A Speech Quality Evaluation Method Based on Auditory Characteristic

Qingxian Li, Liangjiang Liu, Xianyu Zhu
National Center for Urban Energy Measurement (Hunan)
Hunan Institute of Metrology and Test
Changsha, China
zhuxianyu@foxmail.com

Xin Bian, Xin Zhou
Division of Electronics and Information Technology
National Institute of Metrology
Beijing, China
bianx@nim.ac.cn

***Abstract*—In this paper, an objective measurement for speech quality evaluation based on auditory characteristics is proposed. Firstly, the features are extracted by the Gammatone auditory filter bank to calculate the Gammatone Frequency Cepstrum Coefficient (GFCC) of the original speech and the distortion. After getting the GFCC, the average distortion distance is obtained. Then, in order to get an objective estimator for the objective Mean Opinion Score (MOS), it is necessary to map the normalized average distortion distance to the MOS scale. Experiments show that the proposed algorithm can greatly reduce the algorithm complexity. At the same time, the relevancy between the subjective MOS and the objective MOS is close to the Perceptual Evaluation of Speech Quality (PESQ). It is proved that the new algorithm is useful for the real-time online monitoring of the speech quality.**

Keywords—speech quality; objective evaluation; GFCC

I. INTRODUCTION

Compared with telephone, the speech quality of mobile communication system has some congenital defects, because of the impact from the complex and variable wireless environment, which requires the mobile-phone operators to pay more attention to the speech quality than telephone and take more effective method to evaluate that [1]. The current mobile communication charging standard only depends on the user's talk time without consideration of communication quality. If both of them are taken into consideration the charging standard which not only depends on the talk time but also on majority of users' subjective perception and satisfaction degree will become more flexible. Moreover, it will have significant implications for the mobile communication operators to enhance the market competitiveness and further improve the quality of service. Therefore, it is very important to carry out relevant research on speech quality evaluation.

There are two main types of assessments on the speech quality: subjective evaluation and objective evaluation. Subjective evaluation means individuals directly evaluate the quality of the speech. Although it is complex, this evaluation is a real reflection of the speech quality since people are the ultimate recipient of the voice. The Mean Opinion Score (MOS) proposed by ITU in 1996 is widely used in subjective evaluation and is the arithmetic mean of the testers' given scores to directly reflect listeners' perception on the speech quality. One advantage of the subjective evaluation is that it can reflect cognition of people. However, it is time-consuming, expensive and lacks flexibility, repeatability and stability, and it is easily affected by subjective perception. In order to overcome the disadvantages of subjective evaluation, people begin to study the objective evaluation on speech quality. The purpose of doing research into the objective evaluation is not to substitute subjective evaluation, but to make objective evaluation a convenient and accurate way to predict the subjective MOS value [2].

In this paper, an objective speech quality evaluation method based on auditory characteristics is presented. The Gammatone filter bank that accords with human auditory system is used to extract characteristic parameters. And then GFCC of original voice and distorted voice are calculated to find out the average distortion distance, and the mapping relationship between subjective MOS and normalized average distortion distance is established. Finally, the objective MOS value can be obtained from this mapping relationship and the performance of the algorithm can be compared with PESQ.

II. FEATURE EXTRACTION

The study of human ear physiology shows that human auditory system is composed of the external ear, middle ear and inner ear. Speech signals pass though the external ear, middle ear and inner ear in turn and go into the auditory central system after the decomposition of cochlear basilar membrane. Cochlear is the key component of the whole system. When the speech signal is introduced into cochlear basilar membrane, the basilar membrane will generate vibration in the form of traveling wave, and the acoustic response of the basilar membrane is related to the frequency of speech signal. The frequency decomposition of the basilar membrane is an important part of the speech signal processing in the human auditory system [3]. In this paper, we use the Gammatone filter bank to simulate the cochlear model and extract the characteristic parameters that corresponds to subjective perception of the human ear.

A. Gammatone Filter

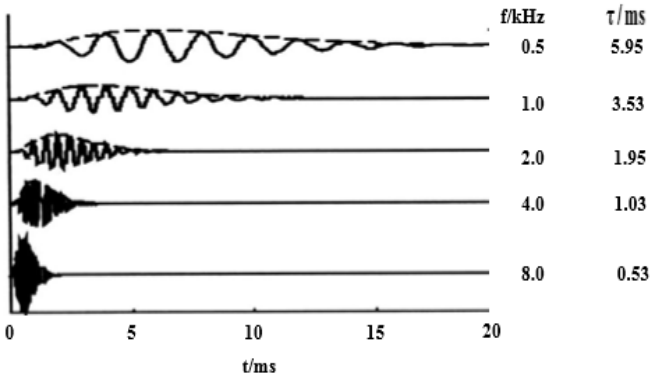
The Gammatone filter can stimulate the filter characteristics of the basilar membrane. The filter's impulse response in time domain is:

This paper was supported by National Sci-Tech Support Plan of China under Grant NO.2014BAK02B05.

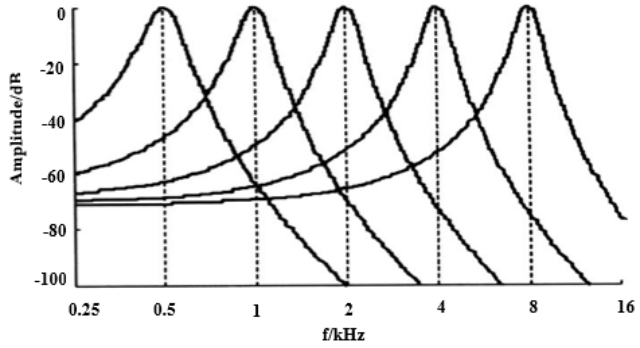
$$g(t) = B^n t^{n-1} e^{-2\pi Bt} \cos(2\pi f_0 t + \varphi) u(t) \quad (1)$$

When $t < 0$, $u(t) = 0$ and $t > 0$, $u(t) = 1$. $B = b_1 ERB(f_0)$, where $ERB(f_0)$ is the equivalent rectangular bandwidth for Gammatone filter(the width of the specified filter through the same energy of the rectangular filter for the same white noise input, referred to as ERB), its relationship with the central frequency of the Gammatone filter is $ERB(f_0) = 24.7 + 0.108 f_0$. The parameter $b_1 = 1.019$ is introduced to make the function more consistent with the physiological data, n is the filter order, when n=4, the Gammatone filter can simulate the filter characteristics of the basilar membrane. parameter φ is initial phase of filter [4].

By conducting the Fourier transform on the impulse response of the Gammatone filter in time domain, frequency response characteristics can be obtained. The time-domain impulse response of the 4 order Gammatone filter with different center frequencies is shown in Fig. 1(a), the corresponding amplitude frequency response curves are shown in Fig. 1(b). The dashed line in Fig. 1(a) indicates the envelope of the time-domain waveform of the Gammatone filter. The value τ represents the time required to reach the maximum value for the vibration envelope from t=0. The amplitude frequency response curve in Fig. 1(b) is based on the logarithmic frequency in the X-coordinate.



(a) time-domain waveform



(b) frequency-domain waveform

Fig. 1. The time-frequency response of the Gammatone filter

From Fig. 1(a), we can find that the Gammatone filter in time domain has the following characteristics: the vibration frequency of the waveform is equal to the center frequency and the envelope is the Gamma function. The higher the center frequency, the shorter the time τ is needed to reach the maximum amplitude (e.g., the maximum envelope). These features are very consistent with the physiological impulse response of the auditory nerve. From Fig. 1(b), we can find that the Gammatone filter is a band pass filter that the maximum amplitude appears to be in the center of each curve and the Gammatone filter with different center frequency has different bandwidth. Besides, it has a sharp frequency selective characteristic because of the steep edges besides the center frequency. The amplitude frequency response accords with the filtering features of the basilar membrane.

B. GFCC

In this paper, the filter bank consisted of 64 Gammatone filters is used to simulate the human auditory model. The cepstrum feature parameters of the model based on the Gammatone filter bank are recorded as GFCC (Gammatone Frequency Cepstrum Coefficient). The detail process of GFCC feature extraction is shown in Fig. 2.

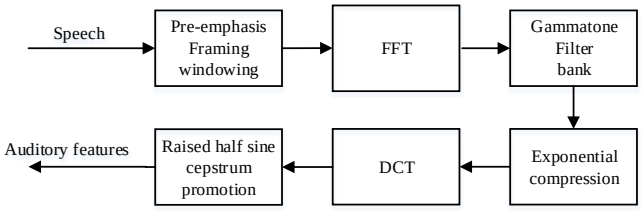


Fig. 2. GFCC feature extraction

It includes 6 steps:

1) Pre-emphasis, framing, and windowing. In order to emphasize high frequency signal, the input high frequency signal is firstly pre-emphasized with the coefficient 0.97. We assume that x(n) is the original speech signal, then the dealt signal y(n) is:

$$y(n) = x(n) - 0.97 \times x(n-1) \quad (2)$$

According to the short-time stationary of the speech signal, it is divided into several frames, the length of each frame is 256 sampling points and the frame shift is 50%.

In order to reduce the edge effect of the speech frame, the speech signal is windowed by Hamming. The Hamming window is given in (3), and the windowed signal $s_w(n)$ can be got by (4).

$$w(n) = \begin{cases} 0.54 - 0.46 \times \cos(\dfrac{2\pi n}{N-1}), & if \quad n = 0,1,\cdots,N-1 \\ 0 & otherwise \end{cases} \quad (3)$$

$$s_w(n) = y(n) \times w(n) \tag{4}$$

2) FFT (Fast Fourier Transform). To get the discrete power spectrum $X(k)$, the windowed signal is transformed from the time-domain into the frequency-domain by the FFT.

3) Filtering by Gammatone filter bank. The energy spectrum is obtained by the square of the power spectrum $X(k)$, and then the Gammatone filter is used for filtering.

4) Exponential compression. The output of each filter is compressed in an exponential form, and we can get a set of energy spectrum $m_1, m_2, \cdots, m_P$.

$$m_i = \sum_{k=1}^{N} [X(k)^2 \times H_i(k)]^{e(f)} \tag{5}$$

In the equation, $e(f)$ is the exponential compression value.

5) DCT (Discrete Cosine Transform). The GFCC is obtained by applying DCT on the energy spectrum after the exponential compression in (6).

$$G_{GFCC}(i) = \sqrt{\frac{2}{N}} \sum_{j=1}^{P} m_j \cos[\frac{\pi i}{P}(j-0.5)], i = 1, 2, \cdots, M \tag{6}$$

M is the dimension of GFCC features, P is the number of filters.

6) Half raised sine cepstrum promotion. The features after DCT are enhanced by the half raised sine window in (7) and (8) [5].

$$w(i) = 0.5 + 0.5 \times \sin(\pi i / N), 1 \le i \le N \tag{7}$$

$$C_{GFCC}(i) = C_{GFCC}(i)_i \times w(i) \tag{8}$$

III. Distortion Calculation and Mos Mapping

We use the two order norm distance as the distortion measure. Distortion distance of the k frame is:

$$D(k) = \sqrt{\sum_{i=1}^{m} [C_o(i,k) - C_d(i,k)]^2}, \quad k = 1, 2, \cdots, N_f \tag{9}$$

In this equation, $C_o(i,k)$ is the coefficients in the k frame with i order of the original speech signal. $C_d(i,k)$ is the coefficients in the k frame with i order of the distorted speech signal. N_f is the total number of frames. m is the highest order of GFCC.

After calculating the distortion distance of each frame, the average distortion distance of all frames is calculated and is used as the total distortion degree.

The last step of the algorithm is to predict the MOS value, and the average distortion distance is fitted by a second order polynomial according to least square method in order to correspond to the objective MOS of the speech quality, that is, the predicted MOS value. The predictive function is as follows:

$$S = \alpha D^2 + \beta D + \gamma \tag{10}$$

Here S is the predicted MOS value, D is the average distortion distance. α, β and γ are the parameters supposed to be calculated, which is determined by the regression function of the subjective test data. And the regression function is obtained by testing a large number of speech file pairs and each pair contains one distorted speech file used to evaluate and one corresponding original speech file with known MOS.

IV. Experiments and Analysis

The speech database used in the experiment is derived from ITU- T Rec.P.Sup23 [6]. The speech is divided into seven categories, including English, French, Japanese and Italian. We select 96 pairs of known subjective MOS score at random to test the speech, and the map between the objective MOS and the normalized average distortion distance can be established with the curve-fitting method by using the relationship between average distortion distance and subjective MOS in each speech pair. It can be seen from Fig. 3 that the conic relationship between the normalized average distortion distance and the subjective MOS score is presented. Therefore, using the second order polynomial, we can get the fitted equation:

$$MOS_o = 4.5402 \times D^2 - 5.9636 \times D + 4.5755 \quad 0 \le D \le 1 \tag{11}$$

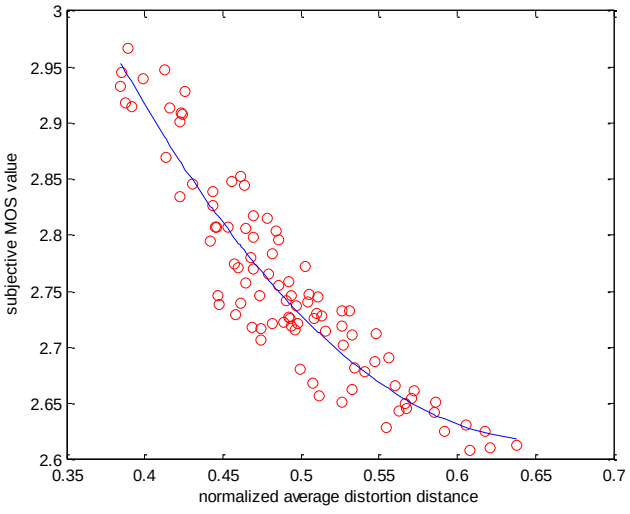


Fig. 3. The relationship between normalized average distortion distance and subjective MOS value

In order to evaluate the performance of the speech quality method, we use the correlation degree and prediction error between the predicted MOS value and the subjective MOS value as the performance index, and the correlation degree is described by the Pearson correlation coefficient ρ in (12). Correlation coefficient describes the linear degree between the objective evaluation and the subjective evaluation of the MOS, the closer the correlation coefficient is to 1, the closer the predicted MOS value is to the subjective MOS value.

$$\rho = \frac{\sum_{i=1}^{N}(MOS_o(i)-\overline{MOS}_o)(MOS_s(i)-\overline{MOS}_s)}{\sqrt{\sum_{i=1}^{N}(MOS_o(i)-\overline{MOS}_o)^2\sum_{i=1}^{N}(MOS_s(i)-\overline{MOS}_s)^2}} \quad (12)$$

The prediction error is expressed by the standard deviation σ in (13). The smaller the σ, the smaller the prediction error. Meanwhile, the better the performance of the objective evaluation.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(MOS_o(i)-MOS_s(i))^2}{N}} \quad (13)$$

In (12) and (13), $MOS_o(i)$ is the objective MOS value of the i pair. $MOS_s(i)$ is the MOS of subjective evaluation, N is the number of data points [7].

Fig. 4 shows the relationship between the subjective MOS value and the predicted MOS value by using ITU speech database. Through the experimental data we can find that the correlation degree ρ between the predicted MOS value of this algorithm and the subjective MOS value is 0.9284, the estimated deviation σ is 0.033. The correlation degree between the objective MOS value of PESQ and the subjective MOS value is 0.935 [8]. In this paper, the complexity of the algorithm is greatly reduced, and at the same time, the correlation degree between subjective MOS value and the objective value is very close to PESQ.

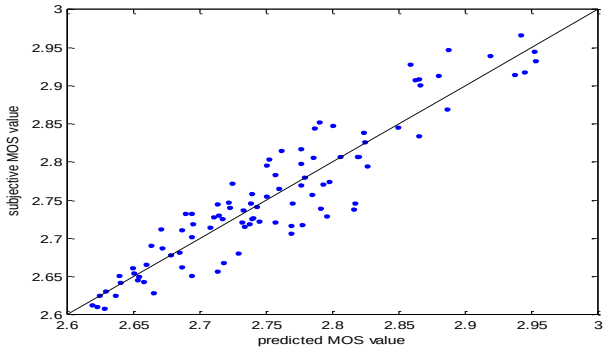


Fig. 4. The relationship between subjective MOS value and predicted MOS value

V. CONCLUSION

In this paper, a speech quality evaluation method based on auditory characteristics is presented. The Gammatone filter bank that accords with human auditory system is used to extract characteristic parameters. And then GFCC of original voice and distorted voice are calculated to find out the average distortion distance. Finally, the map between subjective MOS value and normalized average distortion distance is established to get the objective MOS value. The good relationship between the predicted objective MOS and the subjective MOS indicates that the proposed algorithm can evaluate the subjective speech quality effectively, which can be widely used in the evaluation of a variety of audio systems.

Acknowledgment

We are heartily thankful to Prof. Benshun Yi and Dr. Kang Qiu both from Wuhan University for their support and help.

References

[1] D. Li, GSM Voice quality in cellular mobile communication. Chongqing, China: Chongqing University, 2008.

[2] S. Moller, W.Y. Chan and N. Cô té, "Speech quality estimation: Models and treands," IEEE trans. Signal Processing Magazine, vol. 28, pp. 18-28, 2011.

[3] S. Moller and R.Heusdens, "Objective estimation of speech quality for communication systems," Proceedings of the IEEE, vol. 101, pp.1955-1967, 2013.

[4] H. Xu, L. Lin, and X. Sun, "A new algorithm for auditory feature extraction," IEEE International Conference on Commmunication Systems and Network Technologies, pp.229-232, 2012.

[5] F. Hu and X. Cao, "Auditory feature extraction based on Gammatone filter bank," Computer Engineering, vol. 38, pp.168-170, 2012.

[6] ITU-T Rec.P.Sup23, ITU-T Coded-Speech Database. International Telecommunication Union, Geneva, Switzerland, 1998-02.

[7] W. Yin., B. Yi and D. Wu, "Speech quality evaluation method based on non-uniform spectral coefficient and GMM," Journal of circuits and systems, vol. 15, pp.104-109,2010.

[8] ITU-T R P. 862,"Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union-Telecommunication Standardisation Sector, 2001.