Advances in Intelligent Systems Research (AISR), volume 160
3rd International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2018)

Voice Quality Assessment Method Based on Contribution Degree Analysis

Zhangjian Xuan and Xiaoxia Cai
School of Electronic Defence, National University of Defense Technology, Hefei 230037, China

***Abstract*—In objective assessment methods of speech quality, Mel cepstrum coefficients are usually used as speech feature parameters. However, the effect of the Mel cepstrum order and the component contribution were not involve in the previous research. After determining the optimal order, the contribution of the Mel cepstrum coefficient is studied in combination with the increasing or decreasing component method. The lower Important Mel Frequency Cepstrum coefficient (IMFCC) was obtained after the selection and recombination of the feature parameters. After testing, the assessment effect is improved and the prediction error is reduced.**

Keywords—speech quality assessment; Mel cepstrum coefficient; increase and decrease component method; contribution degree; distortion distance

I. INTRODUCTION

Voice quality assessment refers to a comprehensive overall auditory experience of human voice signal. It is divided into two categories according to the evaluation type: subjective assessment and objective assessment [1]. The subjective assessment methods mainly include the Mean Opinion Score (MOS) method [2] and the articulation index (AI). The objective assessment methods mainly include spectrum measurement distortion distance method and neural network method. However, in practical applications, the subjective assessment methods such as the MOS method require a lot of manpower and material resources. Therefore, researching realistic and effective objective methods of speech quality assessment have became the trends in the field of speech quality assessment.

Objective speech quality assessment began in the 1940s, but it did not begin to develop until the 1970s due to technical limitations. In 1980, Davis and Mermelstein proposed the theory of Mel cepstrum. In 1993, Picone J. W. proposed a speech quality assessment method based on Mel cepstrum coefficients [3]. Cubic transformation was used in document [4] to replace the logarithmic transformation in the extraction process of Mel cepstrum coefficients. Mel cepstrum coefficients were used in document [5] to evaluate the effect of speech interference on ultra-short wave communication interference systems. However, these studies didn't involve the influence of the Mel cepstrum order and the contribution of the component on the speech quality assessment.

II. TRADITIONAL MEL CEPSTRUM COEFFICIENT EXTRACTION METHOD

The correspondence between Mel scale and frequency scale is shown in Equation (1). The process of traditional Mel cepstrum coefficient extraction mainly includes the following aspects.

$$Mel(f) = 2595\lg(1+\frac{f}{700}) \tag{1}$$

A. Frequency Domain Processing

Frequency domain processing includes pre-emphasis, frame division, windowing, and frequency domain analysis. Pre-emphasis is a signal processing method that compensates for high-frequency components of the input signal. Considering that the speech signal approximates a short-time stationary process in the range of 10-30 ms, the speech signal is usually set to a frame of 10-30 ms. The window function usually selects the Hamming window to reduce the spectral leakage caused by the framing. The signal power spectrum is obtained by performing a frequency domain transform on the preprocessed signal through a fast Fourier transform

B. Mel Filtering

The signal power spectrum is simulated by the human auditory audible sound through a set of *N* Mel filters (*N* values take 20-40). The mathematic expression of the Mel filter is given in Equation (2):

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

C. Logarithmic transformation

The Mel-filtered signal is converted to intensity-loudness to obtain logarithmic energy.

D. Discrete Cosine Transform (DCT)

To eliminate the correlation between the characteristic parameters, a discrete cosine transform is required, as shown in Equation (3). C_i is the i-th component of the feature parameter.

$$C_i=\sqrt{\frac{2}{\pi}}\sum_{j=1}^{N}s_j\cos(\frac{\pi i}{N}(j-0.5)) \tag{3}$$

E. Liter Sine Cepstrum Boost

The author in [6] found that the high-dimensional feature parameter components have a better and better grip on the noise environment. The half-liter sine cepstrum improves the proportion of low-dimensional components and improves the effect of high-dimensional components. Half-liter sine cepstrum boost is shown in Equation (4):

$$w(i) = 0.5 + 0.5\sin(\frac{\pi i}{N}), 1 \le i \le N \tag{4}$$

After the extraction of MFCC, the contribution of MFCC dimension is studied.

III. IMPROVED MEL CEPSTRUM COEFFICIENT EXTRACTION METHOD

A. Performance Indicators of Voice Quality Assessment Results

The effect of speech quality assessment is usually measured by the Pearson coefficient ρ and prediction error σ, as shown in Equation (5) and Equation (6). ρ describes the degree of linearity between the objective assessment and the subjective assessment. The correlation coefficient takes a value of 0-1. The higher the value, the closer the objectively predicted MOS value is to the subjective MOS value. σ is usually represented by the standard estimation deviation. The smaller the prediction error, the more accurate the speech quality assessment is.

$$\rho = \frac{\sum_{i=1}^{N}(MOS_o(i) - \overline{MOS_o})(MOS_d(i) - \overline{MOS_d})}{\sqrt{\sum_{i=1}^{N}(MOS_o(i) - \overline{MOS_o})^2 \sum_{i=1}^{N}(MOS_d(i) - \overline{MOS_d})^2}} \tag{5}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(MOS_o(i) - MOS_d(i))^2}{N}} \tag{6}$$

In the experiment, the correlation coefficient is selected as the contribution index.

B. MFCC Order Selection

Under the condition that the number of Mel filters is limited to 24, the MFCC order selection is studied, as shown in Fig. 1.

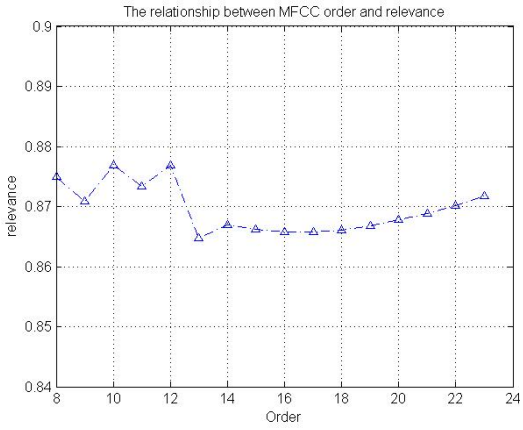

FIGURE I. RELATIONSHIP BETWEEN MFCC ORDER AND CORRELATION DEGREE

It can be seen that when the order of MFCC is between 8 and 12, the characteristic parameters are greatly affected by the noise factors due to the lower order, and the correlation coefficient fluctuation range is 1%. When the order is from 13 to 23, the overall correlation coefficient changes smoothly and the fluctuation range is between 0.6%, but the overall value is lower than the 8th to 12th order. As the order increases or decreases, the correlation coefficient increases slowly. This is because as the order of MFCC increases, the speech information carried by the characteristic parameters increases, but the amount of computation increases with it.

According to Fig.1, the Mel cepstrum coefficient order is selected to be 12 steps.

C. Contribution Analysis

As shown in Equation (7), the increase or decrease component method is a method of calculating the contribution degree of different cepstrum components [7].

$$R(i) = \frac{1}{n}[\sum_{j>i}(p(i,j) - p(i+1,j)) + \sum_{j<i}(p(j,i) - p(j,i-1))] \tag{7}$$

$R(i)$ Indicates the degree of contribution of the *i*-th order component of the feature parameter to the speech quality assessment effect. $p(i,j)$ represents the evaluation effect of the characteristic parameters composed of the cepstrum components of the *i*-th and *j*-th orders. The contribution of MFCC, the first-order and second-order differential MFCC, to the speech quality assessment is shown in Figures 2-4.

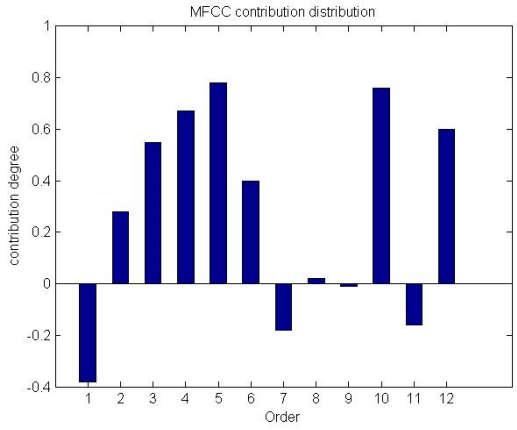

FIGURE II. MFCC CONTRIBUTION DEGREE DISTRIBUTION

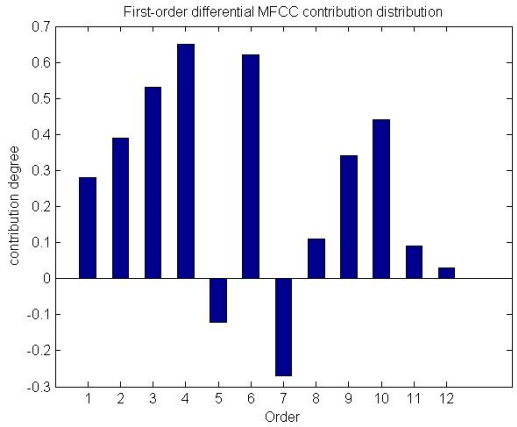

FIGURE III. HE FIRST-ORDER DIFFERENTIAL MFCC CONTRIBUTION

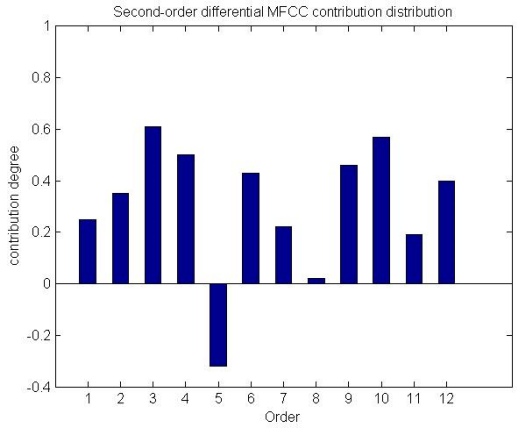

FIGURE IV. THE SECOND-ORDER DIFFERENTIAL MFCC CONTRIBUTION

The component with negative contribution in the feature parameter is removed. MFCC and the first-order and the second-order difference parameters are selected again to form a 25-dimensional MFCC fusion.

IV. EXPERIMENTAL ANALYSIS

The experiment used the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) speech data set. Ten different speech samples were selected from the data set for experimentation. The experiment uses Noisex-92 noise signal library. The experiments selected babble (street mixed vocal), f16, and white noises. The signal to noise ratio ranged from -25dB to 10dB. The voice sample consists of 20 hearing staff (10 men and 10 women, no hearing disease, aged 20-35 years). In a quiet environment, voice signals are scored on a scale of 0-5, and the arithmetic average is performed. The calculation of the spectral distortion measure is shown in Equation (8)

$$D(k) = \sqrt{\sum_{i=1}^{m} [C_o(i,k) - C_d(i,k)]^2} \qquad k = 1, 2, \ldots, N \tag{8}$$

$D(k)$ represents the distortion distance of the *k*-th frame of a pair of speech signals, and $C_o(i,k)$ represents the *i*-th Mel cepstrum coefficient of the *k*-th frame of the original signal. $C_d(i,k)$ denotes the *i*-th Mel cepsrrum coefficient of the first frame of the distorted signal. A non-linear mapping relation of the subjective assessment of the MOS score and the spectrum distortion measure is performed. Usually, the quadratic curve is used to fit the spectral distortion measure and the MOS value according to the least squares criterion. The MOS value is fitted to the normalized distortion distance. The subjective and objective MOS scores are fitted as shown in Figure 5(a) to Figure 5(f).

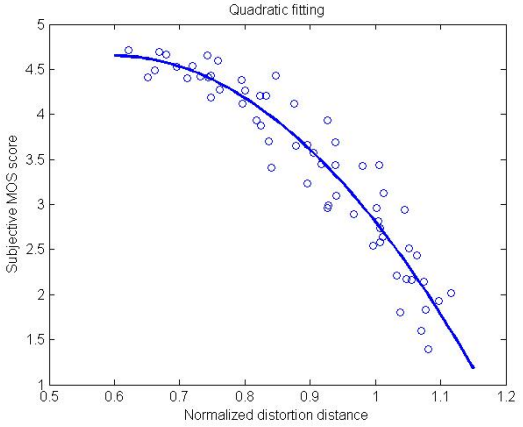

FIGURE V. (A) QUADRATIC FITTING UNDER WHITE NOISE

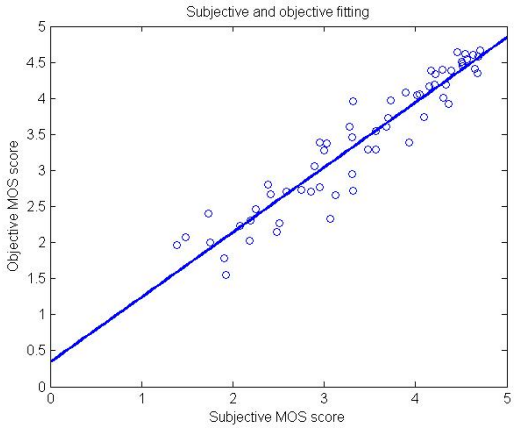

FIGURE V. (B) FITTING OF SUBJECTIVE AND OBJECTIVE MOS VALUES UNDER WHITE NOISE

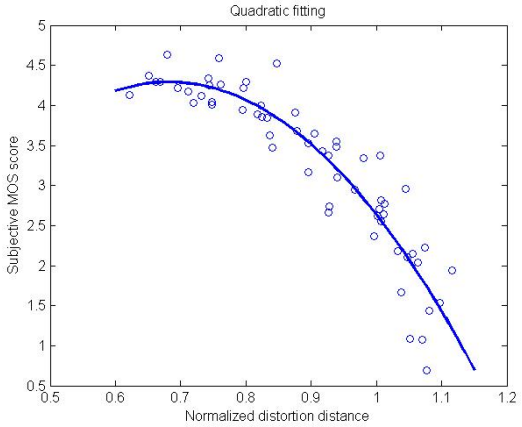

FIGURE V. (C) QUADRATIC FITTING UNDER F16 NOISE

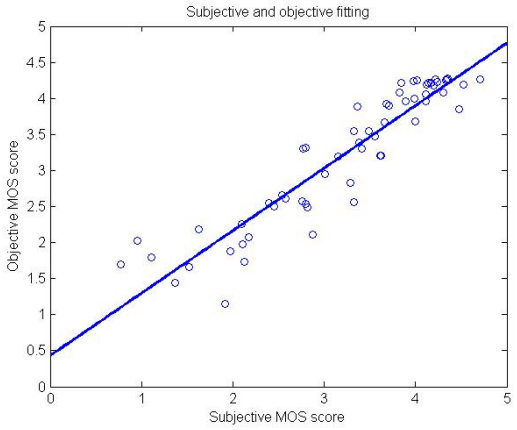

FIGURE V. (D) FITTING OF SUBJECTIVE AND OBJECTIVE MOS VALUES UNDER F16 NOISE

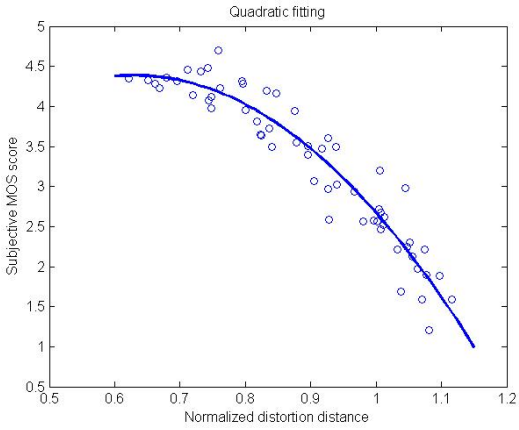

FIGURE V. (E) QUADRATIC FITTING UNDER BABBLE NOISE

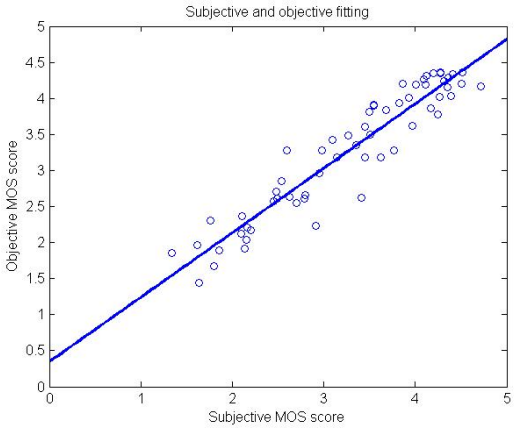

FIGURE V. (F) FITTING OF SUBJECTIVE AND OBJECTIVE MOS VALUES UNDER BABBLE NOISE

The figure shows that in the same SNR range, the overall score of speech quality is: babble noise <white noise <f16 noise. At the same time, the white noise-disturbed speech is evenly distributed among the scores, and the voice quality scores under the interference of f16 noise and babble noise are more fragmented. The analysis shows that this situation is mainly due to babble noise causing speaker recognition interference to the hearing person, while f16 noise and white noise are mechanical noise and natural noise, and the auditory effect and human voice are quite different, so the score is divided. The value is higher. The Pearson correlation coefficient and the prediction error of the original algorithm and the improved algorithm are compared, as shown in Table 1.

TABLE I. COMPARISON OF MFCC AND OUR ALGORITHM IN NOISY ENVIRONMENT

Algorithm	Babble		F16		White	
	ρ	σ	ρ	σ	ρ	σ
MFCC	0.8543	0.3008	0.8719	0.2023	0.8756	0.2166
IMFCC	0.8925	0.1854	0.8912	0.1853	0.8894	0.1914

V. SUMMARY

The objective speech quality assessment algorithm based on improved MFCC cepstrum parameters selects the optimal order by studying the influence of MFCC order on the evaluation effect. After comparing the original algorithm with the improved algorithm under the three noise conditions, the reconstructed IMFCC improves the correlation between the objectively predicted MOS value and the subjective MOS value, reduces the prediction error, and obtains better results. It can be used as an objective voice effective method of quality assessment.

REFERENCES

[1] Chen Guo, Hu Xiulin, Zhang Yunyu, et al. Research Progress on Objective Evaluation Methods of Speech Quality[J]. Chinese Journal of Electronics, 2001, 29(4): 548-552.

[2] TP Barnwell , A Bush Statistical correlation between objective and subjective measures for speech quality.Proc. IEEE ICASSP,1978:586-590

[3] Picone J W. Signal modeling techniques in speech recognition[J]. Proceedings of the IEEE, 1993, 81(9):1215-1247.

[4] Chen Mingyi, Sun Dongmei, He Xiaoyue. Research on Speech Quality Assessment Based on Improved MFCC Speech Feature Parameters[J]. Journal of Circuits and Systems, 2009, 14(3):111-116.

[5] Zhao Lingwei, Zhang Lei. Research on Evaluation Method of Speech Interference Effect Based on Mel Scale[J]. RadioEngineering, 2017, 47(2):32-35.

[6] Yang Ruitian, Zhou Ping, Yang Qing. TEO Energy and Mel Cepstrum Hybrid Parameters for Speaker Recognition[J]. Computer Simulation, 2017, 34(8): 215-219.

[7] Hu Hang. Modern speech signal processing [M]. Beijing. Electronic Industry Press, 2014.