Biomedical Signal Processing and Control 70 (2021) 102973

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Deep connected attention (DCA) ResNet for robust voice pathology detection and classification

Huijun Ding [a], Zixiong Gu [a], Peng Dai [b], Zhou Zhou [c], Lu Wang [d], Xiaoxiao Wu [e,*]

[a] Guangdong Provincial Key Laboratory of Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China
[b] Huawei Technologies Inc., 19 Allstate Pkwy, Markham, ON, Canada L3R 5A4
[c] Department of Otolaryngology Head and Neck Surgery, Shenzhen People's Hospital, Shenzhen, 518020, China
[d] Department of Otolaryngology Head and Neck Surgery, Shenzhen University General Hospital, Shenzhen, 518055, China
[e] School of Electronic Information Engineering, Shenzhen University, Shenzhen, 518060, China

ARTICLE INFO

Keywords:
Voice pathology
Automatic detection
Convolutional neural network
Deep learning

ABSTRACT

The automatic diagnosis method based on speech signal analysis is able to realize the detection and classification of pathological voices. It plays an important role in the early diagnosis and auxiliary treatment of voice pathology, which effectively relief the discomfort of patients and reduce the workload of doctors. Therefore, the automatic diagnosis method based on speech signal analysis is of great research value. Meanwhile, high accuracy, high precision and stability are the pursuit goals. In this paper, a novel computer-aided assessment based on speech signal analysis for pathological voice classification (CS-PVC) system is proposed. This model focuses on the areas with large differences between different pathological voices and healthy voices, while ignore the negative impact of insignificant information on the performance of the model. Two databases were used in the experiments, one is the Saarbruecken Voice database (SVD), and the other is the self-built Shenzhen People's Hospital voice database (SZUPD). The pathological voice detection accuracy of the proposed system on the above two databases are 81.6% and 82.2% respectively. The experimental results show that the proposed framework is not data-dependence. In other words, it has the potential to be universally applicable in medical framework in the future.

1. Introduction

With the change of living habits and the increase of human communication, voice pathology has become a global health problem. The incidence of voice pathology is high, covering a wide range of ages. According to the literature [1,2], it is estimated that 17.9 million U.S. adults aged 18 or older (7.6% of the population) have voice problems in the past 12 months. Nearly 1 in 12 (7.7%) U.S. children aged 3–17 has had a disorder related to voice, speech, language, or swallowing in the past 12 months [3]. Voice pathology causes inconvenience in daily life, resulting in severe social problems. For example, voice pathology may lead to serious mental problems, depression and other related diseases [4].

Currently, voice pathology includes vocal cysts, vocal cord nodules, dysphonia, laryngitis, etc. It is usually accompanied by abnormalties in vocal cord closure, flexibility or symmetry, which cause the voice hoarseness, harshness and weakness [5]. Thus, the voice quality is significantly worse than normal people. From the technical point of view, the recorded audio signals are able to capture the above differences for pathological voice analysis. On the other hand, clinical pathological voice classification approaches are mainly divided into two broad categories, subjective evaluation and objective evaluation [6]. Subjective evaluation is mainly carried out by medical staff or other professionals for visual assessment and auditory–perceptual assessment. However, the penetrating electronic laryngoscope, which is used frequently in visual assessment [7–9], is prone to cause discomfort to the patient, including a sense of foreign invasion, pain, etc. In auditory–perceptual assessment, professionals perform comprehensive scoring by listening to a patient's specified voice and using the patient's pre-defined scoring criteria such as G(overall grade of hoarseness); R (roughness); B(breathiness); A (asthenic) and S (strained quality) of the voice (GRBAS) [10] and consensus auditory–perceptual evaluation of voice (CAPE-V) [11]. The scoring results usually vary from doctor to doctor [12]. In contrast, objective evaluation is able to provide objective quantitative evaluation indicators for doctors to refer to and

* Corresponding author.
E-mail addresses: hjding@szu.edu.cn (H. Ding), guzixiong2018@email.szu.edu.cn (Z. Gu), peng.dai.ca@ieee.org (P. Dai), michelle0304@126.com (Z. Zhou), 694051728@qq.com (L. Wang), xxwu.eesissi@szu.edu.cn (X. Wu).

https://doi.org/10.1016/j.bspc.2021.102973
Received 27 January 2021; Received in revised form 17 June 2021; Accepted 7 July 2021
Available online 5 August 2021

make effective and reliable judgments. It is essential in clinical practice. The computer-aided assessment based on speech signal analysis for pathological voice classification (CS-PVC) provides a convenient non-invasive objective diagnosis scheme, which effectively alleviate the discomfort of patients and reduce the difficulty of operation for doctors in diagnosis. It also utilizes the powerful computing power of computers to achieve rapid diagnosis and consequently reduce the workload of doctors. Therefore, CS-PVC system is worth studying.

CS-PVC system usually consists of two parts, i.e. feature extraction and classification [13–15]. Common clinically interpretable acoustic features include multidimensional voice program (MDVP) [16,17], parameters based on wavelet transform (WT) [18], mel-frequency cepstral coefficients (MFCC) [19] and linear prediction cepstrum coefficient (LPCC) [20]. With the extracted features, a classifier is usually applied to predict the type of voice pathology. Many classifiers have been used for pathological voice detection such as Gaussian mixture model (GMM) [21,21], hidden Markov model (HMM) [22], support vector machines (SVM) [17,23], random forests (RF) [24]. These methods are usually used for small data sets, and thus have a strong dependence on the data set. Their robustness and generalization are poor [25]. With the fast development of machine learning, deep learning methods have been successfully applied in many areas, e.g. speech recognition [26,27], image segmentation [28,29] and object detection [30], etc. It has also been applied in pathological voice detection and has achieved good performance [14]. In addition,several large scale voice pathology databases have been released for relevant research, e.g. Saarbruecken Voice database (SVD) [dataset] [31], Massachusetts Eye and Ear Infirmary database (MEEI) [dataset] [32] and Arabic Voice Pathology database (AVPD) [33].

In this paper, a novel CS-PVC system is proposed, which can automatically detect pathological voices. Fig. 1 shows the diagram of the proposed system. Firstly, the feature extraction is applied. Log mel-frequency spectral coefficients (MFSC) together with its first-order and second-order derivatives are used as acoustic features. The extracted features are then used to train the proposed network. A novel Residual structure-based Deep connected attention model (DCA-ResNet) is proposed. The novel connected attention mechanism enables the model to focus on the differences between different pathological and healthy voices, while ignoring the impact of insignificant information on the model performance. The healthy voice and the pathological voice will be predicted first. Following that, the pathological voice will be further classifier to obtain laryngitis, rekurrensparse, dysphonia and hyperfunktionelle dysphonia. Two datasets are utilized for verification tests, i.e. SVD database and the self-built SZUPD database.

The rest of this paper is arranged as follows. Section 2 will briefly review the related work. Detailed introduction of the proposed system will be given in Section 3. The details of the experiment will be given in Section 4. Section 5 will discuss the experimental results, followed by the conclusion in Section 6.

2. Related work

Pathological voice detection can be formulated as a classification problem, which utilizes acoustic signals as input. Acoustic signals are usually transformed into different kinds of feature embeddings for easier processing. Pathologically, various disease cause functional changes to the larynx, which leads to a wide range of measurable changes to the acoustic signal, such as jitter, shimmer, pitch, etc [37]. Related research in recent years is listed in Table 1. Hemmerling et al. made use of heuristic statistical metrics to measure the above-mentioned changes and obtained very promising results [24]. Then, a complete and convenient voice analysis software MDVP appeared, which not only integrated the above features, but also expanded it [16]. Thirty-three different long-term acoustic parameters with their definitions in MDVP are listed in Arjmandi et al. [38]. Al-nasheri et al. also achieved good results using the multi-dimensional voice features extracted by MDVP [17]. Although hand crafted features lead to decent results, its performance is inherently limited by the quality of the designed features. Motivated by the success of cepstrum in speech signal processing, a number of research proposed to use LPCC [20,39] and MFCC [19,21,25,34] as features. In addition, there is also a lot of studies based on spectrum analysis [35,36], which reduces the steps of discrete cosine transform (DCT) compared with cepstrum. After DCT, the energy is concentrated in the low frequency part. However, there are many high-frequency components in pathological voices that are beneficial to detection, so the time spectrum is more suitable for pathological voice analysis than cepstrum [40,41] (see Table 1).

In 2017, Ail et al. combined MFCC with GMM. However, it has to be noted that the inter-database results are significantly worse than the intra-database results, which indicates poor generalization issue [21]. In [24], Hermmerling et al. proposed a multistep approach. Separate RF models are trained for different gender groups. Hammami et al. adopted feature based on discrete wavelet transformation (DWT) and coupled with SVM and finally achieved 93.10% accuracy [23]. Harar et al. proposed to use XGboost with MFCC features. The models are trained on multiple datasets to improve robustness [25].

With the recent advances in deep learning, it has managed to achieve better-than-human performance on a number of tasks. Deep learning is particularly good at handling classification problems. It can be naturally extended to pathological voice detection. Fang et al. used MEEI to analyze the vowel /a/. Their experiments proved the advantage of deep learning methods over the classical methods, e.g. GMM, SVM, etc [34]. In [35,36], different architectures of convolutional neural network (CNN) are studied and compared on the SVD dataset. Motivated by the previous success of deep CNN, we propose to improve the system by CS-PVC. In addition, the system has also innovatively proposed an attention mechanism. The attention mechanism can effectively extract key information while ignoring irrelevant information [42–44]. The attention mechanism has been widely used in classification tasks, and in these tasks both have achieved satisfactory results [45,46].

3. Method

The proposed CS-PVC system is described in two parts, namely the feature extraction and the network structure.

3.1. Feature extraction

Fig. 2(a) and (b) shows the log mel-frequency spectrogram and MFCC of pathological voice and healthy voice for the vowel /a/. It clearly shows that the frequency change of pathological voice is more unstable than that of healthy voice. Therefore, log mel-frequency spectrogram is used as the input feature of the classification network in order to express the difference between healthy and pathological voices. In addition, because of the DCT used by MFCC, the features are decorrelated and compressed, and the main information is concentrated in the first few vectors. MFSC is smoother in time and frequency domain than MFCC. The MFSC feature make it easier for CNN to discover linear relationships as well as high-order causes of the input data, resulting in a better overall system performance [40]. Therefore, MFSC is used as the input feature of the network in this study. Fig. 3 shows the process of feature extraction. The length of the voice data is between 0 and 3 s. In order to ensure that the voice data is as complete as possible, the amount of calculation is kept small, and CNN needs to have a unified input size. The voice signal is first reframed into 1s long. When the signal is longer than 1s, it needs to be truncated to 1s. When the signal is shorter than 1s, it needs to be filled to 1s by reflection padding. Next, the signal is divided into 40 ms per frame by the Hann window, and there is a 50% overlap between each frame. The time domain signal is converted into a time–frequency domain signal by short-time Fourier transform (STFT), and the spectral coefficients are then sent to the

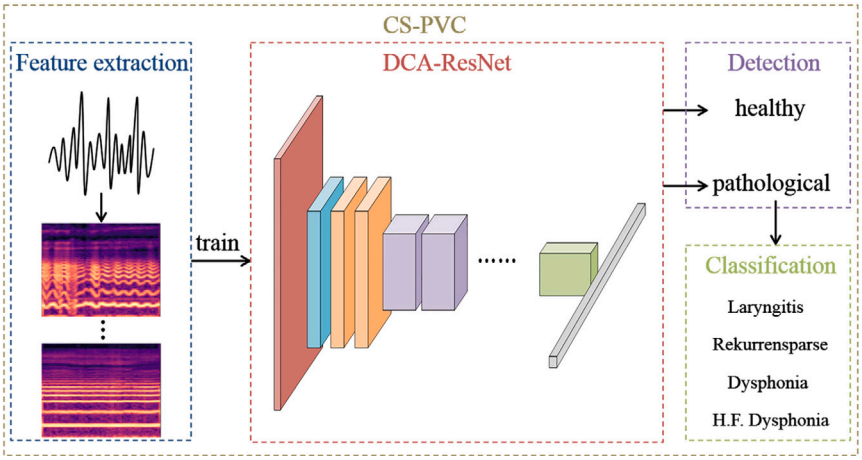

Fig. 1. The diagram of the proposed CS-PVC system.

Table 1
Comparison of related work on pathological speech detection in terms of features, classifiers, accuracy and database.

	First author	Feature	Classifier	Database	Accuracy (%)
Classical Methods	Hemmerling [24]	28-parameters	RF	SVD	SVD:100
	Ali [21]	MFCC	GMM	MEEI, SVD, AVPD	MEEI:94.60/SVD:80.20/AVPD:83.65
	Al-nasheri [17]	MVPD	SVM	MEEI, reSVD, AVPD	MEEI:88.21/SVD:99.68/AVPD:72.53
	Hammami [23]	DWT-feature	SVM	SVD	SVD:93.10
	Harar [25]	MFCC	XGBoost	AVPD, MEEI, PDA, SVD	AVPD+MEEI+PDA+SVD:73.30
Deep Learning	Fang [34]	MFCC	DNN	MEEI	MEEI:99.32
	Wu [35]	Spectrogram	CNN	SVD	SVD:71.00
	Muhammad [36]	Spectrogram	CNN	SVD	SVD:93.50

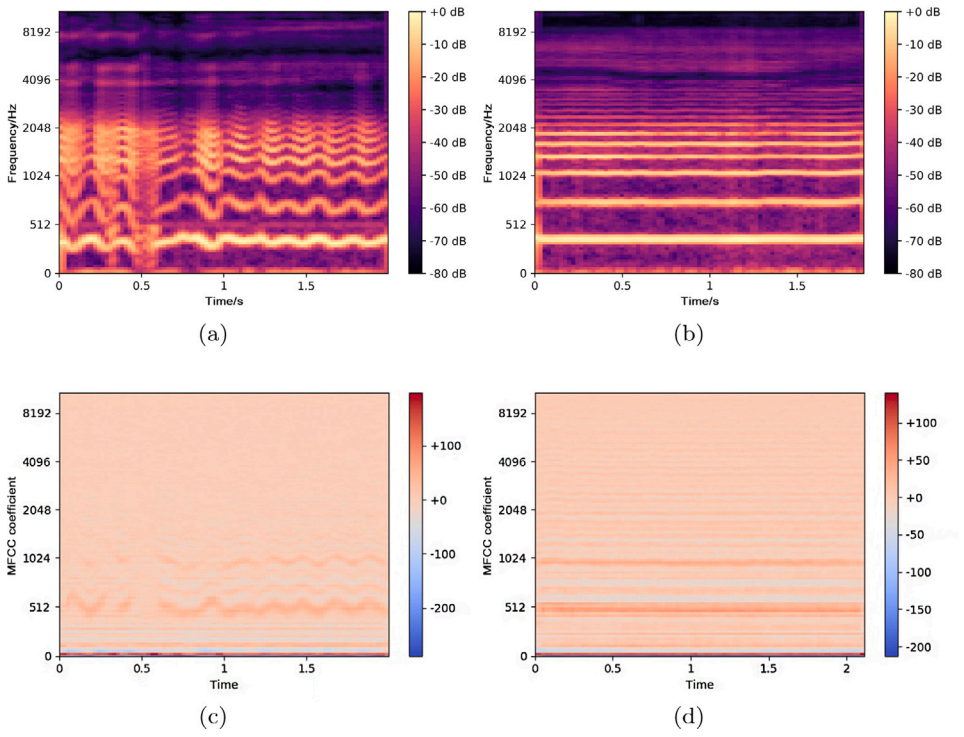

Fig. 2. (a) Log mel-frequency spectrogram of a pathological voice sample for the vowel /a/. (b) Log mel-frequency spectrogram of a healthy voice sample for the vowel /a/. (c) MFCC of a pathological voice sample for the vowel /a/. (d) MFCC of a healthy voice sample for the vowel /a/.

Mel filter bank. The reason for this process is that the frequency scale of the filter bank conforms to the characteristics of human hearing perception. The frequency relationship can be approximated by the

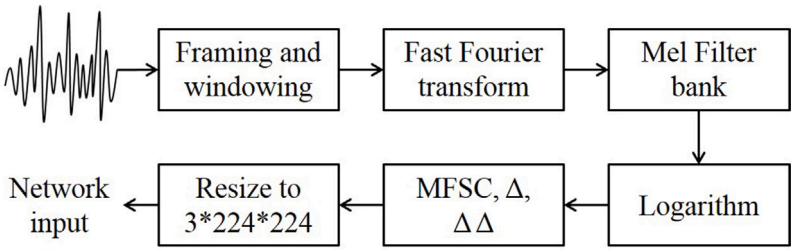

Fig. 3. The flow chart of feature extraction.

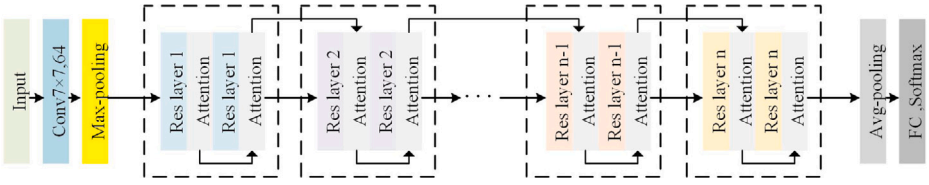

Fig. 4. An overview of DCA-ResNet.

following formula:

$$mel(f) = 2595 \lg(1 + f/700) \tag{1}$$

where f represents the actual frequency of the voice signal, and the unit is Hz. The mel filter bank is represented by the following formula:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ k - f(m-1)/f(m) - f(m-1), & f(m-1) \le k \le f(m) \\ f(m+1) - k/f(m+1) - f(m), & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{2}$$

where the center frequency of mth filter is $f(m)$, and the response at the center frequency is $H_m(k)$. The filter frequency linearly decreases toward 0 on both sides until it reaches the center frequency of two adjacent filters and the interval between $f(m)$ widens as the value of m increases.

The logarithmic operation is applied on mel-spectrogram to get the log mel-spectrogram, and its first-order time derivatives and second-order time derivatives can also be obtained accordingly. The above features express the dynamic relationship of the voice. The first-order time derivatives, d_t, at time t is expressed by the following formula:

$$d_t = \begin{cases} C_{t+1} - C_t, & t < k \\ \sum_{k=1}^{k}(C_{t+1} - C_{t-1})/2 \sum_{k=1}^{k} k^2, & otherwise \\ C_t - C_{t-1}, & t \ge Q - k \end{cases} \tag{3}$$

where k is the difference in time and it take 1 in this paper. C_t is the MFSC coefficient at time t. Q is the maximum number of MFSC coefficients, which is 128 in this paper. After getting the first-order time derivatives from this formula, these results are used as the input and then be processed by the above formula again to get the second-order time derivatives.

In order to match the CNN input size, the MFSC and its first- and second-order time derivatives are adjusted to a size of 224 × 224 by zero-padding, respectively. These three feature images will serve as the input to the three channels of the classification network.

3.2. Network structure

In the paper, the ResNet is used as the backbone network, and the DCA module is integrated on the basis of this backbone network. Therefore, the network is named DCA-ResNet. Its overall network structure is shown in Fig. 4. Each part of the network will be introduced in detail, namely the backbone network and the DCA module.

3.2.1. Backbone network

The feature maps extracted by CNN from MFSC are abstract and complex advanced features and the secondary feature extraction of MFSC is automatically completed. It simplifies the entire CS-PVC system. Since the performance of ResNet is very outstanding in many classification tasks, our study selects ResNet as the backbone network. The core idea of ResNet is to introduce a shortcut connection residual block. This structure solves the problems of gradient vanishing and exploding in the neural network, and reduces the loss error of deep network. At the same time, it can also simplify the optimization process and make the training process fast without adding additional parameters or computational complexity. The residual block structure is shown in Fig. 5, where the weight layer represents a different number convolutional layers. The output and input of residual blocks are denoted by y and x, respectively, and the relationship between x and y is expressed as:

$$y = f(x) + W_3 x \tag{4}$$

where W_3 is the weight of the convolution which makes the input x and $f(x)$ the same number of channels, and $f(x)$ is the output of the second weight layer defined as:

$$f(x) = W_2 \sigma\left(W_1 x\right) \tag{5}$$

where W_1 and W_2 represent the weights of the first and second weight layers in Fig. 4, and σ represents the rectified linear unit (ReLu) function. ResNet has different network layers, and commonly used layers are ResNet18, ResNet34, and ResNet50. They are all implemented by stacking the above residual blocks together [47].

3.2.2. DCA module

In this study, a novel residual attention block is proposed. The importance of each feature channel is automatically obtained by learning. Then according to the importance, the feature enhanced or suppressed to well complete the task. It corresponds to the attention in Fig. 4. The residual attention block is shown in the red dashed box in Fig. 6. First, the features of each channel are compressed into real numbers by global average pooling (GAP). It is defined as:

$$G_i = \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j), G_i \in R^c \tag{6}$$

where H, W represent the length and width of the input, and u_c represents the cth convolution kernel. The correlation between the channels is established by two fully connected (FC) layers, and the

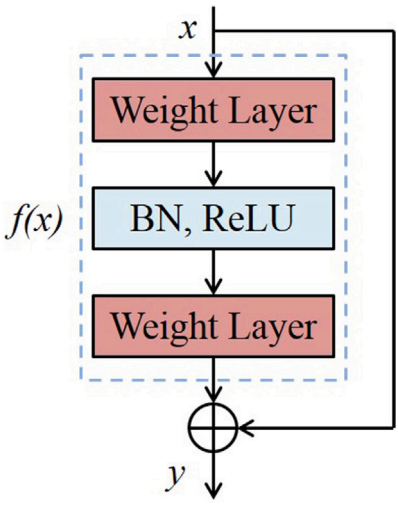

Fig. 5. The structure of the residual block.

normalized weight is obtained by a Sigmoid function. The output of the ith T in Fig. 6 is defined as:

$$T_i = \sigma(W_2 ReLU(W_1 G_i)) \tag{7}$$

where σ represents the sigmoid function, and W_1, W_2 represent the parameters that the network can learn. Finally, the above weights are weighted on the feature map of each channel.

In addition, Fig. 6 shows the connection between the two residual attention blocks. This connection mechanism enables information to flow between attention modules. It effectively avoids frequent changes of information between attention modules, thereby improving the learning ability of attention modules. The connection function is expressed as:

$$f(\alpha G_i, \beta T_{i-1}) = \alpha G_i + \beta T_{i-1} \tag{8}$$

where α and β are learnable parameters, T_{i-1} represents the output of the previous sigmoid function and G_i represents the output of the current GAP process. This designed connection ensures that the current residual attention block is able to learn the information of the previous residual attention block.

4. Experiment

In this section, detailed information of the experimental setup will be introduced, including the dataset, evaluation metrics and implementation details.

4.1. Database

4.1.1. SVD database

Saarbruecken voice database (SVD) was recorded by Institute of Phonetics of Saarland University in Germany. It collected voice recording and electroglottography (EGG) signals from 2041 speakers, which contains 687 healthy persons (428 females and 259 males) and 1356 patients (727 females and 629 males) with 71 different voice pathologies. Each speaker was recorded with the following pronunciations: (1) sustained vowels /a, i, u/ produced at normal, high, low, low–high–low pitch; (2) The German sentence "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?"). All the recordings are sampled at 50 kHz sampling rate with 16-bit resolution. In this paper, damaged and unclear samples are removed. Thus, 1685 sustained normal pitch vowels /a/ were used in our experiment, including 595 healthy recordings and 1090 pathological recordings.

4.1.2. Self-built database

We have established the voice pathology database in cooperation with Shenzhen People's Hospital. This voice database is named SZUPD which is also used in the experiment. It contains recordings of the vowels /a/ of 40 healthy persons and 67 patients with different voice pathologies. All recordings are sampled at 22 kHz sampling rate. Currently, the SZUPD database is still expanding.

4.2. Evaluation metrics

The proposed method is evaluated in terms of accuracy, precision, recall and F1 score. They are respectively defined by the following formula:

$$accuracy = tp + tn/tp + fp + tn + fn \tag{9}$$

$$precision = tp/tp + fp \tag{10}$$

$$recall = tp/tp + fn \tag{11}$$

$$F1 = 2 \cdot precision \cdot recall/precision + recall \tag{12}$$

where tp, fp, tn and fn denote the number of true positives, false positives, true negatives and false negatives, respectively. In this study, pathological samples are the main focus. Therefore, pathological samples are set as positive and healthy samples are set as negative. Accuracy indicates the number of correct samples in the total samples, and can measure the overall performance of the model. Precision indicates the proportion of predicted pathological samples to all pathological samples. It measures the ability of the model to correctly detect pathological sample. Recall indicates the proportion of the predicted pathological samples in the total samples, which measures the model's ability to fully retrieve pathological voices. F1 score is described as the harmonic average of the precision and the recall. The range of the four indicator values is 0 to 1. The larger the index value, the better the detection performance.

4.3. Implementation details

In order to train and verify the model, the dataset is divided into disjoint training set and test set, and the ratio is 8:2. During model training, the training set is divided into 10 equally subsets and 10-fold cross-validation is used to adjust the hyperparameters. The test set is used to evaluate the final performance of the model. Then, the stochastic gradient descent (SGD) is used to optimize the cross-entropy loss function with a batch size of 64 samples. The training parameters of this model are set as follows: the learning rate is 0.005, the momentum is 0.9, and the weight decay is 0.0005. The proposed method is implemented on an Ubuntu server equipped with three GPUs (NVIDIA Titan XP) and the PyTorch is used to build the proposed network with CUDA9.0. The parameters of the network are initialized by default method in PyTorch. In order to compare the performance of different models fairly, they are trained without pre-training.

5. Result and discussion

5.1. Evaluation of MFSC and MFCC

According to the evaluation indicators described in Section 4.2, four evaluation metrics, namely accuracy, precision, recall and F1 score are used to obtain the evaluation results. Different input features, including MFSC and MFCC, and different networks, including ResNet18, ResNet34, and ResNet50, are evaluated. The evaluation results are shown in Table 2. When MFSC is used as the input feature and Resnet18 is used as the network, the index result of recall does not achieve the highest performance. The decrease in recall is probably caused by the increase in precision. This phenomenon is very common

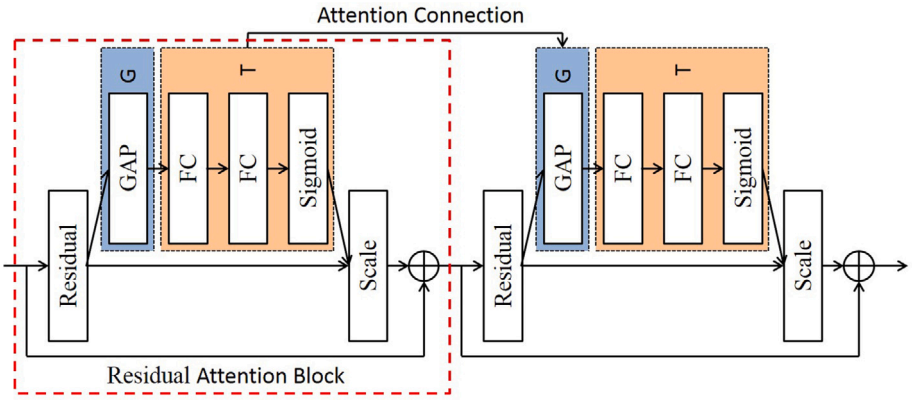

Fig. 6. The structure of the proposed DCA module.

Table 2
Evaluation results using different input features, including MFSC and MFCC, and different depth of the ResNet networks.

Methods	MFSC				MFCC			
	accuracy	precision	recall	F1 score	accuracy	precision	recall	F1 score
ResNet18	**0.786**	**0.838**	0.830	**0.834**	0.751	0.768	**0.881**	0.821
ResNet34	0.766	0.806	0.839	0.822	0.762	0.811	0.826	0.818
ResNet50	0.763	0.792	0.858	0.824	0.745	0.777	0.849	0.811

in classification tasks [48]. However, the F1 score is the harmonic average of precision and recall. It is able to represent the overall performance of these two evaluation indicators, and MFSC combined with ResNet18 has achieved the best F1 results. Therefore, the best overall performance is obtained by MFSC Combined with ResNet18, where the accuracy, precision and F1 score are 0.786, 0.838 and 0.834, respectively.

In the case of the same network depth, the overall performance of MFSC is better than MFCC. The possible reason is that the MFSC features make it easier for CNN to discover linear relations as well as higher order causes of the input data, leading to a better overall system performance [40]. In addition, the Overall performance decreases when ResNet goes deeper. The possible reason is that the original input data size is small and the resolution decreases after resizing. Following that, as the network deepens, some details information may be lost.

5.2. Evaluation of DCA-ResNet

According to the results shown in Table 2, the combination of MFSC and ResNet18 achieves the best performance. Therefore, in order to make a fair comparison, in this part, all models are tested on the basis of using MFSC as the input feature. Five different classification models are trained for comparison to show the performances of different models, namely AlexNet [49], VGG16 [50], ResNet [47], DA-ResNet and DCA-ResNet. Table 3 shows the evaluation results of the above five models in terms of accuracy, precision, recall and F1 score. In this table, the performances of the ResNet-based models are better than AlexNet and VGG16. Compared with the simple ResNet model, the accuracy, precision and F1 score of DA-ResNet are improved by 1.8%, 2.4% and 1.2%, respectively. This outstanding contribution mainly comes from the unique attention module in DA-ResNet. Among all the models, DCA-ResNet achieved the highest results in accuracy, precision, recall and F1 among all models which are 0.816, 0.875, 0.835 and 0.855, respectively. In addition, the region of convergence (ROC) curves are plotted in Fig. 7, where AUC is defined as the area under the ROC. The larger the AUC value, the better the model classification performance. As shown in Fig. 7, the DCA-ResNet achieves the highest AUC value, which is 0.881. This result demonstrates that the DCA module greatly improves the overall performance.

In this experiment, the situation that the proposed DCA-ResNet algorithm cannot predict correctly usually occurs when the distinguished

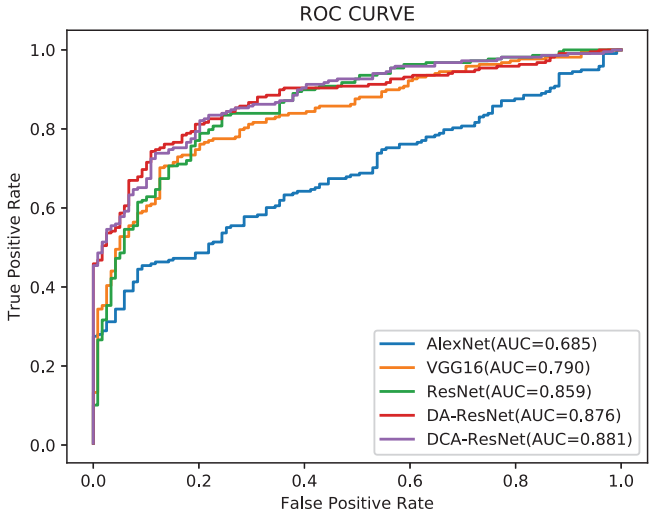

Fig. 7. ROC curves of different models.

samples are from mild voice diseases. The spectrogram example of a mild voice disease incorrectly judged by the proposed algorithm is shown in Fig. 8(b). In this example, the patient has mild laryngitis, and the frequency spectrum is relatively stable in the time–frequency domain. As a comparison, the voice disease spectrum that can usually be correctly predicted is shown in Fig. 8(a). It can be seen that the energy is obviously unstable or jitter in the time–frequency domain, which usually reflects the presence of disconnected pronunciation or hoarseness from the patient. Therefore, the more severe the symptoms, the easier it is to be detected correctly.

5.3. Evaluation of model generalization performance

In this part, two databases, SVD and SZUPD, are used to verify the generalization of the proposed DCA-ResNet. Two sets of experiments are conducted: (1) SVD was used for training while SZUPD was used for testing; (2) SVD was used both for training and testing. Since the sample size of SZUPD is too small (round 107 cases), it is not suitable

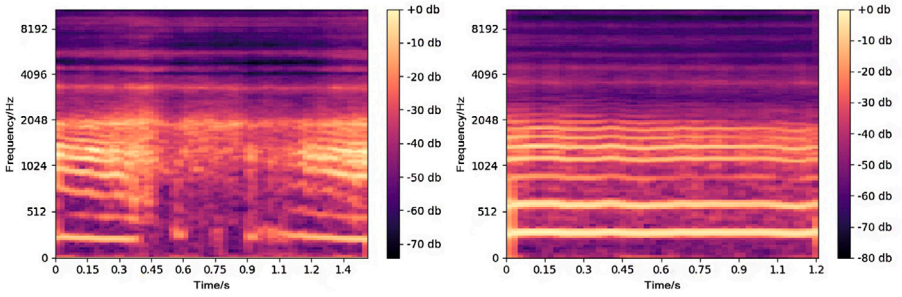

Fig. 8. (a) Log mel-frequency spectrogram of correctly detected voice sample. (b) Log mel-frequency spectrogram of falsely detected voice sample.

Table 3
The comparison of evaluation results using different models.

Meathod	Accuracy	Precision	Recall	F1 score
AlexNet [49][51]	0.721	0.772	0.807	0.789
VGG16 [36][50]	0.766	0.814	0.826	0.820
ResNet [47]	0.786	0.838	0.830	0.834
DA-ResNet	0.804	0.862	0.830	0.846
DCA-ResNet	**0.816**	**0.875**	**0.835**	**0.855**

Table 4
The comparison of evaluation results using different database.

	Accuracy	Precision	Recall	F1 score
Train:SVD Test:SZUPD	0.822	0.980	0.731	0.837
Train:SVD Test:SVD	0.816	0.875	0.835	0.855

Table 5
The number of samples selected from SVD database.

	Laryngitis	Rekurrensparse	Dysphonia	H.F. Dysphonia
male	50	127	42	37
female	33	70	29	115

for the training set. The results in Table 4 show that the proposed DCA-ResNet also achieves good performance on SZUPD. This finding shows that the proposed DCA-ResNet is not data dependent and has good generalization.

5.4. Evaluation of pathological voice classification

After detecting the disease voices, this study further classifies the types of diseases. Several disease types with a large number of samples in the SVD database are considered, and samples with multiple diseases at the same time are excluded. The considered types of diseases are laryngitis, rekurrensparse, dysphonia and hyperfunktionelle dysphonia. Table 5 shows the details of the selected samples. Among them, 70% of the data is randomly selected as the training set, and the remaining 30% is used as the test set. In this part, MFSC is used as the input feature, DCA-ResNet is used as the network, and the confusion matrix (CR) is used to evaluate the test results.

Table 6 is the CR results of four pathological classifications. The rows of the table represent true subjects, and the columns represent prediction subjects. Table 6 shows that rekurrensparse has the highest classification accuracy, reaching 0.627. The classification accuracy of laryngitis, dysphonia and hyperfunktionelle dysphonia is relatively low, and these diseases are all easily misidentified as rekurrensparse. The possible reason is that the effects of the above-mentioned different diseases on the voice are very close, so it is easy to form misjudgment. At the same time, due to the relatively large number of samples of rekurrensparse, the model may be biased toward it. Therefore, other diseases are easily misjudged as rekurrensparse. Finally, the average prediction accuracy of all diseases is 0.470.

6. Conclusion

In this paper, a novel system CS-PVC is proposed for pathological voice detection and classification. Firstly, the MFSC features are extracted from the original voice signals, and then they are passed into the DCA-ResNet to predict the voice pathologies. The attention modules in DCA-ResNet enhances the useful features according to the weights changes in the learning process. In addition, all the attention modules are connected, the attention modules are thus able to exchange information with each other. The experimental results show that DCA-ResNet achieves the best performance in all the compared networks, which proves that the connection mechanism improves the ability of the attention module. In addition, a self-built database (SZUPD) is established to verify the generalization of the proposed model. In the case of using SVD as the training set, the accuracy tested on the SVD database is 0.810, and the accuracy tested on the SZUPD database is 0.822. The above results prove that the system proposed in this paper has good performance and strong generalization in pathological voice detection. After the pathological voice detection, this paper classifies the types of disease, including laryngitis, rekurrensparse, dysphonia and hyperfunktionelle dysphonia. In experiments to classify rekurrensparse the above four types of voice pathologies, rekurrensparse achieves the highest classification accuracy of 0.627, followed by hyperfunktionelle dysphonic, laryngitis and dysphonia. It shows that the proposed system CS-PVC has application potential and room for improvement in pathological voice classification.

The future work is to propose and test some new acoustic features, so that the pathological voices and healthy voices can be distinguished accurately, and the differences between different pathological voices can be effectively amplified. Secondly, this study used the vowel /a/ only for detection and classification. Many voice diseases are not obvious in the pronunciation of a single vowel, but may have obvious

Table 6
The CR results of four pathological classifications.

True	*Prediction*				
	Laryngitis	Rekurrensparse	Dysphonia	H.F. Dysphonia	Class accuracy
Laryngitis	**7**	14	1	3	0.280
Rekurrensparse	3	**37**	11	8	0.627
Dysphonia	0	11	**8**	2	0.381
H.F. Dysphonia	4	17	6	**19**	0.413

characteristics in continuous speech. Therefore, we will try to analyze continuous speech in future work. Finally, the proposed network structure is relatively complicated which requires high computational cost. In response to this problem, we will optimize the network structure to achieve fast calculation speed and good performance, laying the foundation for the proposed system to be used in actual clinical applications in the future.

CRediT authorship contribution statement

Huijun Ding: Writing - reviewing and editing. **Zixiong Gu:** Methodology, Design-implementation and Writing. **Peng Dai:** Writing - reviewing and editing. **Zhou Zhou:** Data collection. **Lu Wang:** Data collection. **Xiaoxiao Wu:** Writing - reviewing and editing.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.bspc.2021.102973.

Acknowledgments

The authors would like to thank Otolaryngology Head and Neck Surgery of Shenzhen People's Hospital for providing pathological voice samples and guidance about pathological voice. This work was supported by the by National Natural Science Foundation of China 61862043 and Natural Science Foundation of Guangdong Province of China 2021A1515011915.

References

[1] N. Bhattacharyya, The prevalence of voice problems among adults in the United States, The Laryngoscope 124 (10) (2014) 2359–2362, http://dx.doi.org/10.1002/lary.24740.

[2] M.A. Morris, S.K. Meier, J.M. Griffin, M.E. Branda, S.M. Phelan, Prevalence and etiologies of adult communication disabilities in the United States: Results from the 2012 National Health Interview Survey, Disabil. Health J. 9 (1) (2016) 140–144, http://dx.doi.org/10.1016/j.dhjo.2015.07.004.

[3] L.I. Black, A. Vahratian, H.J. Hoffman, Communication disorders and use of intervention services among children aged 3-17 years: United States, 2012, NCHS Data Brief (205) (2015) 1–8.

[4] S. Marmor, K.J. Horvath, K.O. Lim, S. Misono, Voice problems and depression among adults in the United States, The Laryngoscope 126 (8) (2016) 1859–1864, http://dx.doi.org/10.1002/lary.25819.

[5] J. Stewart, Ear, nose, and throat diseases, Br. Med. J. 2 (4994) (1956) 701.

[6] D.D. Mehta, R.E. Hillman, Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods, Curr. Opin. Otolaryngol. Head Neck Surg. 16 (3) (2008) 211, http://dx.doi.org/10.1097/MOO.0b013e3282fe96ce.

[7] M.E. Smith, L.O. Ramig, C. Dromey, K.S. Perez, R. Samandari, Intensive voice treatment in parkinson disease: laryngostroboscopic findings, J. Voice 9 (4) (1995) 453–459, http://dx.doi.org/10.1016/s0892-1997(05)80210-3.

[8] R. Speyer, G. Wieneke, P. Dejonckere, Documentation of progress in voice therapy: perceptual, acoustic, and laryngostroboscopic findings pretherapy and posttherapy, J. Voice 18 (3) (2004) 325–340, http://dx.doi.org/10.1016/j.jvoice.2003.12.007.

[9] V. Uloza, A. Vegiene, V. Saferis, Correlation between the quantitative video laryngostroboscopic measurements and parameters of multidimensional voice assessment, Biomed. Signal Process. Control 17 (2015) 3–10, http://dx.doi.org/10.1016/j.bspc.2014.10.006.

[10] M.S. De Bodt, F.L. Wuyts, P.H. Van de Heyning, C. Croux, Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality, J. Voice 11 (1) (1997) 74–80, http://dx.doi.org/10.1016/S0892-1997(97)80026-4.

[11] R.I. Zraick, G.B. Kempster, N.P. Connor, S. Thibeault, L.E. Glaze, Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-v), Am. J. Speech-Lang. Pathol. 20 (1) (2011) 14–22, http://dx.doi.org/10.1044/1058-0360(2010/09-0105).

[12] B.R. Gerratt, J. Kreiman, N. Antonanzas-Barroso, G.S. Berke, Comparing internal and external standards in voice quality judgments, J. Speech Lang. Hear. Res. 36 (1) (1993) 14–20, http://dx.doi.org/10.1044/jshr.3601.14.

[13] V. Mittal, R. Sharma, Glottal signal analysis for voice pathology, in: 2019 2nd International Conference on Innovations in Electronics, Signal Processing and Communication (IESC), IEEE, 2019, pp. 54–59, http://dx.doi.org/10.1109/IESPC.2019.8902368.

[14] J. Rafael Orozco Arroyave, J. Francisco Vargas Bonilla, E. Delgado Trejos, Acoustic analysis and non linear dynamics applied to voice pathology detection: A review, Recent Pat. Signal Process. 2 (2) (2012) 96–107, http://dx.doi.org/10.2174/2210686311202020096.

[15] G. Muhammad, M. Alsulaiman, Z. Ali, T.A. Mesallam, M. Farahat, K.H. Malki, A. Al-nasheri, M.A. Bencherif, Voice pathology detection using interlaced derivative pattern on glottal source excitation, Biomed. Signal Process. Control 31 (2017) 156–164, http://dx.doi.org/10.1016/j.bspc.2016.08.002.

[16] K. Elemetrics, Multi-Dimensional Voice Program (MDVP) [Computer Program], Pine Brook, NJ: Author, 1993.

[17] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, T.A. Mesallam, M. Farahat, K.H. Malki, M.A. Bencherif, An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification, J. Voice 31 (1) (2017) 113–e9, http://dx.doi.org/10.1016/j.jvoice.2016.03.019.

[18] E.S. Fonseca, R.C. Guido, A.C. Silvestre, J.C. Pereira, Discrete wavelet transform and support vector machine applied to pathological voice signals identification, in: Seventh IEEE International Symposium on Multimedia (ISM'05), IEEE, 2005, pp. 5–pp, http://dx.doi.org/10.1109/ISM.2005.50.

[19] R. Fraile, N. Sáenz-Lechón, J.I. Godino-Llorente, V. Osma-Ruiz, C. Fredouille, Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex, Folia Phoniatr. Logop. 61 (3) (2009) 146–152, http://dx.doi.org/10.1159/000219950.

[20] J.I. Godino-Llorente, S. Aguilera-Navarro, P. Gómez-Vilda, Lpc, LPCC and MFCC parameterisation applied to the detection of voice impairments, in: Sixth International Conference on Spoken Language Processing, 2000.

[21] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, A. Al-Nasheri, T.A. Mesallam, M. Farahat, K.H. Malki, Intra-and inter-database study for Arabic, English, and German databases: do conventional speech features detect voice pathology? J. Voice 31 (3) (2017) 386–e1, http://dx.doi.org/10.1016/j.jvoice.2016.09.009.

[22] J.D. Arias-Londoño, J.I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, G. Castellanos-Domínguez, An improved method for voice pathology detection by means of a HMM-based feature space transformation, Pattern Recognit. 43 (9) (2010) 3100–3112, http://dx.doi.org/10.1016/j.patcog.2010.03.019.

[23] I. Hammami, L. Salhi, S. Labidi, Voice pathologies classification and detection using EMD-DWT analysis based on higher order statistic features, IRBM (2020) http://dx.doi.org/10.1016/j.irbm.2019.11.004.

[24] D. Hemmerling, A. Skalski, J. Gajda, Voice data mining for laryngeal pathology assessment, Comput. Biol. Med. 69 (2016) 270–276, http://dx.doi.org/10.1016/j.compbiomed.2015.07.026.

[25] P. Harar, Z. Galaz, J.B. Alonso-Hernandez, J. Mekyska, R. Burget, Z. Smekal, Towards robust voice pathology detection, Neural Comput. Appl. (2018) 1–11, http://dx.doi.org/10.1007/s00521-018-3464-7.

[26] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 6645–6649, http://dx.doi.org/10.1109/ICASSP.2013.6638947.

[27] Y. Jin, B. Wen, Z. Gu, X. Jiang, X. Shu, Z. Zeng, Y. Zhang, Z. Guo, Y. Chen, T. Zheng, et al., Deep-learning-enabled MXene-based artificial throat: Toward sound detection and speech recognition, Adv. Mater. Technol. 5 (9) (2020) 2000262, http://dx.doi.org/10.1002/admt.202000262.

[28] S. Nema, A. Dudhane, S. Murala, S. Naidu, RescueNet: An unpaired GAN for brain tumor segmentation, Biomed. Signal Process. Control 55 (2020) 101641, http://dx.doi.org/10.1016/j.bspc.2019.101641.

[29] H. Ding, Z. Pan, Q. Cen, Y. Li, S. Chen, Multi-scale fully convolutional network for gland segmentation using three-class classification, Neurocomputing 380 (2020) 150–161, http://dx.doi.org/10.1016/j.neucom.2019.10.097.

[30] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2016) 1137–1149, http://dx.doi.org/10.1109/TPAMI.2016.2577031.

[31] B. Woldert-Jokisz, Saarbruecken Voice Database, Institut für Phonetik, Universität des Saarlandes, 2007.

[32] M. Eye, E. Infirmary, Elemetrics Disordered Voice Database (Version 1.03), Voice and Speech Lab, Boston, MA, 1994.

[33] T.A. Mesallam, M. Farahat, K.H. Malki, M. Alsulaiman, Z. Ali, A. Al-Nasheri, G. Muhammad, Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms, J. Healthc. Eng. 2017 (2017) http://dx.doi.org/10.1155/2017/8783751.

[34] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, C.-T. Wang, Detection of pathological voice using cepstrum vectors: A deep learning approach, J. Voice 33 (5) (2019) 634–641, http://dx.doi.org/10.1016/j.jvoice.2018.02.003.

[35] H. Wu, J. Soraghan, A. Lowit, G. Di Caterina, A deep learning method for pathological voice detection using convolutional deep belief networks, Interspeech 2018, 2018, http://dx.doi.org/10.21437/Interspeech.2018-1351.

[36] M. Alhussein, G. Muhammad, Voice pathology detection using deep learning on mobile healthcare framework, IEEE Access 6 (2018) 41034–41041, http://dx.doi.org/10.1109/ACCESS.2018.2856238.

[37] M.A. Kiliç, F. Öğüt, G. Dursun, E. Okur, I. Yildirim, R. Midilli, The effects of vowels on voice perturbation measures, J. Voice 18 (3) (2004) 318–324, http://dx.doi.org/10.1016/j.jvoice.2003.09.007.

[38] M.K. Arjmandi, M. Pooyan, M. Mikaili, M. Vali, A. Moqarehzadeh, Identification of voice disorders using long-time features and support vector machine with different feature reduction methods, J. Voice 25 (6) (2011) e275–e289, http://dx.doi.org/10.1016/j.jvoice.2010.08.003.

[39] A. Akbari, M.K. Arjmandi, Employing linear prediction residual signal of wavelet sub-bands in automatic detection of laryngeal pathology, Biomed. Signal Process. Control 18 (2015) 293–302, http://dx.doi.org/10.1016/j.bspc.2015.02.008.

[40] A.-r. Mohamed, Deep Neural Network Acoustic Models for ASR (Ph.D. thesis), 2014.

[41] H. Ding, Y. Soon, C.K. Yeo, A DCT-based speech enhancement system with pitch synchronous analysis, IEEE Trans. Audio Speech Lang. Process. 19 (8) (2011) 2614–2623, http://dx.doi.org/10.1109/TASL.2011.2156785.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, arXiv preprint arXiv:1706.03762.

[43] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141, http://dx.doi.org/10.1109/TPAMI.2019.2913372.

[44] X. Ma, J. Guo, S. Tang, Z. Qiao, Q. Chen, Q. Yang, S. Fu, DCANet: Learning connected attentions for convolutional neural networks, 2020, arXiv preprint arXiv:2007.05099.

[45] Y. Zhou, C. Li, B. Xu, J. Xu, J. Cao, Hierarchical hybrid attention networks for Chinese conversation topic classification, in: International Conference on Neural Information Processing, Springer, 2017, pp. 540–550, http://dx.doi.org/10.1007/978-3-319-70096-0_56.

[46] X. Mei, E. Pan, Y. Ma, X. Dai, J. Huang, F. Fan, Q. Du, H. Zheng, J. Ma, Spectral-spatial attention networks for hyperspectral image classification, Remote Sens. 11 (8) (2019) 963, http://dx.doi.org/10.3390/rs11080963.

[47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[48] M. Buckland, F. Gey, The relationship between recall and precision, J. Am. Soc. Inf. Sci. 45 (1) (1994) 12–19.

[49] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1097–1105, http://dx.doi.org/10.1145/3065386.

[50] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[51] M. Alhussein, G. Muhammad, Automatic voice pathology monitoring using parallel deep models for smart healthcare, IEEE Access (2019) 1, http://dx.doi.org/10.1109/ACCESS.2019.2905597.