Multi-task convolutional network voice enhancement with speaker feature constraints

ZHANG Linpu, LONG Hua, SHAO Yubin, DU Qingzhi

(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, China)

E-mail :1797662488@qq.com

Abstract: In order to solve the problem of speaker recognition in the environment of noise interference, a speech enhancement method based on multi-task learning was proposed as the front-end of the speaker recognition system. On the basis of Convolutional Neural Network (CNN), a multi-task learning model of fusion network based on speech enhancement and speaker recognition is constructed, and Mel Spectral Cepstrum Coefficient (MFCC) and fundamental tone periodic features are spliced at the input and output ends as auxiliary tasks, and the loss weight is adaptively adjusted by using the isoskesticity uncertainty. Experimental results show that compared with the CNN and DNN models that only input logarithmic power spectrum (LPS), the CNN model with auxiliary tasks can improve the performance of speech enhancement. In addition, the joint training of speech enhancement and speaker recognition task can enhance the speaker recognition effect under noise interference and improve the robustness of the model.

Keywords: speech enhancement; multi-task learning; speaker recognition; Convolutional neural networks

CLC Number: TP391.4 **Document Identification Code**: A **Article Number**: 2020-0648

A Multi-task Convolution Network with Speaker Feature Constraint for Speech Enhancement

LONG Hua，ZHANG Lin-pu， SHAO Yu-bin，DU Qing-zhi

（Faculty of Information Engineering and Automation，Kunming University of Science and Technology, Kunming 650500, China）

Abstract: Aiming at the problem of speaker recognition in noisy environment, a speech enhancement method based on multi task learning is proposed as the front end of speaker recognition system. Based on convolutional neural network (CNN), a fusion multi-task learning model of speech enhancement and speaker recognition is built. At the same time, Mel frequency cepstrum coefficient (MFCC) and pitch features are concatenated with the input and output as auxiliary tasks, and the loss weight is adaptively adjusted by same variance uncertainty. The experimental results show that the CNN model with auxiliary tasks can improve the performances of speech enhancement compared with the CNN and DNN models which only use logarithmic power spectrum (LPS) as input. In addition, the joint training of speech enhancement and speaker recognition tasks can improve the performances of speaker recognition under noise interference and the robustness of the model.

Key words: speech enhancement; multi-task learning; speaker recognition; convolution network

1 Introduction

Speech Enhancement (SE) is an important task and topic in acoustic research, which aims to recover clean speech without noise as much as possible under the condition of giving noisy speech. There are many kinds of speech enhancement methods, and the traditional mono speech enhancement algorithms are mainly divided into time domain and frequency domain methods. Time-domain methods, such as parameter- and filtering-based methods, mainly use filters to estimate the channel parameters and excitation parameters of the vocal organs [1]. Frequency domain methods are mainly based on short-time spectral

Received:2020-10-1 Funds:The National Natural Science Foundation of China (No.61761025) About author:**LONG Hua,** female, born in 1963, Ph.D., professor, CCF member, graduate supervisor, research direction is wireless network and audio signal processing; **Zhang Linpu,** male, born in 1996, master's degree candidate, research direction is audio signal processing, speech recognition; **Shao Yubin,** male, born in 1970, master's degree, professor, graduate supervisor, research direction of mobile communication and personal communication systems and signal processing; **Du Qingzhi,** male, born in 1977, master's degree, senior engineer, graduate supervisor, research direction is signal processing, communication and information systems

estimation, such as spectral subtraction [2], Wiener filtering [3], and minimum mean square error [4]. Wait.

In recent years, deep learning has become a research hotspot in the field of speech. Deep neural networks (DNNs) are widely used in speech recognition, speech synthesis, speaker recognition, and other fields. Because of the great success of deep learning in the field of image recognition, researchers have begun to apply similar methods in the direction of speech enhancement and achieved success [5]. Although the neural network greatly improves the quality of the noise reduction frequency in speech enhancement, in these models, the enhancement process not only suppresses the background noise, but also causes great damage to the original signal.

Speaker Recognition (SR), also known as voiceprint recognition, is also a direction with great research value in speech recognition tasks. Voiceprint recognition has become an important means of identity authentication, and the biometric feature of voiceprint has been successfully applied to surveillance, secure unlocking, voice-activated operation of smartphones and smart appliances, and judicial authentication. Currently, neural networks include d-vector [6] and x-vector [7] and other models have been successful in the field of speaker recognition. These models have achieved satisfactory results under ideal interference-free or high signal-to-noise ratio. However, the recognition rate of speaker recognition tasks in the interference environment, especially under the interference condition of low signal-to-noise ratio, decreases rapidly. Although X-Vector adds noise and reverberation to the dataset [8], this method is more for data augmentation considerations, and the noise energy added is relatively small, and the effect is still not ideal in the face of low signal-to-noise ratio environments. Therefore, in the face of complex noise types and signal-to-noise ratio environments in real situations, it is necessary to preprocess and enhance speech to improve the robustness of the speaker recognition system in the noisy environment. In Ref. [9], the DNN network was used to fit the nonlinear function relationship of the i-vector feature vectors of clean speech and noisy speech, and the approximate representation of the i-vector of clean speech was obtained, which reduced the influence of noise on speaker recognition. In Ref. [10], the Mel frequency cepstrum coefficient (MFCC) and the Ideal Binary Mask for noisy speech were comparedIBM) is spliced together with the logarithmic power spectrum and fed into the DNN network to predict these three characteristics corresponding to clean voice. Compared with the single-task model that only estimates the logarithmic power spectrum of pure speech, the framework adds additional constraints to the objective function to improve the effect of speech enhancement. In Ref. [11], a denoising autoencoder was designed, and a deep network was superimposed on it to form a deep structure, and the Gaussian hybrid model of the traditional i-vector was replaced by this neural network structure.

Inspired by the above work, this paper proposes a convolutional neural network speech enhancement model based on multi-task learning. In the multitasking framework, the model not only learns the mapping relationship between noisy speech and the logarithmic power spectrum of clean speech, but also takes discrete speaker labels as another output of the network. In addition, the continuous features of speech (such as MFCC, fundamental tone period) are used as auxiliary tasks in multi-task learning, hoping to provide more information to the network, and as a limiting term, the output power spectrum is constrained, so as to improve the speech enhancement and speaker recognition effect.

2 Spectrum mapping model based on CNN

Convolutional neural networks (CNNs) have achieved great success in the field of image recognition in recent years and have been widely used in commercial projects [12]. The principle of CNN is based on a simulation of biological visual habits and neural networks, and an approximation of local perception in the cerebral cortex. Compared with the standard fully connected DNN model, CNN can better adapt to the changes in the time and frequency domains of audio information, and overcome the influence of unstable environment and non-stationary noise in speech signals [13]. In addition, because the convolutional layer of CNN adopts the principle of parameter sharing and sparse joining, the number of parameters of the CNN model is also greatly reduced compared with DNN, which improves the training and operation speed, and makes it run better on devices with weak performance.

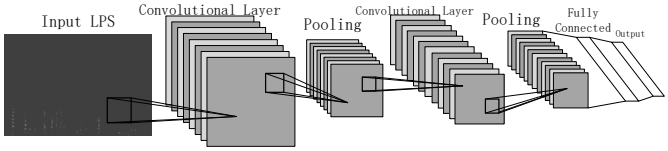

Fig.1 Structure of CNN model

Suppose there is a clean voice and a noise voice, the additive noisy speech obtained from them is:

$$y = x + n \tag{1}$$

The goal of speech augmentation is to y estimate the signal of clean speech as accurately as possible when the input is noisy speech x . The prediction of clean speech by the augmented model is denoted as $\hat{x}$: Speech enhancement based on spectrum mapping is to calculate the amplitude spectrum of clean speech based on the amplitude spectrum of noisy speech. The amplitude spectrum represents the intensity of the signal at different times and frequencies, generally speaking, the horizontal axis of the amplitude spectrum represents the time, the vertical axis represents the frequency, and the Z axis represents the amplitude of the corresponding signal.

The short-time Fourier transform x (STFT) of y the speech signal and the noisy speech after framing $X(n,k)$ is denoted as sum $Y(n,k)$, where the $n=1,2,\ldots\ldots,\ N$ number of frames is depresented. $k=1,2,\ldots\ldots,\ K$ Representing the band dimension, the number of points of the Fourier transform is D , because the Fourier transform is symmetrical, so the effective band dimension of the amplitude spectrum $K=D/2+1$. The complex sequence of STFTs of speech signals is as follows:

$$X\left(n,k\right)=X_r\left(n,k\right)+jX_i\left(n,k\right) \tag{2}$$

X_r and are X_i the real and imaginary parts on the STFT domain, respectively. The formula F for calculating its φ amplitude and phase is:

$$X_r(n,k)+jX_i(n,k)=\sqrt{X_r(n,k)^2+X_i(n,k)^2}e^{j\theta} \tag{3}$$

$$\varphi_x(n,k)=\arctan\frac{X_i(n,k)}{X_r(n,k)} \tag{4}$$

$$F_x\left(n,k\right)=\left|X\left(n,k\right)\right| \tag{5}$$

Because the human ear's perception of audio signals is nonlinear, the logarithmic power spectra (LPS) is obtained by taking the logarithm of the amplitude spectrum to enhance the weak part of the amplitude

$$P_x\left(n,k\right)=\ \log\left(F_x\left(n,k\right)\right) \tag{6}$$

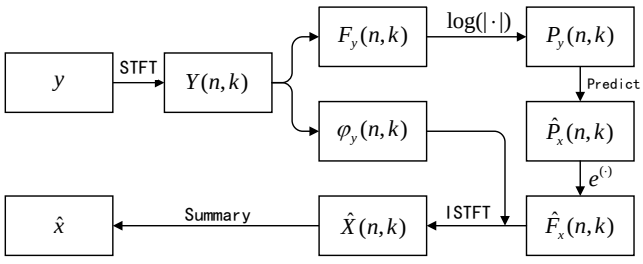

Fig. 2 Process of spectrum feature extraction and speech reconstruction

$Y(n,k)$ The LPS extraction method is the $X(n,k)$ same. In order to avoid the influence of extreme values and improve the convergence speed of the model, we normalize the logarithmic power spectrum of noisy speech $P_y\left(n,k\right)$ and the logarithmic power spectrum of clean speech $P_x\left(n,k\right)$. The logarithmic power spectrum of the noisy speech $P_y\left(n,k\right)$ is the input of the CNN enhancement model, and the model's estimation of the LPS of clean speech is the output of the model, which is denoted as $\hat{P}_x\left(n,k\right)$. The training goal of the CNN speech augmentation model is to minimize the loss function as follows L :

$$L=\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\hat{P}_x\left(n,k\right)-P_x\left(n,k\right)\right)^2 \tag{7}$$

First, the estimated LPS needs to be exponentially reverted back to the power spectrum:

$$\hat{F}_x(n,k)=\exp\left(\hat{P}_x\left(n,k\right)\right) \tag{8}$$

Because the phase estimation of the LPS for clean speech is difficult, and the experiment shows that the human ear does not perceive the phase change obviously, we use the phase of the noisy voice to reconstruct the clean speech:

$$\hat{X}(n,k)=\hat{F}_x(n,k)e^{j\varphi_y(n,k)} \tag{9}$$

$$\hat{x}(n)=\frac{1}{K}\sum_{k=1}^{K}\hat{X}(n,k)e^{j2\pi kn/K} \tag{10}$$

where is the K effective dimension of the Fourier transform mentioned above.

3 CNN acoustic model based on multi-task learning

3.1 Converged network multi-task learning

The concept of multi-task learning (MTL) is proposed relative to the standard single-task learning model. The traditional single-task model optimizes only one objective function at a time, for a single task. Multi-task learning, on the other hand, obtains correlation information between multiple tasks by sharing some parameters of the hidden layer [14].

Noisy speech contains not only speech and noise information, but also information such as pitch and timbre differences between different speakers. The speech enhancement task focuses on extracting the similarities between clean speech and distinguishing between speech and noise. Speaker recognition focuses on distinguishing the personal characteristics of different speakers in speech signals, such as pronunciation habits, timbre, etc. Therefore, by using the parameter sharing mechanism of multi-task learning, the neural network model enables the speech enhancement task and the speaker recognition task to obtain the implicit information between each other. The characteristics and commonalities between the two tasks are also reflected through this mechanism, which can help with their respective training.

Different from the traditional multi-task neural network model that shares all parameters, this paper uses a cross-stitch

unit [15], through which the hidden layer parameters can be shared between multiple tasks of the model, and the degree of sharing can be adjusted by adaptation through backpropagation. If the model has two tasks A B, the h_A sum h_B is the l output of the activation function of the first layer in the model, and h_A^{ij} the output at h_B^{ij} a specific position in the output matrix, respectively (i, j). Combine h_A and h_B input into the linear combination function to obtain the input sum of the next convolutional layer $h_A^{'}$ $h_B^{'}$

$$\begin{pmatrix} h_A^{'ij} \\ h_B^{'ij} \end{pmatrix} = \begin{pmatrix} \Lambda_{AA} & \Lambda_{AB} \\ \Lambda_{BA} & \Lambda_{BB} \end{pmatrix} \begin{pmatrix} h_A^{ij} \\ h_B^{ij} \end{pmatrix} \tag{11}$$

The arguments of these linear functions are noted that Λ the network can end the sharing by setting Λ_{AB} the sum Λ_{BA} to 0; Λ_{AB} Λ_{BA} Increase the degree of sharing by attaching and attaching higher values. Note that we only share parameters at the pooling layer or the fully connected layer of the CNN.

Λ are parameters that can be learned through training. Since the functions within the elements are linear combinations, the L partial derivative of the loss function for the pair can be calculated as follows: Λ

$$\begin{pmatrix} \frac{\partial L}{h_A^{ij}} \\ \frac{\partial L}{h_B^{ij}} \end{pmatrix} = \begin{pmatrix} \Lambda_{AA} & \Lambda_{BA} \\ \Lambda_{AB} & \Lambda_{BB} \end{pmatrix} \begin{pmatrix} \frac{\partial L}{h_A^{'ij}} \\ \frac{\partial L}{h_B^{'ij}} \end{pmatrix} \tag{12}$$

$$\begin{aligned} \frac{\partial L}{\partial \Lambda_{AA}} &= \frac{\partial L}{\partial h_A^{'ij}} h_A^{ij}, \frac{\partial L}{\partial \Lambda_{AB}} = \frac{\partial L}{\partial h_B^{'ij}} h_A^{ij} \\ \frac{\partial L}{\partial \Lambda_{BA}} &= \frac{\partial L}{\partial h_A^{'ij}} h_B^{ij}, \frac{\partial L}{\partial \Lambda_{BB}} = \frac{\partial L}{\partial h_B^{'ij}} h_B^{ij} \end{aligned} \tag{13}$$

As shown in Figure 3, the CNN model designed in this paper contains multiple convolutional layers and fully connected layers, in which each convolutional layer and fully connected layer need to perform nonlinear transformation operations, and we choose the Rectified Linear Unit (ReLU) as the activation function. The size of the convolution kernels is 3×3, and the number of convolution kernels is 64-64-128-128-256, the step size is 1×1. The pooling layer adopts Max-Pooling, with a size of 2×2. The convolutional layer is followed by two fully connected layers, each with 512 nodes. Finally, the output layer of the speech enhancement task is a fully connected layer containing 101 nodes, and the output layer of the speaker recognition task is also a fully connected layer, and the number of nodes is consistent with the number of speakers in the training set, and finally the output is through the softmax function.

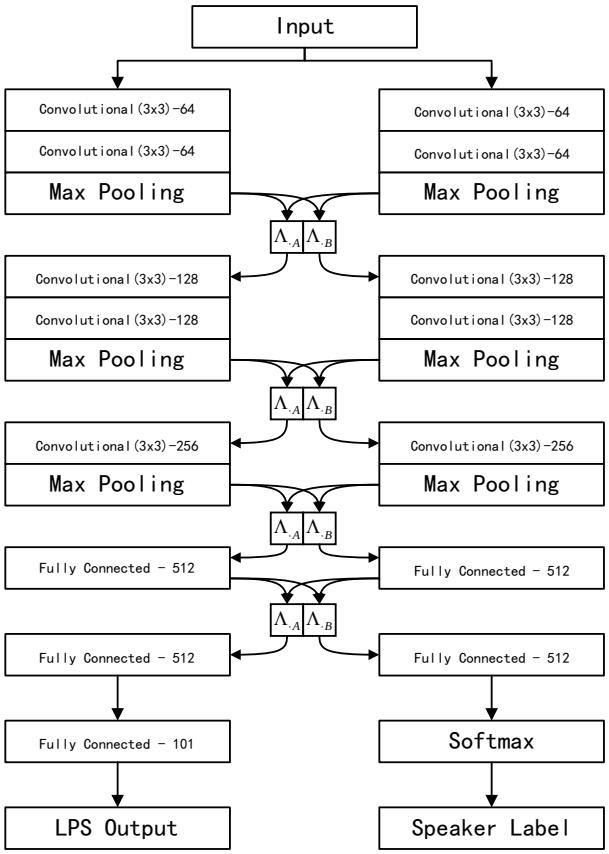

Fig.3 Structure of fusion network multi-task learning CNN model

3.2 Feature Joint Assisted Training

In CNN-based speech enhancement, the optimized objective function is the minimum mean square error between input and output on the logarithmic power spectrum [16, 17]. In the logarithmic power spectrum, different frequencies are assumed independently, the interrelationship between various dimensions is not considered, and the prediction of the model lacks constraints, which is not conducive to the simulation and perception of the auditory characteristics of the human ear [18].

In this section, we introduce the method of Auxiliary Learning to indirectly optimize the objective function. In the multi-task learning framework, if the data dimension of the main task is high and there are many irrelevant features, it will cause more difficulties in fitting the model. Helper tasks add constraints to the training of the model, allowing the model to focus more on the features that are closely related to the outcome. Therefore, the introduction of auxiliary tasks allows the model to learn not only the LPS of clean speech, but also continuous features such as Mel spectral cepstrum coefficient and fundamental period.

MFCC is a common speech feature used for tasks such as speech recognition, voiceprint recognition, and emotion recognition. In the calculation of the MFCC, a Mel filter set that is equidistant from the pitch change perceived by the human ear is applied to the cepstral spectrum, highlighting the low-frequency part of the sound and emphasizing the connection

between adjacent frames. In MFCC, the Meier triangle filter smooths the speech spectrum, eliminates the effect of harmonics, and highlights the formant peaks of speech. Therefore, the MFCC does not reflect the pitch or pitch of a speech, so if the MFCC is used as an input feature of the speech recognition system, the result will not be affected by the pitch of the input speech. But in fact, the change of pitch can indicate the difference in the pronunciation habits of different speakers, which is an important feature to describe speech stimulation.

The pitch reflects the periodicity of the vocal cords when a person pronounces voiced sounds, and the pitch period is the reciprocal of the frequency of the vocal cord vibration. The thinness, toughness, and length of the speaker's vocal cords have a lot to do with the fundamental period, so the fundamental period largely reflects the personality of the speaker's voice.

In this paper, the wavelet transform method is used to extract the fundamental period. Because the frequency and temporal resolution characteristics of the wavelet transform pair signal are very similar to the time-frequency analysis characteristics of the human ear, and the wavelet transform extreme points of the speech signal correspond to the opening and closing points of the glottis. So the fundamental period can be estimated using the distance between adjacent extreme points in the wavelet transform. The location of the mutation in the signal is reflected at the zero or extreme point. Therefore, according to the singularity in the wavelet signal, the detection of the fundamental tone period can be realized.

3.3 Adaptive Loss

In multi-task learning, multiple regression and classification tasks improve efficiency, prediction accuracy, and generalization ability by learning multiple objectives from shared representations. However, the scales are different between different tasks, which involves the joint learning of the objective functions of different unit-scale tasks in multi-task learning. Therefore, a very important problem in multi-task learning is how to design the loss function to balance different types of tasks and avoid the whole model being dominated by a single task in the training process. This involves assigning different weights to the loss functions of different tasks, and unifying the losses of different tasks into a single loss function. The conventional method is to simply add up the losses of each task or set a uniform weight, as follows:

$$loss = \frac{1}{N}\sum_{i=0}^{N} L_i \tag{14}$$

Further, the weights may be adjusted manually, resulting in the final model performing well on some tasks and less well on others.

Ref. [19] describes a method to adaptively adjust the weights of different loss functions by using the Homoscedastic Uncertainty. The uncertainty of isoskesticity is a contingent uncertainty that captures the confidence level of correlation between different tasks, so this uncertainty can be used as a measure for assigning the weight of loss to different tasks.

Assuming $f_w(x)$ that the output of the neural network is the input and x the model parameters are the w y correct output. In the case of multitasking, the K maximum likelihood of getting a discrete regression task is:

$$p(y_i \mid f_w(x)) = N(f_w(x), \sigma_i^2) \tag{15}$$

$$p(y_1, y_1, \cdots, y_k \mid f_w(x)) = \prod_{i=1}^{K} p(y_i \mid f_w(x)) \tag{16}$$

Each model follows a σ Gaussian distribution with noise scalars. For classification problems, it is usually output via the softmax function, as shown in the following formula:

$$p(y \mid f_w(x)) = \text{softmax}(f_w(x)) \tag{17}$$

So the estimation of maximum likelihood can be expressed as the negative log-likelihood of the minimized model in the following equation:

$$-\log p(y_1, y_1, \cdots, y_k \mid f_w(x)) \propto \sum_{i=1}^{K} \frac{1}{2\sigma_i^2} \| y_i - f_w(x) \|^2 + \log \sigma_i \tag{18}$$

Obtain the likelihood estimation of the softmax classification:

$$\log p(y = c \mid f_w(x), \sigma) = \frac{1}{\sigma^2} f_w^c(x) - \log \sum_c \exp\left(\frac{1}{\sigma^2} f_w^c(x)\right) \tag{19}$$

Taking the two output continuity y_1 and discrete types y_2 as examples, the Gaussian distribution and softmax are used to model the loss function:

$$\begin{aligned} L(w, \sigma_1, \sigma_2) &= N(f_w(x), \sigma_1^2) \cdot \text{softmax}(f_w(x)) \\ &= \frac{1}{2\sigma_1^2} L_1(w) + \frac{1}{\sigma_2^2} L_2(w) + \log \sigma_1 + \log \frac{\sum_c \exp\left(\frac{1}{\sigma_2^2} f_w^c(x)\right)}{\left(\sum_c \exp(f_w^c(x))\right)^{\frac{1}{\sigma_2^2}}} \\ &\approx \frac{1}{2\sigma_1^2} L_1(w) + \frac{1}{\sigma_2^2} L_2(w) + \log \sigma_1 + \log \sigma_2 \end{aligned} \tag{20}$$

Specify the loss function for each task $L_i(w) = \| y_i - f_w(x) \|^2$, and the joint loss of the final multi-task model is:

$$L(w, \sigma_{1:K}) = \sum_{i=1}^{K} \frac{1}{2\sigma_i^2} L_i(w) + \log \sigma_i \tag{21}$$

When the noise σ increases, the corresponding weight decreases; Conversely, as the noise σ decreases, the corresponding weight increases.

4 Experimental design and analysis of results

4.1 Experimental Setup

The speech data used in this experiment is the Free ST Chinese Mandarin Corpus Chinese voice library, which is resampled to 8000Hz frequency and single-channel clean voice. The Free ST Chinese Mandarin Corpus voice library contains 120 voices per person for 855 speakers, with a total of more than 100,000 Chinese Mandarin voice data. The noise data used is derived from the ESC-50 noise library [20], which includes 50 classes of 40 noises each.

Regarding the synthesis of noisy voices, the clean voices in each voice library are mixed with the noise by -5dB, -2dB, 0dB, 5dB, and 10dB respectively. Among them, the noise is randomly selected from the 50 categories of noise of the ESC-50, and the clip of the same length as the clean speech is randomly intercepted, the noise energy is adjusted according to the signal-to-noise ratio, and then mixed with the original speech.

In the speech preprocessing part, the spectral entropy method was used to detect the endpoint (SAD) of the original speech. After that, the signal is framed, and the frame length is consistent with the number of STFT points, which is 200 points (25ms). The frame shift is 80 points (10ms). A hamming window is then added to each frame and the STFT is calculated. The STFTs in the MFCC transform are the same as above, and the number of Meier filters is 40.

The clean voices in the test set were derived from 254 voices from 127 speakers in Free ST Chinese Mandarin Corpus that did not overlap with the training set; The noise comes from the 3 types of noise in the ESC-50 that do not overlap with the training set, namely the Laughing, Wind, and Train noises. The clean speech of the test set was mixed with randomly intercepted noise of different kinds according to the signal-to-noise ratio of -5dB, 0dB and 10dB, respectively, and a total of 4572 noisy test sets were obtained.

Perceptual Evaluation of Speech Quality (PESQ), Short Time Objective Intelligibility (STOI) and Segmented Signal-to-Noise Ratio (SSNR) were used in this experiment) as an evaluation index for speech enhancement results. PESQ focuses on enhancing the overall quality of speech, which is a commonly used standard method to evaluate voice quality, with a score between -0.5 and 4.5, with higher representing better voice quality. STOI is a commonly used index to evaluate speech intelligibility in the field of speech enhancement in recent years, with a score between 0 and 1, with higher representing better intelligibility. SSNR represents the signal-to-noise ratio of the enhanced voice, and the higher the voice, the cleaner the enhanced voice.

4.2 Experimental Testing

In order to verify the effect of the proposed model, the baseline of the DNN-based speech enhancement model is designed, and the model uses LPS as the input-output feature, where each input is an LPS containing a total of 11 frames of context, and the output is the middle frame of the corresponding clean spectrum. The model has 4 hidden layers, each hidden layer contains 2048 nodes, and the activation function is ReLU. CNN models that only include the two main tasks of speech enhancement and speaker recognition are denoted as CNN-Mul. The proposed CNN model with two main and auxiliary tasks is denoted as CNN-Mul-Aux. In the auxiliary training of the model CNN-Mul-Aux, we spliced the MFCC and fundamental tone period of noisy speech with the LPS at the input end, and spliced the MFCC and fundamental tone period of clean speech at the output end of speech enhancement.

Table 1 Performance test of speech enhancement under Laughing noise

model	Enter	SQ	STANDS	SSNR
DNN	-5dB	1.2368	0.4983	-2.4697
	0dB	1.6196	0.5968	-0.4677
	10dB	2.1627	0.7276	0.8542
CNN-Mul	-5dB	1.7425	0.6247	-1.3697
	0dB	2.0125	0.7607	-0.0356
	10dB	2.3647	0.8044	1.2456
CNN-Mul-Aux	-5dB	1.9465	0.6623	-1.8697
	0dB	2.2354	0.7709	0.2155
	10dB	2.4853	0.8153	1.5687

Table 2 Performance test of speech enhancement under Wind noise

model	Enter	SQ	STANDS	SSNR
DNN	-5dB	1.8075	0.6148	-1.2716
	0dB	2.2734	0.7963	-0.2536
	10dB	2.5180	0.8461	2.3215
CNN-Mul	-5dB	1.9863	0.7035	-1.0456
	0dB	2.3058	0.7812	0.6820
	10dB	2.5266	0.8240	3.0145
CNN-Mul-Aux	-5dB	2.1969	0.7791	-0.0184
	0dB	2.3561	0.8041	0.7713
	10dB	2.6710	0.8575	3.4791

Table 3 Performance test of speech enhancement under Train noise

model	Enter	SQ	STANDS	SSNR
DNN	-5dB	1.7669	0.6572	-1.1330
	0dB	2.0001	0.7356	-0.0849
	10dB	2.3630	0.8060	3.2812
CNN-Mul	-5dB	1.8849	0.5866	-0.9616
	0dB	2.2314	0.7379	0.4135

	10dB	2.4812	0.8316	2.9145
CNN-Mul-Aux	-5dB	1.9624	0.7037	-0.2131
	0dB	2.3633	0.8258	1.1040
	10dB	2.6073	0.8464	3.7687

Table 1, Table 2, and Table 3 compare the performance of the model under Laughing, Wind, and Train noises, respectively. It can be seen from Table 1~Table 3 that the CNN enhancement model achieves good results under the three noises, except for the SSNR under the 10dB Train noise, the indicators of CNN are consistently better than DNN in different SNR levels and environments. In addition, CNN also achieves better results than DNN in noise filtering, indicating that CNN does make better use of the correlation between LPS frequency domain and time domain, so that it can better estimate stationary and non-stationary noise in speech. In addition, the CNN-Mul-Aux model with auxiliary tasks is better than the CNN-Mul model and DNN without auxiliary tasks in all indicators, indicating that the splicing of MFCC features and fundamental tone periods at the input and output ends can indeed significantly improve the enhancement effect of the model, and can avoid the overfitting of the model on a separate LPS task, thereby enhancing the robustness of the model.

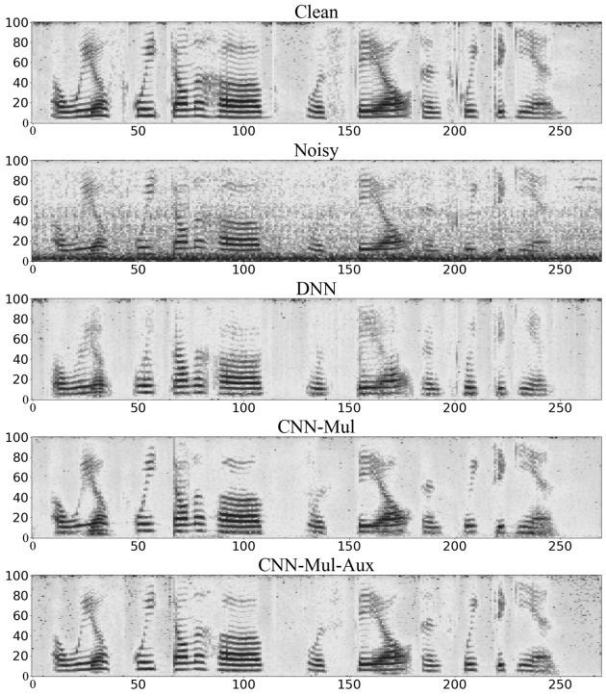

Fig. 4 Comparison of logarithmic power spectrum results

As shown in Figure 4, the logarithmic power spectra of the results of each model after the enhancement of noisy speech can be more intuitively reflected in the difference of the enhancement effect. We take a piece of speech with -5dB Train noise and augment it with each model, from top to bottom, clean voice, noisy voice, and logarithmic power spectrum after model enhancement. The horizontal axis represents the number of frames after the split, the vertical axis represents the frequency, and the Z axis represents the amplitude of the signal at the corresponding frequency.

From the comparison of the logarithmic power spectra of several algorithms in Figure 4, it can be seen that the enhanced voice purity of the CNN algorithm is higher than that of DNN, and the noise suppression effect is better. In addition, compared with the CNN-Mul model without auxiliary tasks, CNN-Mul-Aux retains the information of the original speech better.

In this paper, the X-Vector speaker recognition algorithm is used to evaluate the degree of distortion of the enhanced speech on the speaker information. The model framework is based on the Kaldi toolbox, and the training set uses the AISHELL-1 Chinese speech dataset [21]。 In the test stage, both the x-vector and CNN speaker recognition models took the vectors before softmax and performed Probabilistic Linear Discriminant Analysis (PLDA) to score them. The registration set was 1270 voices from 127 speakers in Free ST Chinese Mandarin Corpus that did not overlap with the training set, and another 2540 voices from the same speaker in the test set. The performance evaluation indexes used are equal error rate (EER) and minimum detection cost function (DCF). The DCF function is calculated as follows:

$$DCF = C_{FR}E_{FR}P_{target} + C_{FA}C_{FA}(1-P_{target}) \tag{22}$$

C_{FR} with C_{FA} penalty coefficients for false rejection and false acceptance, respectively, P_{target} and with a $(1-P_{target})$ priori probabilities for the real speaking test and impersonation test, respectively. We use NIST SRE 2016 to set $C_{FR}=1$ $C_{FA}=1$ $P_{target}=0.001$ this set of parameters. When these three values are selected, select a set of FRR and FAR values to make the DCF the smallest, and the DCF is minDCF. In this paper, the minDCF with NIST SRE 2016 parameters is denoted as minDCF16. Because minDCF16 not only considers the different costs of the two errors, but also takes into account the prior probability of the test situation, it is more reasonable than EER to evaluate the performance of the speaker recognition model.

In order to verify the speaker recognition performance of the proposed model, several different methods are compared. Among them, X-Vector is the result of directly identifying noisy speech with X-Vector model. DNN-x is the result of the speech enhanced by DNN on the x-vector model. CNN-ori-x is the result of the enhanced speech of the multi-task CNN model without auxiliary tasks on the x-vector model. CNN-x is the result of the speech on the x-vector model enhanced by the CNN model proposed in this paper. CNN-direct is the result of the CNN model proposed in this paper for direct speaker recognition. Table 4 shows the EER and minDCF16 of each algorithm under different signal-to-noise ratios.

Table 4 Results of speaker recognition performance test

model	Enter the	HONOR	minDCF16
x-vector	-5dB	23.12	0. 9683
	0dB	19.30	0.9473
	10dB	12.73	0.8854
DNN-x	-5dB	17.75	0.8968
	0dB	11.96	0.8303
	10dB	4.68	0.7526
CNN-ori-x	-5dB	10.56	0.8501
	0dB	6.77	0.8321
	10dB	3.94	0.7194
CNN-x	-5dB	8.63	0.8625
	0dB	5.91	0.8244
	10dB	3.66	0.6912
CNN-direct	-5dB	7.21	0.8245
	0dB	5.83	0.7569
	10dB	3.47	0.6830

As can be seen from Table 4, the recognition effect of X-Vector is greatly affected in the noisy environment, especially in the case of low signal-to-noise ratio, indicating that it is necessary to preprocess noisy speech. After DNN enhancement, the recognition rate of speech on x-vector has been greatly improved, but the effect is still not ideal. In addition, it can be seen that, especially in the case of -5dB, the recognition rate of CNN enhancement is significantly higher than that of the DNN model, indicating that the speaker recognition task of the fusion network provides more speaker information to the enhanced model through parameter sharing. Among them, the effect of CNN-x is better than that of CNN-ori-x, indicating that the auxiliary task helps CNN speech enhancement retain more speaker information. The CNN-direct model is slightly better than the CNN-x model, which may be due to the better recognition performance of the CNN than the TDNN in the x-vector, the Statistic Pooling layer in the x-vector, and the channel differences due to the different training sets.

5 Conclusion

In this paper, aiming at the problem of ignoring speaker information in traditional speech augmentation models, the speech augmentation technology under noise interference and its impact on speaker recognition are studied, and a convolutional neural network speech augmentation method based on multi-task and auxiliary task constraints is proposed. By constructing a multi-task learning model of fusion network for speech enhancement and speaker recognition, splicing MFCC and fundamental tone periodic features at the input and output ends, and adaptively adjusting the loss weight by using the isoskesticity uncertainty, the effect of speech enhancement is significantly improved, the distortion of denoised speech is reduced, and excellent performance is achieved in the speaker recognition task under noise interference.

References:

[1] Gannot S, Burshtein D, Weinstein E. Iterative and sequential Kalman filter-based speech enhancement algorithms[J]. IEEE Transactions on speech and audio processing, 1998, 6(4): 373-385.

[2] Karam M, Khazaal H F, Aglan H, et al. Noise removal in speech processing using spectral subtraction [J]. Journal of Signal and Information Processing, 2014, 5(2): 32

[3] Xia B, Bao C. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification[J]. Speech Communication, 2014, 60(7): 13-29.

[4] Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator[J]. IEEE Transactions on acoustics, speech, and signal processing, 1984, 32(6): 1109-1121.

[5] Xu Y, Du J, Dai L R, et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal processing letters, 2013, 21(1): 65-68.

[6] Variani E, Lei X, McDermott E, et al. Deep neural networks for small footprint text-dependent speaker verification[C]. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014: 4052-4056.

[7] Snyder D, Garcia-Romero D, Povey D, et al. Deep Neural Network Embeddings for Text-Independent Speaker Verification[C]. Interspeech, 2017: 999-1003.

[8] Snyder D, Garcia-Romero D, Sell G, et al. X-vectors: Robust dnn embeddings for speaker recognition[C].2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 5329-5333.

[9] Zhang H R. Research on Robustness of Speaker Recognition in Noisy Environment [D]. Jiangsu: Nanjing University of Posts and Telecommunications, 2018.

[10] Xu Y, Du J, Huang Z, et al. Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement[C]. Interspeech, 2015: 1508-1512.

[11] Tan Z, Mak M W, Mak B K W, et al. Denoised senone i-vectors for robust speaker verification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(4): 820-830.

[12] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C].

Advances in neural information processing systems, 2012: 1097-1105.

[13] Park S R, Lee J W. A Fully Convolutional Neural Network for Speech Enhancement[J]. Evaluation, 2017, 10(2): 5.

[14] Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning[C]. Advances in neural information processing systems, 2007: 41-48.

[15] Misra I, Shrivastava A, Gupta A, et al. Cross-stitch networks for multi-task learning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 3994-4003.

[16] Xu Y, Du J, Dai L R, et al. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 23(1): 7-19.

[17] Fu S W, Tsao Y, Lu X. SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement[C]. Interspeech, 2016: 3768-3772.

[18] Bregman A S. Auditory scene analysis: The perceptual organization of sound[M]. Massachusetts: MIT press, 1994.

[19] Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 7482-7491.

[20] Piczak K J. ESC. Dataset for environmental sound classification[C]. Proceedings of the 23rd ACM international conference on Multimedia, 2015: 1015-1018.

[21] Bu H, Du J, Na X, et al. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline[C]. 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, 2017: 1-5.

Attached Chinese references:

[9] Zhang Hongran. Research on the robustness of speaker recognition in noisy environment[D].Jiangsu:Nanjing University of Posts and Telecommunications,2018