AUC of the proposed method in this paper are higher than those of the two methods based on AMS and RASTA-PLP under two different networks. Besides, for the same method, DNN network shows the better classification performance than SVM. Thus, the combination of three features including E-BLSP has better classification performance than AMS and RASTA-PLP features for normal and pathological multi-vowels recognition. In the case of DNN network, the performance of the method based on AMS feature is the worst. And compared with the method based on RASTA-PLP feature, the method in this paper has less computational complexity and higher classification accuracy.

Fig.6 plots ROC curves of all the methods mentioned in this paper under DNN and SVM networks. In all cases, combining the three features including E-BLSP, LPCC and MFCC using DNN network, the performance is the best with respect to the performance obtained using other parameter sets or other methods based on AMS and RASTA-PLP features. Due to this fact, combination parameter set including E-BLSP feature is preferred for further analysis combining classifiers. Besides, there is a clear improvement in the performance of classification system introducing E-BLSP feature. From the nine ROC curves mentioned above, it is clear that combination of three features including E-BLSP using DNN network serves as a discriminant between normal and pathological multi-vowels.

IV. CONCLUSION

In this study, we introduce a new feature called E-BLSP, evaluate the classification performance of three features (E-BLSP, LPCC, MFCC) on DNN and SVM networks, and then proposed an effective method based on combination of three features including E-BLSP for the classification of normal and pathological multi-vowels. For the three features (E-BLSP, LPCC, MFCC) studied in this paper, the obtained accuracies vary from different features in the same classification, and also vary from one classification network to another with the same feature. The best acquired accuracies for DNN and SVM network are 98.6190% and 96.2693%, respectively, while the largest values of AUCs can reach 0.9925 and 0.9868, respectively, using the three features together including E-BLSP, LPCC and MFCC. In addition, the classification performance of the method in this paper is better than that of other two methods based on AMS and RASTTA-PLP features by comparison.

The method based on the combination of three features including E-BLSP greatly improves the accuracies of normal vowels /a/, /i/, /u/ and pathological vowels /a/, /i/, /u/. Especially, the recognition performance of this method for pathological vowel /i/ is better than that of the other traditional features for pathological vowel /a/, which provides a new idea for pathological voice recognition research and also for voice repair of various vowels and more complex words even sentences.

Regarding the future work, we will explore the classification of pathological voice based on more complex vowels and study other more efficient classification features and recognition algorithms. Besides, experiments in this paper are only conducted on SVD voice disorders database. More experiments with other databases will be carried out to supplement this aspect in the future.

REFERENCES

[1] N. Roy, R. M. Merrill, S. Thibeault, R. A. Parsa, S. D. Gray, and E. M. Smith, "Prevalence of voice disorders in teachers and the general population," *J. Speech, Lang., Hearing Res.*, vol. 47, no. 2, pp. 93–281, Apr. 2004.

[2] K. H. Malki, "Voice disorders among Saudi teachers in Riyadh city," *Saudi J. Oto-Rhinolaryngology Head Neck Surg.*, 2010.

[3] *National Institute on Deafness and Other Communication Disorders: Voice, Speech, and Language: Quick Statistics*. Accessed: Oct. 2018. [Online]. Available: http://www.nidcd.nih.gov/health/statistics/vsl/Pages/stats.aspx

[4] (Oct. 2018). *Research Chair of Voicing and Swallowing Disorders*. [Online]. Available: http://c.ksu.edu.sa/vas/en/vsb

[5] Cancer Research U.K. Accessed: Oct. 2018. [Online]. Available: http://cancerresearchuk.org

[6] H. R. Sharifzadeh, I. V. Mcloughlin, and F. Ahmadi, *Speech Rehabilitation Methods for Larynegctomised Patients*. Amsterdam, The Netherlands, Springer, 2010, pp. 597–607.

[7] R. Rong, *Research on Speech Recognition Based on Artificial Neural Network*. Shandong Normal Univ., Shandong Sheng, China, 2005.

[8] X. Song, *Vowel Recognition Based on K-L Transform and Standardization of Formant Parameters*. Shanghai Normal Univ., Shanghai Shi, China, 2010.

[9] S. Sharma and P. K. Das, "Reduced feature sets for vowel recognition," in *Proc. 8th Int. Conf. Elect. Comput. Eng.*, Dec. 2014, pp. 116–119.

[10] A. H. Unnibhavi and D. S. Jangamshetti, "LPC based speech recognition for Kannada vowels," in *Proc. Int. Conf. Elect., Electron., Commun., Comput., Optim. Techn. (ICEECCOT)*, Mysuru, India, Dec. 2017, pp. 1–4.

[11] T. A. Mesallam *et al.*, "Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms," *J. Healthcare Eng.*, vol. 2017, Oct. 2017, Art. no. 8783751.

[12] N. Moritz, J. Anemüller, and B. Kollmeier, "Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2011, pp. 5492–5495.

[13] S. M. Al-Qaraawi and S. S. Mahmood, "Wavelet transform based features vector extraction in isolated words speech recognition system," in *Proc. 9th Int. Symp. Commun. Syst., Netw. Digit. Sign*, Jul. 2014, pp. 847–850.

[14] M. D. O. Rosa, J. C. Pereira, and M. Grellet, "Adaptive estimation of residue signal for voice pathology diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 1, pp. 96–104, Jan. 2000.

[15] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, "Discrimination of pathological voices using a time-frequency approach," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 3, pp. 421–430, Mar. 2005.

[16] M. K. Arjmandi, M. Pooyan, M. Mikaili, M. Vali, and A. Moqarehzadeh, "Identification of voice disorders using long-time features and support vector machine with different feature reduction methods," *J. Voice*, vol. 25, no. 6, pp. e275–e289, Nov. 2011.

[17] G. Muhammad and M. Melhem, "Pathological voice detection and binary classification using MPEG-7 audio features," *Biomed. Signal Process. Control*, vol. 11, pp. 1–9, May 2014.

[18] G. Muhammad, T. A. Mesallam, K. H. Malki, M. Farahat, A. Mahmood, and M. Alsulaiman, "Multidirectional regression (MDR)-based features for automatic voice disorder detection," *J. Voice*, vol. 26, no. 6, pp. 817.e19–817.e27, Nov. 2012.

[19] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and Mel-Cepstral coefficients," *IEEE J. Trans. Bio-Med. Eng.*, vol. 58, no. 2, pp. 370–379, Feb. 2011.

[20] A. Al Nasheri *et al.*, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2017.

[21] S.-H. Fang *et al.*, "Detection of pathological voice using Cepstrum vectors: A deep learning approach," *J. Voice*, to be published.

[22] K. Ezzine and M. Frikha, "Investigation of glottal flow parameters for voice pathology detection on SVD and MEEI databases," in *Proc. 4th Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, Sousse, Tunisia, Mar. 2018, pp. 1–6.

[23] A. Castellana, A. Carullo, S. Corbellini, and A. Astolfi, "Discriminating pathological voice from healthy voice using cepstral peak prominence smoothed distribution in sustained vowel," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 3, pp. 646–654, Mar. 2018.

[24] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, no. S1, p. S35, Apr. 1975.

[25] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, "Efficient encoding of mel-generalized cepstrum for CELP coders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1997, pp. 1355–1358.

[26] K. Tokuda, T. Kobayashi, S. Imai, and T. Chiba, "Spectral estimation of speech by mel-generalized cepstral analysis," *Electron. Commun. Jpn.*, vol. 76, no. 2, pp. 30–43, Jul. 1993.

[27] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[28] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[29] T. N. Sainath *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 965–979, May 2017.

[30] C.-C. Chang and C.-J. Lin, "A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, Apr. 2011, Art. no. 27.

[31] R. Solera-Urena *et al.*, "Real-time robust automatic speech recognition using compact support vector machines," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1347–1361, May 2012.

[32] S. Cumani and P. Laface, "Large-scale training of pairwise support vector machines for speaker recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 11, pp. 1590–1600, Nov. 2014.

[33] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[34] W. J. Barry and M. Pützer. Saarbrucken Voice Database. Institute of Phonetics, University of Saarland, Saarbrücken, Germany. Accessed: Oct. 2018. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/

TAO ZHANG received the M.S. degree from the School of Electronic Information Engineering, Tianjin University, Tianjin, China, in 2001, and the Ph.D. degree from Tianjin University, in 2004, where he is currently an Associate Professor with the Texas Instruments DSP Joint Lab, School of Electrical and Information Engineering. His current interests include adaptive signal processing, acoustic signal processing, auditory model, image, and video.

YAQIN WU received the B.S. degree from the School of Information Science and Technology, Dalian Maritime University, China, in 2017. She is currently pursuing the M.S. degree with the School of Electrical and Information Engineering, Tianjin University. Her research interest includes acoustic signal processing.

YANGYANG SHAO received the B.S. degree from the School of Electronic Information Engineering, Hebei University, Hebei, China, in 2018. She is currently pursuing the M.S. degree with the School of Electrical and Information Engineering, Tianjin University. Her research interest includes acoustic signal processing.

MINGYANG SHI received the B.S. degree from the School of Electronic Information Engineering, Tianjin Polytechnic University, Tianjin, China, in 2018. He is currently pursuing the M.S. degree with the School of Electrical and Information Engineering, Tianjin University. His research interest includes neuroimage.

YANZHANG GENG received the M.S. degree from the School of Mechanical and Power Engineering, North University of China, Taiyuan, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University. His research interest includes acoustic signal processing.

GANJUN LIU received the B.S. degree from the College of Electronic Information Engineering, Southwest University, Chongqing, China, in 2016. He is currently pursuing the M.S. degree with the School of Electrical and Information Engineering, Tianjin University. His research interest includes audio signal processing.