Human speech emotion recognition via feature selection and analyzing

Xiangmin Lun¹,², Fang Wang³, Zhenglin Yu¹*
¹College of Mechanical and Electrical Engineering, Changchun University of Science and Technology, Changchun, Jilin, 130000, China
²School of Automation Engineering, Northeast Electric Power University, Jilin, Jilin, 132000, China
³Corresponding author’s e-mail: yuzhenglin@cust.edu.cn

Abstract. Speech emotion recognition is one of the important research topics in the field of multimedia processing and human-machine interface. To obtain the most influential features of the speech data for emotion recognition, in this paper, 64 statistical features of the speech signal including short-term energy, pitch, frame, format, and spectrum energy were extracted with speech emotion database. Mean Impact Value (MIV) and the improved Correlation-based Feature Selection (CFS) were employed to select the most influential feature set. BP neural network (BPNN) was used to identify the accuracy. The proposed MIV-CFS method selected the features related to speech emotion, with less recognition error, the recognition accuracy all higher than 88%, and the highest recognition accuracy is 91.61%.

1. Introduction
Recently, human-computer interaction system is developing rapidly. Thence, multimedia processing has become one of the important research fields of speech recognition technology, speech recognition technology is widely used as its specialized branch. Speech emotion recognition has been applied as in communication with machine ‘using’ emotion, which is why machines are taught to use sound signals to distinguish human emotions.

A large number of speech features including short-term energy [1], pitch[2], formants[3] are required to be extracted in speech emotion recognition. However, not all features can improve speech emotion recognition. Dimensionality reduction such as PCA [4], FDA[5] and KPCA[6] and feature selection such as Relief[7], mRMR[8] and CFS [9] are effective methods to improve accuracy and efficiency of speech recognition. In this paper, we aim to develop a new method to reduce feature dimension and further improve the speech emotion recognition rate.

2. Method
2.1. Feature extraction
In this paper, a set of 64 features including short-term energy, pitch, formant and energy spectrum features were extracted to recognize human speech emotion, which are listed below in details.

Features 1-10: including short-term energy, the maximum, minimum, mean, median value of the energy and variance of its first order differences.
Features 11-26: including pitch range, pitch frequency, the maximum, minimum, mean, median and the variance of its first and second order differences.

Features 27-31: including voiced frames number, unvoiced frames number, the ratio of them, the ratio of voiced frames number and the total frames number, the number of voiced regions.

Features 32-61: including the first, second, third formants, the maximum, minimum, mean, median, and variance of their first and second order differences.

Features 62-64: including the percentage of the spectrum energy below 250Hz, above 650Hz, and above 4KHz.

2.2. Feature selection

2.2.1. Mean Impact Value algorithm.
MIV algorithm is regarded as the best evaluation method for relevance between variables of the neural networks. The input features which could influence the BPNN greatly are chosen as the perfect input vectors[10]. Firstly, the original sample data P was used to train neural network[11]. Secondly, new sample P1 and P2, which were got from original value P by plus or minus 10% respectively to selected feature in P, are employed as the test data. Then we got two results called A1 and A2. D is used to express the difference between A1 and A2 and regarded as the influence of changing this feature in P. Finally, when all the features in P are changed, the MIV of them will be obtained respectively [12]. The features with bigger MIVs will be chosen as the most influential feature set [13].

2.2.2. Correlation based Feature Selection.
CFS calculates the correlation value between every two features and feature-class to select the features related to classes most closely[14], which is calculated by:

\[ M_S = \frac{k \bar{r}_{cf}}{\sqrt{k + k(k-1) \bar{r}_{ff}}} \]  

Where \( M_S \) is the evaluation of the feature subset \( S \) by CFS, \( S \) contains \( k \) features, \( \bar{r}_{cf} \) represents the average correlation between feature \( f(f \in s) \) and category \( c \), \( \bar{r}_{ff} \) represents the average correlation between features and features [15].

3. Experiments and results analysis

3.1. Experiment design
In this paper, we used CASIA Chinese language sentiment database as the source of data in the experiments. This database was recorded by Institute of Automation of Chinese Academy of Sciences which including 4 professional enunciators (2 male and 2 female) speech in a satisfied environment (SNR 35dB). Respectively, it contains 50 sentences in 6 different emotions types (happy, surprise, anger, fear, sad, neutral). Experiment process was shown in Figure 1.

(1) First, a 64-feature statistical vector of each sentence were extracted while corresponding emotion was used as corresponding label. 240 sentence data was randomly selected from 300 sentences as train data, and the remaining 60 sentences were used as test data. BPNN with 64 inputs, 6 outputs and 65 hidden layers were trained and tested.

(2) MIV was employed to reduce feature dimension. In this process, 32 features were selected by MIV in Table 1 while others were discarded because of features’ influential percentage is 0.

(3) According to the feature set in Table 1, 50 sentences of each emotion was used as experimental data while 40 sentences were randomly selected as training data and other 10 sentences were used as testing data. And correct recognition were calculated by BPNN. This procedure was repeated for 100 times and the mean value of correct recognition rates was regarded as the final result of this part.
(4) To prove the effectiveness of MIV feature selection method, LDA (Linear Discriminant Analysis) and MCFS (Multi Cluster Feature Selection) were also employed to reduce feature’s dimension from 64 to 32. Then, features with labels of 300 sentences were used in BPNN classification as the experimental data in this part.

(5) Correlation-Based Feature Selection (CFS) was employed for 32-features dimension reduction. Also, correct recognition after CFS was calculated by BPNN in the previous steps.

3.2. Experiment results

3.2.1. Mean Impact Value algorithm. The selected 32 features by MIV were shown in Table 1, which arrayed by their MIVs. Higher MIVs, higher relationship with human emotional state.

3.2.2. Comparison of dimension reduction methods. Table 2 displayed the average correct recognition rates of selected 32 features by MIV, which is higher than original 64 features. For further verification of dimension reduction effectiveness of MIV, LDA [11] and MCFS [6] are employed in this experiment. In table 2, the recognition results of selected 32 features by MIV > the recognition results of selected 32 features by LDA > the recognition results of selected 32 features by MCFS. Also, the correct recognition rates of LDA are similar to MIV in all the emotional state except “Sad”, which is 67.94% and much lower than MIV. The correct recognition rate of MCFS fluctuates greatly also unstable. Although the highest accuracy rate is 91.89% of "happy", the lowest is only 46.05% of "fear".

Table1 The selected 32 dimension features and their mean impact values

| Feature order | Feature                              | MIV   | Feature order | Feature                              | MIV   |
|---------------|--------------------------------------|-------|---------------|--------------------------------------|-------|
| 11            | Maximum value of pitch               | 9.11% | 25            | Variance of pitch's second order difference | 2.06% |
| 22            | Minimum value of pitch's second order difference | 6.87% | 20            | Variance value of pitch's second order difference | 2.06% |
| 21            | Maximum value of pitch's second order difference | 6.53% | 15            | Variance value of pitch               | 2.06% |
| 16            | Maximum value of pitch's first order difference | 6.19% | 28            | Unvoiced frames                       | 1.89% |
| 13            | Mean value of pitch                  | 6.19% | 36            | Variance value of the first formant    | 1.72% |
| 14            | Median value of pitch                | 6.01% | 32            | Maximum value of the first formant     | 1.03% |
| 26 | Range of pitch | 5.84% | 30 | Ratio of voiced frames and total frames | 1.03% |
| 24 | Median value of pitch's second order difference | 5.82% | 56 | Variance value of the second formant's first order difference | 0.68% |
| 19 | Median value of pitch's first order difference | 5.82% | 51 | Variance value of the first formant's first order difference | 0.52% |
| 23 | Mean value of pitch's second order difference | 5.33% | 12 | Minimum value of pitch | 0.52% |
| 27 | Voiced frames | 4.98% | 64 | Percentage of the energy spectrum above 4KHz | 0.17% |
| 18 | Mean value of pitch's first order difference | 4.98% | 38 | Minimum value of the second formant | 0.17% |
| 17 | Minimum value of pitch's first order difference | 4.98% | 37 | Maximum value of the second formant | 0.17% |
| 41 | Variance value of the second formant | 2.41% | 35 | Median value of the first formant | 0.17% |
| 46 | Variance value of the third formant | 2.34% | 34 | Mean value of the first formant | 0.17% |
| 61 | Variance value of the third formant's first order difference | 2.06% | 33 | Minimum value of the first formant | 0.17% |

Table 2 Recognition results comparison of feature dimension reduction methods

| Algorithm | Happy | Angry | Surprise | Sad | Fear | Neutral | Average |
|-----------|-------|-------|----------|-----|------|---------|---------|
| Original 64 features | 80.49% | 80.36% | 77.34% | 93.48% | 74.49% | 86.58% | 82.12% |
| Selected 32 features by MIV | 80.85% | 89.83% | 77.08% | 89.78% | 84.32% | 89.61% | 85.24% |
| Selected 32 features by LDA | 83.00% | 88.12% | 78.30% | 67.94% | 89.60% | 89.63% | 82.77% |
| Selected 32 features by MCFS | 91.89% | 88.39% | 91.22% | 76.42% | 46.05% | 73.45% | 77.9% |
| Selected 10 features by MIV | 99.19% | 43.15% | 90.15% | 89.95% | 41.24% | 85.08% | 74.79% |

3.2.3. Selecting features using CFS. From Table 1, pitch features, formant features and voiced frames features are all important in emotion recognition. In contrast, short-term energy features and spectrum energy features are not so closely related to speech emotions. Thence, we plan to further select features including pitch features (15 ones), formant features (5 ones) and voiced frames features (3 ones) according to Table 1 by MIV.

We reduced 15 pitch features to 3-8 ones, and 5 formant features to 2-3 ones. Then, the selected features will be combined with 3 voiced frames features (27th, 28th and 30th) to compose a feature set. There were totally 8 to 14 features in this set, which was used as the speech emotion recognition data source.

Feature selection results including feature numbers and feature orders were listed in Table 3. All feature combinations and recognition rates were shown in Table 4. The combination form is “pitch features + voiced frames features + formant features”. For example, “3+3+2” means feature combination is “21st + 22nd + 25th + 27th + 28th + 30th + 41st + 36th” according to Table 3.
Table 3 Feature selecting results using CFS

| Selected feature number | Selected features’ orders |
|-------------------------|---------------------------|
| about pitch             |                           |
| 3                       | 21,22,25                 |
| 4                       | 11,22,20,26              |
| 5                       | 18,19,20,23,24           |
| 6                       | 14,15,17,22,23,24        |
| 7                       | 13,14,15,19,20,25,26     |
| 8                       | 15,17,18,19,20,21,24,25 |

| Selected feature number | Selected features’ orders |
|-------------------------|---------------------------|
| about formant           |                           |
| 2                       | 41,36                     |
| 3                       | 41,36,32                  |

Table 4 Recognition results of different feature combination

| Feature combination | Accuracy  | Feature combination | Accuracy  |
|---------------------|-----------|---------------------|-----------|
| 3+3+2               | 88.70%    | 3+3+3              | 89.68%    |
| 4+3+2               | 89.28%    | 4+3+3              | 89.82%    |
| 5+3+2               | 90.32%    | 5+3+3              | 90.46%    |
| 6+3+2               | 90.56%    | 6+3+3              | 90.67%    |
| 7+3+2               | 90.62%    | 7+3+3              | 91.15%    |
| 8+3+2               | 91.61%    | 8+3+3              | 90.85%    |

From table 4, the correct recognition rate were all higher than 88%, which is satisfactory. The recognition errors of 60 randomly selected sentences were shown in Figure 2, 13 errors in 64 features without feature selection, 15 errors in 32 features selected by LDA, 10 errors in 32 features selected by MIV. However, only 4 errors in 32 features selected by MIV and CFS. It is clear that the feature set selected by MIV and CFS is more satisfactory for speech emotion recognition than the other methods mentioned in this paper.

Figure 2 Misclassification before and after experiment
4. Conclusions
The results of MIV and CFS feature extraction and selection were compared with LDA and MCFS. From the comparison results, we can draw the following conclusions. Among the 32 features selected by MIV, the pitch feature is at the forefront, but relying solely on the pitch feature without combining other features (such as voiced frames and formants) cannot obtain a good recognition result. The recognition rate of feature extraction with MIV is higher than that with LDA and MCFS. After further feature selection using CFS on the extracted features, the recognition rate has improved significantly. The selected features mainly include 13 features related to the pitch and formants.

References
[1] Prukkanon, N., Chamnongthai, K., & Miyanaga, Y. (2016). F0 contour approximation model for a one-stream tonal word recognition system. AEUE - International Journal of Electronics and Communications, 70(5), 681-688.
[2] Vegesna, V. V. R., Gurugubelli, K., & Vuppala, A. K. (2018). Prosody modification for speech recognition in emotionally mismatched conditions. International Journal of Speech Technology, 21(3): 521-532.
[3] Michael, A, Kraus, S. (2019). Basic auditory processing deficits and their association with auditory emotion recognition in schizophrenia. Schizophrenia Research, 204, 155-161.
[4] Razzak I, Saris R A. (2020). Integrating joint feature selection into subspace learning: a formulation of 2dpca for outliers robust feature selection. Neural Networks, 121, 441-451.
[5] Patridge, E. V., Gareiss, P. C., Kinch, M. S., & Hoyer, D. W. (2015). An analysis of original research contributions toward fda-approved drugs. Drug Discovery Today, 20(10), 1182-1187.
[6] Neffati, S, Ben Abdellafou, K., Taouali, O., & Bouzrara, K. (2019). A new bio-cad system based on the optimized kpca for relevant feature selection. The International Journal of Advanced Manufacturing Technology, 102(1-4): 1023-1034.
[7] Dalai S, Chatterjee B, Dey D. (2012). Rough-set-based feature selection and classification for power quality sensing device employing correlation techniques. IEEE Sensors Journal, 13(2): 563-573.
[8] Ahmad, A. O., Tajudin, K. A., Azmi, A. B. M., & Awadallah, M. A. (2018). A novel gene selection method using modified mrmr and hybrid bat-inspired algorithm with β -hill climbing. Applied Intelligence, 48, 4429-4447.
[9] Liu, X., & Tang, J. (2014). Mass classification in mammograms using selected geometry and texture features, and a new svm-based feature selection method. IEEE Systems Journal, 8(3), 910-920.
[10] Meng, Q., Li, K., & Zhao, C. (2019). An improved particle filtering algorithm using different correlation coefficients for nonlinear system state estimation. Big Data, 7(2), 114-120.
[11] Swain, M., Sahoo, S., Routray, A., Kabisatpathy, P., & Kundu, J. N. (2015). Study of feature combination using hmm and svm for multilingual odia speech emotion recognition. International Journal of Speech Technology, 18(3), 387-393.
[12] Bolón-Canedo, V., Porto-Díaz, I., Sánchez-Maroo, N., & Alonso-Betanzos, A. (2014). A framework for cost-based feature selection. Pattern Recognition, 47(7), 2481-2489.
[13] Hautamäki R G, Kinnunen T, Hautamäki V. (2015). Automatic versus human speaker verification: the case of voice mimicry. Speech Communication, 72, 13-31.
[14] Kerkeni L, Serrestou Y, Raof K, (2019). Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. Speech Communication, 114: 22-35.
[15] Rajisha, T. M., Sunija, A. P., & Riyas, K. S. (2016). Performance analysis of malayalam language speech emotion recognition system using ann/svm. Procedia Technology, 24, 1097-1104.