EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH

Open Access

Research on monaural speech segregation based on feature selection

Xiaoping Xie, Yongzhen Chen*, Rufeng Shen and Dan Tian

Abstract

Speech feature model is the basis of speech and noise separation, speech expression, and different styles of speech conversion. With the development of signal processing methods, the feature types and dimensions increase. Therefore, it is difficult to select appropriate features. If a single feature is used, the representation of the speech signal will be incomplete. If multiple features are used, there will be redundancy between features, which will affect the performance of speech separation. The feature described above is a combination of parameters to characterize speech. A single feature means that the combination has only one parameter. In this paper, the feature selection method is used to select and combine eight widely used speech features and parameters. The Deep Neural Network (DNN) is used to evaluate and analyze the speech separation effect of different feature groups. The comparison results show that the speech segregation effect of the complementary feature group is better. The effectiveness of the complementary feature group to improve the performance of DNN speech separation is verified.

Keywords Feature selection, Group lasso, Deep Neural Network (DNN), Monaural speech segregation, Complementary feature group

1 Introduction

With the continuous development of artificial intelligence, increasing feature models are emerging to describe speech more accurately [1]. In the data era, the number of features and parameters is growing rapidly. Therefore, it is necessary to select the feature parameter for models based on statistical characteristics. The main way is to retain the features that contribute a lot to the representation model and delete redundant or irrelevant features. This method effectively reduces the feature parameter set of the model, thereby improving the performance of the model [2]. Therefore, feature selection is often used in the preprocessing of classification or other systems.

In the 1970s, Hoer proposed a biased estimation method of ridge regression by adding L2 regularization to the residual sum of squares (RSS) based on the defect of the least RSS method [3]. But this method does not realize the feature selection. In 1996, Tibshirani proposed Least Absolute Shrinkage and Selection Operator (LASSO), which is realized by adding L1 regularization to RSS. This method not only can select the best subset but also was the stability in ridge regression [4]. However, the Lasso method did not get attention and development until 2005 when the emergence of least angle regression (LARS) broke this situation [5]. In 2005, Zou proposed a new feature selection method based on regularization, that is, elastic net [6]. This method promoted the formation of group effects between features often retained or abandoned together with high correlation. When the input feature exists in the form of a group, the previous methods cannot achieve the selection well. Therefore, in 2006, Yuan and others proposed group lasso by extending the regularization of previous methods. This method realized feature selection from the perspective of the group. But it only works when the design matrix is orthogonal [7]. In 2010, Friedman

*Correspondence:
Yongzhen Chen
yz_chen@hnu.edu.cn
State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, Hunan University, Changsha, Hunan Province 410082, China

combined lasso with group lasso to solve the limitation of group lasso and proposed the sparse group lasso. This method gives a new way to solve the convex problem of lasso's algorithm. Moreover, it can be used to solve the problem when the design matrix is not orthogonal [8]. The feature selection method was applied to the speech field in 2013. Wang and others selected speech features and parameters based on group lasso. Finally, the optimal complimentary feature combination of amplitude modulation spectrogram (AMS) + Mel-frequency cepstral coefficient (MFCC) + relative spectral transform perceptual linear prediction (RASTA-PLP) was obtained [9]. In 2016, Wang proposed a Bark wavelet packet transform feature extraction algorithm based on Fisher's ratio to select feature parameters, which solved the problem of low recognition rate of MFCC feature parameters in noisy environments [10]. In the same year, some related scholars conducted research on two-dimensional feature representation and extraction. Wan and others proposed a feature recognition method based on block two-dimensional MMC, which can accurately identify and extract features through the overall fusion of the features extracted from each sub-block [11]. Yan proposed a new two-dimensional image feature extraction algorithm based on the Bayesian shape model algorithm. The experimental results show that the algorithm has greatly improved the accuracy of feature recognition [12]. In 2017, Zhao and others proposed a voice hybrid feature extraction and feature enhancement method based on the multi-layer Fisher (Multi-Fisher) criterion, which has achieved accurate recognition of the voice commands of specific doctors. Compared with the traditional Mel-frequency cepstral coefficient (MFCC) feature parameters, the feature parameters filtered by the Multi-Fisher criterion increase the accuracy from 86.1 to 94.2% [13]. In 2022, Liu proposed an enhanced feature fusion method based on correlation and deep learning to generate high-dimensional features, and use the Pearson correlation coefficient to select the optimal feature combination. The results show that the overall feature selection accuracy can reach 99.84% [14]. In 2022, Chen proposed a dynamic correlation-based feature selection (DRFS) algorithm. The algorithm uses conditional mutual information to measure the conditional correlation of selected features and categories. Compared with existing algorithms, the proposed algorithm can effectively improve the classification accuracy of feature selection [15]. However, with the emergence of new speech feature models, especially the multi-resolution cochleagram (MRCG) method, it is necessary to analyze and evaluate the feature combination. Then a more reasonable feature combination can be selected.

This paper will select the speech features and parameters based on the group lasso method. The separation effect of eight different speech feature groups is verified to obtain the feature group with the best separation effect. Firstly, eight widely used speech features and parameters are selected and combined. Then, the speech separation effect of different feature groups is evaluated and analyzed by DNN. Finally, the feature group with the best speech separation effect is determined by comparing them with each other. The experimental results show that the speech effect of the complementary feature group is better. The research results of this paper provide a certain reference for solving the feature extraction problem of speech signals in engineering applications.

2 Group lasso method

In recent years, more and more speech feature models have been discovered, such as MFCC, MRCG, etc. This section will combine features based on the above parameters to better describe speech. In speech separation applications, feature selection is to select the feature that can best express a target from multiple features. It not only reduces the number of features of the description object but also reduces the redundancy between different features to a certain extent. It achieves the best expression of the description object with the least features in the most effective way. Effective feature selection can not only effectively reduce the dimension of features and parameters, but also improve the expression of the model and the performance of the system. Group lasso takes the feature group of a feature as a single variable. In feature selection, if the regression coefficient of the variable is large, the whole group of features will be retained. On the contrary, if the regression coefficient corresponding to the variable is small, the whole group of features will be eliminated. Thus, the sparsity between different feature groups is realized, while the sparsity is not considered in the feature group. The definition of group lasso is as follows:

$$\widehat{\beta}^{GLasso} = \operatorname{argmin}\left\{ \|Y - X\beta\|^2 + \lambda \sum_{j=1}^{J} \left\|\beta_j\right\|_{K_j} \right\} \quad (1)$$

where $[K]$ is the kernel matrix, for vector $\vec{\eta} \in R^d$ and when $d \geq 1$, $[K]$ is defined as follows:

$$\|\eta\|_K = \sqrt{(\eta^T K \eta)} \quad (2)$$

where the dimension of $[K]$ is $d \times d$, and $[K]$ is a symmetric positive definite matrix. Generally, in order to simplify the calculation, we usually take $K_j = I_{pj}$, where $j = 1, ..., J$. It is obvious when $p_1 = ... = p_J = 1$, group lasso is simplified to the lasso method [16].

3 DNN

DNN is essentially an extension of the perceptron. Its basic structure is shown in Fig. 1.

The figure shows a DNN structure with 2 hidden layers. The two outermost layers in the DNN are called input and output layers, and the rest are hidden layers. Circle symbols represent neurons. There is a connection between every neuron in the previous layer and every neuron in the next layer. As the number of hidden layers and neurons increases, the structure of DNN becomes more complex. Assuming there are *m*input variables, the input–output relationship is:

$$y = \sigma(z) = \sigma\left(\sum_{i}^{m} \omega_i^* x_i + b\right) \tag{3}$$

where x is the input layer. y is the output layer. $\sigma(\bullet)$ is the activation function. ω is the weight value. b is the offset value.

The training and learning process of DNN mainly realizes the initialization of weight W and bias b through forward propagation. Then, the initialized weight W and bias b are modified and iteratively calculated by back propagation. When the changes of the weight W and the bias b are less than the set threshold, the iteration is stopped, and the trained DNN can be obtained [17].

4 Feature extraction

The extraction of speech feature parameters is crucial for expressing the processing of speech signals as supervised learning based on machine learning. Good features can greatly improve the performance of speech signal processing. This section mainly uses the newly proposed MRCG feature and eight feature parameters such as AMS, RASTA-PLP and MFCC in the widely used optimal complementary feature group for the subsequent group lasso-based feature selection process. The feature parameter results extracted below are all from the same audio data in TIMIT [18].

(1) AMS

Step 1: Full-wave rectification is carried out for each frequency channel to obtain envelope features, and quarter sampling is carried out.

Step 2: Framing and windowing are carried out. Then fast Fourier transform (FFT) is performed.

Step 3: In each frequency band, the FFT amplitude is multiplied by 15 triangular windows equally divided in the range of 15.6–40 Hz, and the AMS characteristic parameter can be obtained by summation [19].

(2) MFCC

Step 1: Pre-emphasis, framing, and windowing are performed.

Step 2: A short-time Fourier transform is performed.

Step 3: The Mel triangular filter bank is used to filter, which is transferred from the frequency domain to the Mel domain. It can smooth the spectrum, eliminate harmonics, and so on. The MFCC can be obtained.

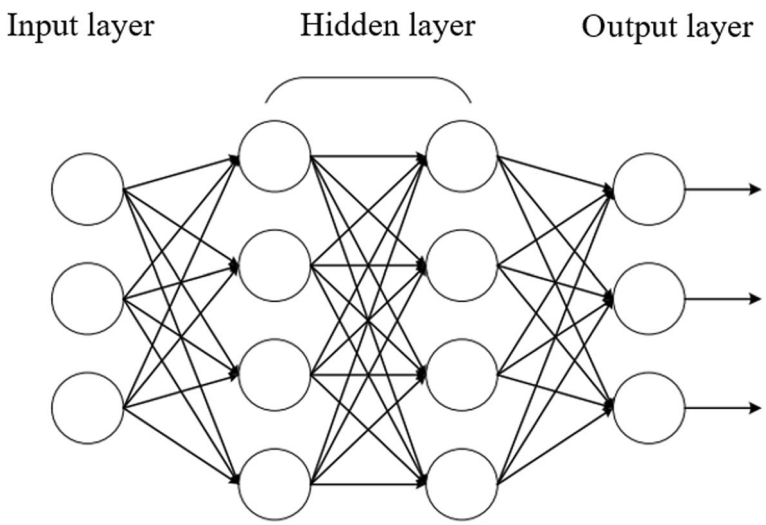

Fig. 1 DNN basic structure diagram

(3) RASTA-PLP

Step 1: For each time frame, the critical frequency band power spectrum is calculated as the same as PLP. The short-term energy spectrum of the speech signal is converted to the bark spectrum which conforms to the characteristics of human hearing.

Step 2: The power spectrum amplitude of static compression nonlinear conversion is a logarithm operation.

Step 3: The logarithmic spectral components in each frequency band are filtered by using an equivalent band-pass RASTA filter.

Step 4: To suppress the constant and slow-changing parts of the frequency band, the filtered speech representation is transformed by extending the static nonlinear transformation. That is, the inverse logarithm operation is performed on it.

Step 5: In simultaneous interpreting the traditional PLP, it is used to simulate the sensitivity of the human ear to different frequencies by multiplying the equal-loudness curve. And then the 0.33 power is obtained to compress the loudness amplitude.

Step 6: Continue the remaining steps such as discrete Fourier transform, obtaining all pole models, and other operations [20].

(4) Gammatone coefficient (GC)

Step 1: The signal is filtered by gammatone filter banks.

Step 2: The output of each filter is resampled at a sampling frequency of 100 Hz.

Step 3: The gammatone frequency (GF) parameter is obtained by suppressing the amplitude through the cube root.

(5) Gammatone frequency cepstral coefficient (GFCC)

Step 1: The previous steps are the same as the GF extraction method.

Step 2: The amplitude suppression results are processed by discrete cosine transform to get the GFCC parameter [21].

(6) Linear prediction coefficient (LPC)

The differential operation is performed on the framed and windowed signal, and then the corresponding filter coefficients are obtained according to the minimum mean square error criterion, which is the LPC parameter.

(7) Linear prediction cepstral coefficient (LPCC)

Similar with the LPC extraction process, the LPCC can be obtained by converting the LPC results into the cepstrum domain.

(8) MRCG

Step 1: The signal cochleagram CG1 is extracted, and the logarithm operation is performed on CG1, in which the frame length is 20 ms and frameshift is 10 ms.

Step 2: The operation is similar to the first step in extracting the signal of cochleagram CG2. The difference is that the frame length is 200 ms and the frameshift is 10 ms.

Step 3: CG3 is obtained by smoothing CG1 through a window with 11-time frames and frequency channels centered on a given timing and frequency unit.

Step 4: The CG4 calculation method is similar to the previous step, except that the size of the square window is 23.

Step 5: Connect CG1 to CG4 to get MRCG [22].

MRCG parameter is obtained from four cochleagrams processed in different ways. The high-resolution cochleagram can highlight the local information of the signal, while the low-resolution cochleagram can capture more extensive spectrum and time background information. Figure 2 shows the spectrogram and the corresponding eight speech feature parameter maps of the same speech signal.

5 Experimental results and analysis

5.1 Experimental data and setup

The speech database and noise database used in this paper are from the DARPA TIMIT Acoustic–Phonetic Continuous Speech Corpus (TIMIT) [18] and NOISEX-92 [23], respectively. The TIMIT corpus was built for the acquisition of acoustic speech knowledge (model training) and the evaluation of automatic speech recognition systems (ASR). A total of 6300 pure audios are included in TIMIT. They are composed of 10 sentences each spoken by 630 speakers from 8 major dialect regions in the United States. Each sentence is named by the speaker details code and sentence code. It is subdivided into two parts: TRAIN and TEST. NOISEX-92 is a noise database widely used in speech enhancement tasks, which contains 15 types of noise. The mixed signal used in the experiment consists of pure speech signal and noise signal. The training process randomly selected 600

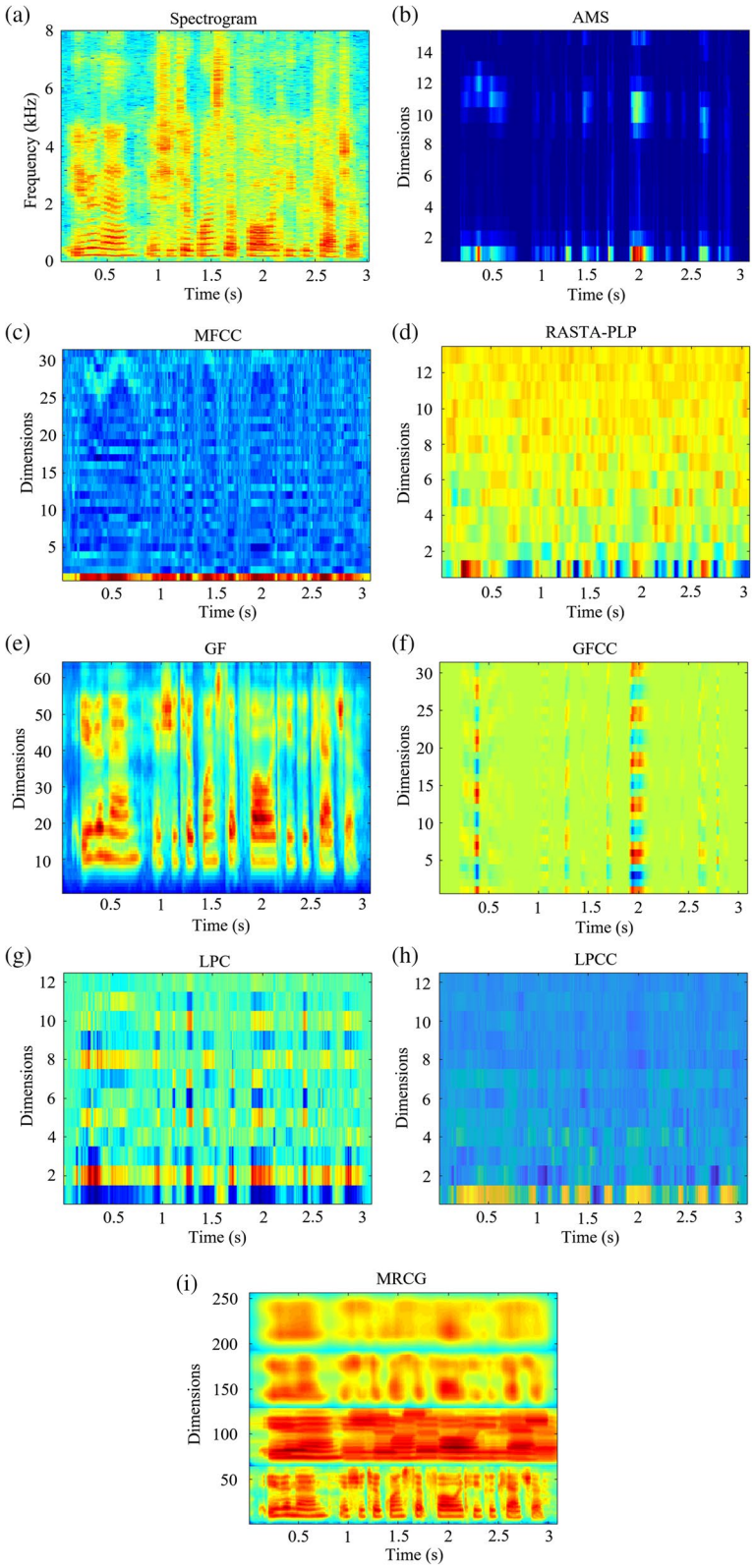

Fig. 2 Different feature maps of a speech signal: **a** spectrogram, **b** AMS, **c** MFCC, **d** RASTA-PLP, **e** GF, **f** GFCC, **g** LPC, **h** LPCC, and **i** MRCG

pure speech data from the TIMIT-TRAIN folder, and the testing process randomly selected 120 pure speech data from the TIMIT-TEST folder. These pure speech signals are mixed with random noise signal in NOISEX-92 at randomly signal-to-noise ratio (SNR) selected from $[-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5]$. Thus, we obtained the training set and test set used in our experiments, respectively. In our experiments, all selected audio signals were resampled to 16 kHz. The number of hidden layers of the DNN is 4, and each layer selects 1024 neurons.

5.2 Analysis of group lasso experimental results

This section performs feature selection on the above 8 feature parameters based on the Group LASSO method. Figure 3 shows the corresponding regression coefficients for each parameter. In Fig. 3, the regression coefficients of AMS, MFCC, GF, and MRCG are larger than other characteristic parameters. It indicates that these 4 characteristic parameters have a greater contribution to the model than others. Therefore, AMS + MFCC + GF + MRCG is selected as the optimal complimentary feature group.

5.3 Analysis of DNN verification results

Each feature in the complementary feature set can provide complementary information to better represent the speech signal and improve the separation performance of the system. This section conducts comparative experiments between a single feature and its different feature groups to verify the advantages of complementary feature groups.

The task of speech separation is to distinguish the unit belonging to the target signal from the interference unit. Therefore, it is more convenient to use classification accuracy to measure the performance of the system. Because the classification accuracy treat the unlabeled target unit and the labeled target unit equally, HIT, FA, and HIT-FA are used to evaluate the performance of the system. Where HIT refers to the percentage of accurately classified target signals in the masking matrix to the dominant action units. FA refers to the percentage of misclassified interference signals in the masking matrix to the dominant action units. HIT-FA refers to the difference between the HIT and FA [24, 25]. Normal cells are either marked 1 (speech) or 0 (noise). There are two types of wrongly labeled units: one is originally 1 but marked as 0; the other is 0 marked as 1. Therefore, classification accuracy means the percentage of correctly labeled units in the total number of labeled units. But the two wrong labels cannot be treated the same. The above is the reason for adopting HIT-FA in this paper.

Figure 4 shows the statistical results of HIT, FA, and HIT-FA in the noise mismatching, and Fig. 5 shows the corresponding statistical results in the noise matching

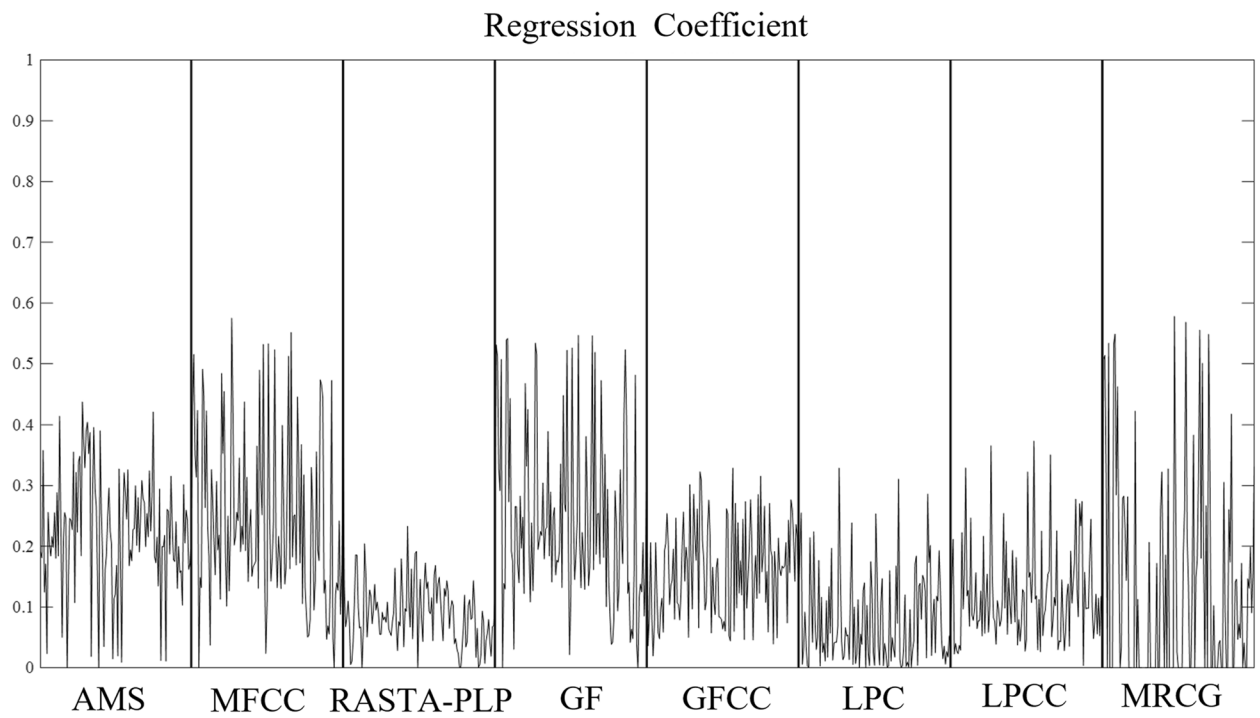

Fig. 3 Group lasso results

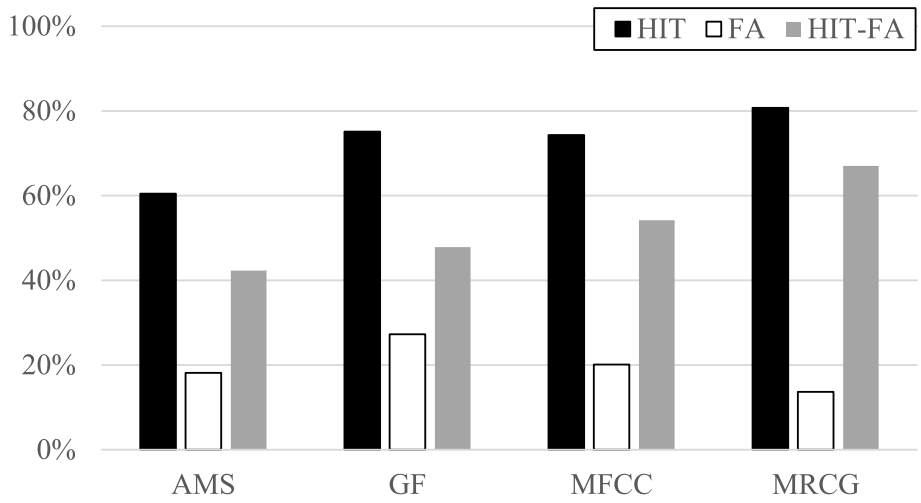

Fig. 4 HIT, FA, and HIT-FA statistical results in the noise mismatching condition

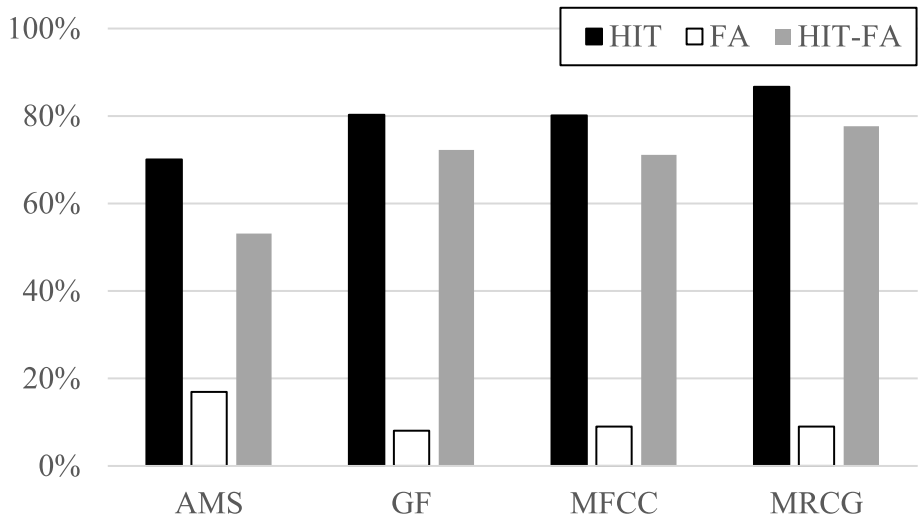

Fig. 5 HIT, FA, and HIT-FA statistical results in the noise matching condition

condition. Through analysis of Figs. 4 and 5, it is found that HIT-FA of all features is minimum in the noise mismatching condition because of its higher FA value. However, under the condition of noise matching, the FA value of the single feature parameter is low, and the difference mainly comes from the HIT value. It is not difficult to find that AMS does not perform well either under the condition of noise mismatching or noise matching. AMS is not good at labeling the unit containing the target signal information. MRCG feature is best than the other three single feature parameters in all cases.

The AMS + MFCC + GF + MRCG complementary feature group obtained in this chapter is compared with other feature groups. In the case of noise matching and

Table 1 HIT-FA results for feature groups in the noise mismatching and matching condition

Feature and feature group	Mismatching	Matching
AMS + GF	65.90%	77.80%
AMS + GF + MRCG	77.12%	79.41%
AMS + MFCC	69.65%	76.48%
AMS + MFCC + GF	74.02%	79.20%
AMS + MFCC + GF + MRCG	**77.57%**	**79.85%**
AMS + MFCC + MRCG	77.31%	79.44%
AMS + MRCG	76.78%	78.73%
GF + MRCG	76.84%	78.87%
MFCC + GF	72.64%	78.25%
MFCC + GF + MRCG	77.04%	79.45%
MFCC + MRCG	76.71%	79.12%

The bold data represents the optimal value of each feature group under this condition

mismatching, the corresponding statistical results are shown in Table 1.

It can be seen from Table 1 that HIT-FA of all feature groups has been significantly improved compared with that of the single feature. In comparison with the single feature, feature combination can effectively improve the classification accuracy of DNN speech separation system.

To verify the advantages of complementary feature group more comprehensively, single feature and different feature groups are trained using ideal binary mask (IBM) and ideal ratio mask (IRM) as objectives. Four metrics are used to evaluate the effectiveness of the AMS + MFCC + GF + MRCG feature set in improving system performance under different training target conditions. These metrics include short-term objective intelligibility (STOI) [26], perceptual evaluation of speech quality (PESQ) [27], signal-to-distortion ratio (SDR), and SNR.

Table 2 shows the calculation results in the noise mismatching condition when IBM is used as the training target. Through the comparison of the longitudinal data in Table 2, it is not difficult to find that AMS has the worst performance in features under the four evaluation criteria, and MRCG has the best comprehensive performance. In the feature group, AMS + MFCC + GF + MRCG performs best in STOI and performs better in PESQ, SDR, and SNR. Hence, AMS + MFCC + GF + MRCG is the best group in the single features and feature groups. Compared with AMS, AMS + MFCC + GF + MRCG increased STOI, PESQ, SDR, and SNR by 8.6%, 8.1%, 26.7%, and 25.7%, respectively. Compared with MRCG, the four evaluation criteria are improved by about 0.7%, 0.2%, 1.4%, and 2.4%, respectively.

Table 2 Calculation results in noise mismatching condition when IBM is the training target

Feature and feature group	STOI	PESQ	SDR	SNR
AMS	0.7517	2.2208	8.3039	8.2443
AMS + GF	0.8119	2.4371	10.5176	10.2655
AMS + GF + MRCG	0.8148	2.4167	10.6053	10.4037
AMS + MFCC	0.8088	**2.4427**	10.3002	10.0512
AMS + MFCC + GF	0.8132	2.4169	10.5283	10.3024
AMS + MFCC + GF + MRCG	**0.8166**	2.4006	10.5186	10.3628
AMS + MFCC + MRCG	0.8149	2.3970	**10.6519**	**10.4322**
AMS + MRCG	0.8127	2.4184	10.3728	10.2319
GF	0.8076	2.3951	10.2246	9.8945
GF + MRCG	0.8108	2.3995	10.4422	10.1793
MFCC	0.8067	2.4295	10.1689	9.7345
MFCC + GF	0.8096	2.4206	10.4220	10.0669
MFCC + GF + MRCG	0.8111	2.3881	10.5219	10.2396
MFCC + MRCG	0.8108	2.3769	10.6505	10.2960
MRCG	0.8110	2.3964	10.3754	10.1218

Bold data represents the optimal value of each feature group for this indicator

Table 3 Calculation results in the noise matching condition when IBM is the training target

Feature and feature group	STOI	PESQ	SDR	SNR
AMS	0.7527	2.2276	8.4709	8.4105
AMS + GF	0.8202	2.4656	10.8801	10.6167
AMS + GF + MRCG	0.8277	2.4594	10.9181	10.7706
AMS + MFCC	0.8208	2.4422	10.5386	10.3851
AMS + MFCC + GF	0.8259	2.4537	10.7574	10.6132
AMS + MFCC + GF + MRCG	**0.8779**	2.4314	**10.9985**	**10.8134**
AMS + MFCC + MRCG	0.8270	2.4574	10.9764	10.7877
AMS + MRCG	0.8250	2.4420	10.8013	10.6818
GF	0.8175	2.4599	10.7795	10.3668
GF + MRCG	0.8242	2.4439	10.8542	10.6339
MFCC	0.8170	2.4585	10.4417	10.0997
MFCC + GF	0.8225	**2.4759**	10.6328	10.2979
MFCC + GF + MRCG	0.8265	2.4323	10.8243	10.6493
MFCC + MRCG	0.8252	2.4542	10.9840	10.6970
MRCG	0.8201	2.4485	10.5185	10.3349

Bold data represents the optimal value of each feature group for this indicator

Table 3 shows the calculation results in the noise matching condition when IBM is used as the training target. Through the comparison of the longitudinal data in Table 3, it is not difficult to find that under the four evaluation criteria, AMS is still the worst in feature, and GF has the best comprehensive performance. In the feature group, AMS + MFCC + GF + MRCG has the best comprehensive performance. Compared with AMS, AMS + MFCC + GF + MRCG increased STOI, PESQ, SDR, and SNR by 16.6%, 9.1%, 29.8%, and 28.6%, respectively. Compared with GF, STOI, SDR, and SNR are increased by 7.4%, 2.0%, and 4.3%, respectively. In PESQ, GF was better than the characteristic group.

Table 4 shows the calculation results in the noise mismatching condition when IRM is used as the training target. In the single feature, AMS is still the worst, and MRCG is the best. In the feature group, AMS + MFCC + GF + MRCG has the best comprehensive performance. Compared with AMS, AMS + MFCC + GF + MRCG increased STOI, PESQ, SDR, and SNR by 8.7%, 13.4%, 29.0%, and 28.1%, respectively. Compared with MRCG, the four evaluation criteria are improved by about 0.8%, 1.3%, 2.0%, and 2.7%, respectively.

Table 5 shows the calculation results in the noise matching condition when IRM is used as the training target. In the single feature, AMS is still the

Table 4 Calculation results in the noise mismatching condition when IRM is the training target

Feature and feature group	STOI	PESQ	SDR	SNR
AMS	0.7538	2.2349	8.2144	8.1006
AMS + GF	0.8117	2.5034	10.4221	10.1191
AMS + GF + MRCG	0.8178	2.5258	**10.6274**	10.3807
AMS + MFCC	0.8108	2.4820	10.2555	10.0639
AMS + MFCC + GF	0.8160	2.5220	10.5311	10.2775
AMS + MFCC + GF + MRCG	**0.8193**	**2.5349**	10.5937	10.3745
AMS + MFCC + MRCG	0.8184	2.5227	10.5991	**10.4131**
AMS + MRCG	0.8168	2.5169	10.5676	10.3178
GF	0.8092	2.4863	10.2725	9.8783
GF + MRCG	0.8163	2.5130	10.4567	10.2289
MFCC	0.8077	2.4494	10.0376	9.7046
MFCC + GF	0.8133	2.5225	10.5097	10.1582
MFCC + GF + MRCG	0.8164	2.5303	10.5790	10.3323
MFCC + MRCG	0.8151	2.5228	10.5821	10.1904
MRCG	0.8131	2.5036	10.3848	10.0971

Bold data represents the optimal value of each feature group for this indicator

Table 5 Calculation results in the noise matching condition when IRM is the training target

Feature and feature group	STOI	PESQ	SDR	SNR
AMS	0.7554	2.2546	8.5116	8.4045
AMS + GF	0.8218	2.5484	10.8552	10.5734
AMS + GF + MRCG	0.8289	2.5875	11.1200	10.8836
AMS + MFCC	0.8210	2.5083	10.5937	10.3795
AMS + MFCC + GF	0.8277	2.5638	10.9731	10.7261
AMS + MFCC + GF + MRCG	**0.8324**	**2.6112**	**11.1327**	**10.8991**
AMS + MFCC + MRCG	0.8303	2.5799	11.0467	10.8180
AMS + MRCG	0.8270	2.5666	10.9469	10.7455
GF	0.8190	2.5563	10.8528	10.3649
GF + MRCG	0.8266	2.5774	10.9900	10.7394
MFCC	0.8178	2.4870	10.3623	10.0485
MFCC + GF	0.8260	2.5636	10.9189	10.5575
MFCC + GF + MRCG	0.8291	2.5972	11.1015	10.8133
MFCC + MRCG	0.8271	2.5767	10.9487	10.6630
MRCG	0.8227	2.5593	10.8100	10.5759

Bold data represents the optimal value of each feature group for this indicator

worst, and MRCG is the best. In the feature group, AMS + MFCC + GF + MRCG has the best comprehensive performance. Compared with AMS, AMS + MFCC + GF + MRCG increased STOI, PESQ, SDR, and SNR by 10.2%, 15.8%, 30.8%, and 29.7%, respectively. Compared with MRCG, the four evaluation criteria are improved by about 1.2%, 2.0%, 3.0%, and 3.1%, respectively.

To summarize, AMS + MFCC + GF + MRCG feature group has the best comprehensive performance under the evaluation criteria of STOI, PESQ, SDR, and SNR, whether under the training objectives of IBM or IRM or under the condition of noise matching or mismatch. In particular, the SDR criteria increased the most, up to 30.8%.

6 Conclusion

In this paper, eight widely used speech feature parameters AMS, MFCC, RASTA-PLP, GF, GFCC, LPC, LPCC, and MRCG are selected and combined based on group lasso. The effectiveness of AMS + MFCC + GF + MRCG complementary feature group is verified by DNN in the noise matching and mismatch condition using different training targets of IRM and IBM. The experimental results show that AMS + MFCC + GF + MRCG improve the accuracy of system classification as well as perform well in STOI, PESQ, SDR, and SNR. This paper not only verifies the effectiveness of the group lasso method, but also improves the accuracy of feature extraction and classification in engineering applications. According to the new speech feature model such as MRCG, the feature combination is carried out and good results are obtained. It is more representative of the development trend of speech than the original model and provides a valuable reference for the feature extraction and processing of speech signals.

Abbreviations

DNN	Deep Neural Network
RSS	Residual sum of squares
LASSO	Least Absolute Shrinkage and Selection Operator
LARS	Least angle regression
AMS	Amplitude modulation spectrogram
MFCC	Mel-frequency cepstral coefficient
DRFS	Dynamic correlation-based feature selection
RASTA-PLP	Relative spectral transform perceptual linear prediction
MRCG	Multi-resolution cochleagram
FFT	Fast Fourier transform
GC	Gammatone coefficient
GF	Gammatone frequency
GFCC	Gammatone frequency cepstral coefficient
TIMIT	The DARPA TIMIT Acoustic–Phonetic Continuous Speech Corpus
ASR	Automatic speech recognition systems
LPC	Linear prediction coefficient
LPCC	Linear prediction cepstral coefficient
SNR	Signal-to-noise ratio
IBM	Ideal binary mask
IRM	Ideal ratio mask
STOI	Short-time objective intelligibility
PESQ	Perceptual evaluation of speech quality
SDR	Signal-to-distortion ratio

Acknowledgements

Not applicable.

Authors' contributions

XX and YC have jointly participated in proposing the ideas, discussing the results, and writing and proofreading the manuscript. YC designed the core

methodology of the study and carried out the implementation. RS and DT carried out the implementation of the algorithms and the experiments. All authors read and approved the final manuscript.

Authors' information

Xiaoping Xie received a bachelor's degree in engineering. In 2003, he obtained a master's degree in vehicle engineering from Central South University of Forestry and Technology. In 2014, he received the PhD degree in signal processing from Hunan University. Since 2013, he has been working in the School of Mechanical and Vehicle Engineering, Hunan University, as the Director of the Automotive Laboratory of the School of Mechanical and Vehicle Engineering, where he is currently an associate professor. His research interests include signal processing and analysis, speech separation and conversion, vibration and noise, automotive test methods, and experiments, CAE simulation, and optimization. Yongzhen Chen, Dan Tian, and Rufeng Shen received their bachelor's degrees in engineering in 2016. They are currently pursuing a master's degree at the School of Mechanical and Vehicle Engineering of Hunan University. Their research interests focus on speech signal analysis, automotive NVH, non-stationary signal decomposition, and coherence analysis.

Funding

This work was supported by the Natural Science Foundation of Hunan Province, 2022JJ30147, Features extraction and tracing of unsteady signals in mechanical systems.

Availability of data and materials

All datasets used in this paper are publicly available and include the TIMIT database [22] and NOISEX-92 [23].

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 26 August 2021 Accepted: 6 February 2023
Published online: 16 February 2023

References

1. Z.X. Li, *Research on Single-Channel Speech Separation Method Based on Autoregressive Deep Neural Network, University of Science and Technology of China* (2019), pp.2–4
2. H. Li, *Single-channel Speech Separation Based on Deep Learning, Inner Mongolia University* (2017), pp.5–8
3. E. A. Hoer, Kennard, et al. Ridge Regression: Biased Estimation for Non-orthogonal Problems. Technometrics (1970), pp.55-59
4. R. Tibshirani, Regression shrinkage and selection via the LASSO. J. Roy. Stat. Soc. **73**(1), 273–282 (1996)
5. B. Efron, T. Hastie, J.R. Tibshirani, Least angle regression. Ann. Stat. **32**(2), 407–451 (2004)
6. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. **67**(5), 768 (2005)
7. M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Series. B. Stat. Methodol. **68**(1), 49–67 (2006)
8. J. Friedman, T. Hastie, R. Tibshirani, *A Note on the Group Lasso and a Sparse Group Lasso* (2010), pp.1-8. (Statistics)
9. W. Wang, K. Han, D. Wang, Exploring monaural features for classification -based speech segregation. IEEE Trans. Audio. Speech. Lang. Process. **21**(2), 270–279 (2013)
10. X.H. Wang, L. Qu et al., Speech feature extraction algorithm based on Fisher's ratio Bark wavelet packet transform. J. Xi'an. Polytech. Univ. **30**(4), 452–457 (2016). https://doi.org/10.13338/j.issn.1674-649x.2016.04.008
11. M.H. Wan, G.F. Lu, Feature extraction method for block two-dimensional maximum distance criterion. Small. Microcomputer. Syst. **37**(09), 2088–2092 (2016)
12. Y. Yan, F. Jin, H.X. Lu, A new two-dimensional image feature extraction algorithm. Image. Process. **22**(05), 1008 (2006)
13. X. Zhao, X.D. Chen et al., Speech mixed feature extraction and feature enhancement based on multi-fisher criterion. Nanotechnol. Precis. Eng. **15**(4), 317–322 (2017). https://doi.org/10.13494/j.npe.20160044
14. H.X. Liu, Y.N. Dong, X.H. Qiu, Network flow classification based on correlation feature selection and deep learning. J. Nanjing. Univ. Posts. Telecommunications. (Natural Science Edition). **42**(4), 75–84 (2022). https://doi.org/10.14132/j.cnki.1673-5439.2022.04.011
15. Y.B. Chen, Q.Q. Li, Y.G. Liu, Feature selection algorithm based on dynamic correlation. Comp. Appl. **42**(1), 109–114 (2022)
16. T.F. Lee, P.J. Chao et al., Using multivariate regression model with least absolute shrinkage and selection operator to predict the incidence of xerostomia after intensity-modulated radiotherapy for head and neck cancer. PLoS. One. **9**(2), 700 (2014)
17. A.B. Hussein, M.E. Sinan et al., Modified speech separation deep learning network based on Hamming window. IOP. Conf. Ser. Mater. Sci. Eng. **1076**(1), 1–11 (2021)
18. J. S. Garofolo, L. F. Laurel, et al., DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1–1.1 [J]. NASA STI/Recon technical report n, 93:27403 (1993). DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (1993), https://www.kaggle.com/mfekadu/darpa-timit-acousticphonetic-continuous-speech. Accessed 20 Mar 2013
19. M. David, M. Lavandier et al., Sequential stream segregation of voiced and unvoiced speech sounds based on fundamental frequency. Hear. Res. **344**, 235–243 (2018)
20. M. Li, Research on feature selection methods and algorithms. Comput. Technol. Dev. **23**(12), 16–21 (2013)
21. Y. Shao, D.L. Wang, Robust speaker identification using auditory features and computational auditory scene analysis, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA*. 22(12), 1993-2002 (2008)
22. J. Chen, Y. Wang, D.L. Wang, A feature study for classification-based speech separation at very low signal-to-noise ratio. *IEEE International Conference on Acoustics*. (2014)
23. A. Varga, H.J. Steeneken, Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. **12**(3), 247–251 (1993)
24. J. Chen, Y. Wang, D. Wang, A feature study for classification-based speech separation at low signal-to-noise ratios. IEEE/ACM. Trans. Audio. Speech. Lang. Process. **22**(12), 1993–2002 (2014)
25. X. Zhang, D. Wang et al., Deep learning based binaural speech separation in reverberant environments. IEEE/ACM. Trans. Audio. Speech. Lang. Process **25**(5), 1075–1084 (2017)
26. T. Nguyen, T. Duong et al., Autism blogs: expressed emotion, language styles and concerns in personal and community settings. IEEE Trans. Affect. Comput. **6**(3), 312–323 (2015)
27. T.M. Minipriya, R. Rajavel, Review of ideal binary and ratio mask estimation techniques for monaural speech separation, in *2018 Fourth International Conference on Advances in Electrical, Electronics, Information, Communication*. (2018), pp.1-5

Publisher's Note