Automatic voice quality evaluation method of IVR service in call center based on Stacked Auto Encoder

Li Wang1*, Zongwei Wang1, Guoyi Zhao1, Yuan Su1, Jinli Zhao2 and Leilei Wang2
1 State Grid Customer Service Center, Tianjin 300309, China
2 BEIJING CHINA-POWER INFORMATION TECHNOLOGY CO., LTD., Tianjin 300309, China
*E-mail: li-wang@sgcc.com.cn

Abstract. The basic features extracted by traditional methods for speech quality evaluation are not clear, which leads to the small correlation coefficient of subjective and objective evaluation value. Therefore, an automatic voice quality evaluation method for IVR service in call center based on stackable automatic encoder is proposed. All kinds of devices are used to simulate the real use of IVR service voice of call center and collect IVR service voice of call center. According to the process of sampling quantization frame preemphasis window processing, the IVR service voice data of call center is preprocessed. Based on the structure of stackable automatic encoder, the reconstruction process of coding and decoding is designed to extract the basic features of business speech. BP neural network is introduced to establish an automatic speech evaluation model to evaluate speech quality automatically. Experimental results: compared with the traditional method, the average correlation coefficients of subjective evaluation value and objective evaluation value are 0.023517 and 0.02258 respectively, and the average deviation of correlation coefficient is 0.048775 and 0.03485 respectively.

1. Introduction
In recent years, with the rapid development of digital network and VLSI technology, all kinds of audio and video processing technology and audio and video transmission technology are changing rapidly [1]. Network voice transmission technology, especially call center IVR service, has become an important application and service project in the field of Internet and telecommunications, undertaking more and more voice communication tasks [2]. However, the voice service provided by IVR service will be affected by packet loss and other network factors in the transmission process, resulting in the low quality of IVR service [3]. Therefore, how to accurately and reliably measure and evaluate the voice quality of IVR service is a key problem in network measurement and network planning.

In modern telephone and network communication, there are various voice systems, which provide many kinds of different services [4]. The performance differences between these systems exist objectively, so the evaluation of various systems to improve the overall performance has become an important factor in the smooth flow of information exchange [5]. One of the main criteria to evaluate the system performance is to see the impact of these systems on the speech quality, so it is necessary to accurately evaluate the speech quality of the system output [6]. In addition, the subjective feelings of communication users and consumers for audio and video ultimately depend on the quality of audio and video [7]. So audio and video quality evaluation is becoming an important research topic.
Reference [8] extracted parameters to distinguish silence and speech segments from data packets, and then used the trained random forest model to evaluate speech quality, which achieved better evaluation effect with MOS value correlation. In reference [9], by constructing a two-dimensional dynamic topology map to represent the classification information and the sample relationship information between clusters, we can incremental identify clusters of arbitrary shape. Reference [10] adopts E-Model model for the purpose of network transmission planning, ignores the network damage of call center, and synthesizes the negative influence of several factors in the process of voice transmission as parameter R, and the higher the R value is, the better the voice quality is, and it does not need to input the original voice signal. Based on the above research results, this research introduces Stacked Auto Encoder to extract the essential features of voice, reduce the error of voice quality evaluation, and puts forward the research of automatic voice quality evaluation method of call center IVR service based on Stacked Auto Encoder.

2. Automatic voice quality evaluation method of IVR service in call center based on Stacked Auto Encoder

2.1. IVR service voice data acquisition and preprocessing

2.1.1. Voice data acquisition
The call center IVR service voice data acquisition equipment is composed of full anechoic room, processing room, sound source array, sound source, digital equalizer, computer, pulse multi-function analyzer, simulated head, hearing aid and other instruments. The acoustic signal acquisition unit is composed of a simulated head, IVR service voice of call center and pulse multi-function analyzer. In the process of audio acquisition, the head simulator of B&K Company 4128D is selected. It can simulate the diffraction and reflection of sound wave through the human head and trunk, and get the sound field around the ears. In the process of audio data acquisition, the voice output signal of IVR service in call center is collected by using pulse multi-function analyzer through simulating human head as the pure voice and noise voice.

2.1.2. Speech data preprocessing
In order to better analyze the voice signal, according to the voice data structure diagram of IVR service in call center collected by audio equipment shown in Figure 1, the collected voice signal is preprocessed. The preprocessing process is as follows:

1. Sampling quantization: Through sampling and quantization, the analog signal with continuous time and amplitude is transformed into corresponding digital signal with discrete time and amplitude. This process is called A/D conversion process.

2. Framing: Due to the short-term stationary characteristics of speech signal, the analysis method of stationary process can be used to frame the speech signal. This is to ensure the smoothness of the change between the feature parameters.

3. Pre emphasis: The pre emphasis processing of speech signal is realized by digital signal through first-order filter, and its emphasis process is as follows: using pre emphasis coefficient to increase the original signal sequence to form pre emphasis sequence. Among them, the pre emphasis coefficient threshold is [0.9, 1.0].

4. Windowing: Rectangular window has good smoothing performance, but it lacks high-frequency components of speech signal, which will lead to the loss of some detail information in speech waveform. In order to avoid the loss of information in speech waveform, Hamming window is selected as the window function of the system. Here, the basic information of sampling rate, frame length and frame shift in the preprocessing process will be given. The specific information is shown in Table 1.
Table 1 Parameter information of speech sample preprocessing

| Parameter name            | Parameter information | Parameter name            | Parameter information |
|---------------------------|-----------------------|---------------------------|-----------------------|
| Sampling frequency        | 8000 Hz               | Quantization bits         | 8 bits                |
| Frame length              | 30 ms                 | Frame shift time          | 10 ms                 |
| Frame length points       | 240 points            | Number of frame shifts    | 80 points             |
| Pre emphasis filter       | 1-0.9375 z^{-1}       | Window function           | Hamming window        |

2.2. Extracting essential features of speech based on Stacked Auto Encoder

Using Stacked Auto Encoder, the essential features of IVR service voice are extracted. The process of encoding, decoding and reconstruction is as follows:

1. In this study, the selected code function of Stacked Auto Encoder is as follows:

\[ f_\phi(x) = c(Mx + a) \]  

(1)

In formula (1), \( f \) is the vector of hidden layer; \( x \) is the logarithm power spectrum characteristic data of speech after given preprocessing; \( M \) is the weight matrix of coding; \( a \) is the bias vector of coding; \( c \) is the slope; \( \varphi = \{M, a\} \) is the set of weight matrix and bias vector of coding. In order to avoid "covering" a few samples, the weight of formula (1) is increased, and the Stacked Auto Encoder mapping function is obtained as follows:

\[
\begin{align*}
H(x, s_j) &= \frac{x}{s_j} + 1 \\
\frac{N}{N_j} &+ 1 \\
\end{align*}
\]

\[ f_\varphi(x) = c\left(M \left[ x + H(x, s_j) \right] \right) + b \]  

(2)

In formula (2), \( s_j = \sum_{j=1}^{N} x_j \), it represents the \( j \)-th sample in the \( N \) samples. If the number of samples with a value of 1 is more, it means that the \( s_j \) value is larger, and the \( H(x, s_j) \) value is smaller, the importance of the voice feature is lower, and vice versa; \( H(x, s_j) \) represents the mapping function of samples; \( N \) represents the total number of samples; \( \gamma \) represents the percentage of a small number of samples in the total samples.

2. The representation \( f \) of hidden layer is used as input, and the output layer represents \( o \). the decoding function of stacked auto encoder is selected as follows:

\[ o_\varphi(f) = c(M'y + a') \]  

(3)

In formula (3), \( M' \) is the weight matrix of decoding; \( a' \) is the bias vector of coding; \( \varphi' = \{M', a'\} \) is the set of weight matrix and bias vector of decoding.

3. From the above two steps, we can see that each input vector \( x^{(i)} \) will be mapped to a corresponding reconstruction vector \( o^{(i)} \) and hidden vector \( f^{(i)} \). \( (i) \) is the \( i \)-th sample in the \( N \) samples, then the optimal solution of \( \varphi \) and \( \varphi' \) should minimize the mean square error. The calculation formula is as follows:

\[
\begin{align*}
\varphi' = \arg\min_{\varphi, \varphi'} \frac{1}{N} \sum_{i=1}^{N} l(x^{(i)}, o^{(i)}) = \arg\min_{\varphi, \varphi'} \frac{1}{N} \sum_{i=1}^{N} l(x^{(i)}, o_\varphi\left[f_\varphi\left(x^{(i)}\right)\right])
\end{align*}
\]  

(4)
In formula (4), \( l \) is the loss function whose value is the distance of Bernoulli distribution. In this case, the discrete topological structure expression of speech logarithm power spectrum characteristic data \( x \) and output layer data \( o \) is as follows:

\[
I(x,o) = -\sum_{i=1}^{N} [x_i \log o_j + (1-x_i) \log (1-o_j)]
\]  

(5)

In formula (5), \( I(x,o) \) is the negative log likelihood function of \( x \). At this time, Stacked Auto Encoder is a deep neural network that trains the superimposed network layer by layer and obtains the required features in the last layer.

2.3. Building evaluation model to evaluate voice quality of IVR service

Based on the Stacked Auto Encoder model, the extracted essential features of voice are introduced into BP neural network to build an automatic voice evaluation model for IVR service in call center.

![Figure 1 Call center IVR service voice quality automatic evaluation model](image)

In Figure 1, \( E \) represents the standard deviation of network delay jitter; \( K \) represents the packet loss rate; \( M \) represents the speech quality measurement standard; * represents the number of network iterations; \( d_k \) represents the number of neurons with the network delay jitter standard deviation; \( p_k \) represents the number of neurons with network packet loss rate; \( k \) represents the number of neurons, and \( k = 1, 2, \ldots, n \), \( n \) represents the total number of neurons. According to the evaluation model shown in Figure 1, the voice quality of call center IVR service is evaluated. The standard deviation of network delay jitter and network packet loss rate are taken as inputs, and the output result is only one, which is the standard value of voice quality. Then the model outputs the calculation process of speech quality measurement standard as follows:

\[
M = \sum_{k=1}^{n} d_k(x) p_k(x)
\]  

(6)

By substituting formula (6), the result of IVR service voice quality evaluation can be obtained through the call center IVR service voice quality evaluation model shown in Figure 1.

3. Simulation experiment

3.1. Experimental preparation

In this experiment, the IVR service voice database of the call center contains the original voice, the processed distorted voice and the corresponding MOS score. In order to prove the effectiveness of the research method and evaluate the performance of speech quality, the selected training and test speech database is consistent with ITU-T P.563 standard.
Among them, 8000 sentences needed for testing and training are 16KHz, 16bit quantization and 8s in length. Each distortion statement has indicated the subjective MOS score, the score range is 1-5. In this experiment, 90% of the statements were selected as training data and 10% of the statements were used as test data.

3.2. Experimental result
Based on the speech sequences designed in this experiment, three groups of evaluation methods are used to evaluate 8000 speech sequences. The correlation coefficient $R$ between the objective evaluation and the subjective evaluation is calculated, and the mean square error $S$ is calculated to characterize the deviation degree of the correlation coefficient between the subjective evaluation value and the objective evaluation value. The calculation formula is as follows:

$$ R = \frac{C(X,Y)}{\sqrt{D(X)D(Y)}} $$
$$ S = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (X_i - Y_i)^2} $$

In formula (7), $C$ is the covariance; $D$ is the variance; $X$ is the subjective MOS score; $Y$ is the objective MOS score; $X_i$ is the subjective MOS score of the $i$th statement; $Y_i$ is the objective MOS score of the $i$th statement; $N$ is the number of test statements. It can be seen from equation (7) that the threshold value of correlation coefficient $R$ is $[0,1]$. The closer the value is to 1, the closer the objective speech quality evaluation is to the subjective speech quality evaluation. The threshold value of mean square error $S$ is $[0,1]$. The closer the value is to 0, the better the performance of the algorithm is.

Based on the calculation formula shown in equation (7), three groups of evaluation methods are obtained to evaluate 8000 speech sequences, as shown in Table 2.

| Method Speech sequence | Conventional method 2 Correlation coefficient | Mean square error | Conventional method 1 Correlation coefficient | Mean square error | Research method Correlation coefficient | Mean square error |
|------------------------|----------------------------------------------|------------------|----------------------------------------------|------------------|----------------------------------------|------------------|
| 8000                   | 0.9341                                       | 0.2660           | 0.9255                                       | 0.2463           | 0.9624                                 | 0.1163           |
| 7000                   | 0.9442                                       | 0.2555           | 0.9516                                       | 0.2436           | 0.9620                                 | 0.1132           |
| 6000                   | 0.9512                                       | 0.2435           | 0.9524                                       | 0.2235           | 0.9687                                 | 0.0995           |
| 5000                   | 0.9513                                       | 0.2209           | 0.9533                                       | 0.2164           | 0.9748                                 | 0.0964           |
| 4000                   | 0.9530                                       | 0.2146           | 0.9655                                       | 0.2044           | 0.9772                                 | 0.810            |
| 3000                   | 0.9698                                       | 0.1956           | 0.9589                                       | 0.1843           | 0.9888                                 | 0.0701           |
| 2000                   | 0.9582                                       | 0.1894           | 0.9634                                       | 0.1760           | 0.9930                                 | 0.0488           |
| 1000                   | 0.9671                                       | 0.1845           | 0.9658                                       | 0.1641           | 0.9948                                 | 0.0255           |

It can be seen from Table 2 that the more the number of speech sequences, the higher the correlation between objective evaluation and subjective evaluation, and the lower the deviation of correlation coefficient between subjective evaluation and objective evaluation. The average correlation coefficient between objective evaluation and subjective evaluation is 0.953613, and the average deviation between subjective evaluation and objective evaluation is 0.22125. It can be seen that the average correlation coefficient between the objective evaluation and the subjective evaluation of 8000 speech sequences evaluated by the research method is 0.023517 and 0.02258 larger than that of the conventional methods, and the average deviation of the correlation coefficient between the subjective
4. Conclusion
In summary, this study proposes a Stacked Auto Encoder based call center IVR service voice quality automatic evaluation method, which makes full use of the Stacked Auto Encoder to extract the essential eigenvalues from the voice sequence, which can better represent the voice and enhance the performance of the method. Through rigorous experiments, the method of this study is verified, and its performance is better than that of conventional methods. In the future research, it is necessary to establish a real-time evaluation method for the audio quality of network transmission, and evaluate the audio experience quality of users by monitoring the network transmission performance when the system is running.

References
[1] Hikmatullah M R, Haryadi S. (2017) Perceptual evaluation of speech quality over the top call service[C]//2017 3rd International Conference on Wireless and Telematics (ICWT). IEEE, 2017: 181-185.
[2] Steinmetzger K, Zaar J, Relao-Iborra H, et al. (2019) Predicting the effects of periodicity on the intelligibility of masked speech: An evaluation of different modelling approaches and their limitations[J]. The Journal of the Acoustical Society of America, 146(4):2562-2576.
[3] Wallace D, Cheer J. (2020) Design and evaluation of personal audio systems based on speech privacy constraints[J]. The Journal of the Acoustical Society of America, 147(4):2271-2282.
[4] Dong X, Williamson D S. (2020) Towards real-world objective speech quality and intelligibility assessment using speech-enhancement residuals and convolutional long short-term memory networks[J]. The Journal of the Acoustical Society of America, 148(5):3348-3359.
[5] Nogueira W, Abel J, T F ingscheidt. (2019) Artificial speech bandwidth extension improves telephone speech intelligibility and quality in cochlear implant users[J]. The Journal of the Acoustical Society of America, 145(3):1640-1649.
[6] Wallace D, Cheer J. (2020) Design and evaluation of personal audio systems based on speech privacy constraints[J]. The Journal of the Acoustical Society of America, 147(4):2271-2282.
[7] Hsu C Y, Tu S Y, Yang C T, et al. (2020) Digital audio signal watermarking using minimum-energy scaling optimisation in the wavelet domain[J]. IET Signal Processing, 14(10):791-802.
[8] Abou Haibeh L, Hakem N, Safia O A. (2017) Performance evaluation of VoIP calls over MANET for different voice codecs[C]//2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2017: 1-6.
[9] Bakeva V, Popovska-Mitrovikj A, Mechkaroska D, et al. (2019) Gaussian channel transmission of images and audio files using cryptcoding[J]. IET Communications, 13(11):1625-1632.
[10] Su Z, Li W, Zhang G, et al. (2020) A steganographic method based on gain quantization for iLBC speech streams[J]. Multimedia Systems, 26(2):223-233.