

Research on the Objective and Subjective Evaluation Methods of Speech Quality Driven by Feature Fusion

Pengzhen Yang^{1, a *}, Tao Qi^{1, b}, and Xin Lv^{1, c}

¹ 63879 Unit, Huayin, Weinan, Shaanxi 714200, China

^a pzycover@163.com, ^b 849344143@qq.com, ^c 365636798@qq.com

Abstract. Aiming at the problem of inaccurate speech quality evaluation caused by the previous subjective evaluation-dominated and insufficient coverage of objective evaluation indicators in military communication equipment speech testing, this paper proposes a subjective and objective comprehensive evaluation method based on feature fusion. It focuses on extracting feature parameters such as range, variance, skewness, and kurtosis to characterize the transmission characteristics of speech signals. Through the automatic extraction of speech signal features and objective score calculation, and finally through a large number of subjective and objective joint tests, the mapping relationship between objective scores and subjective evaluation levels was analyzed, completing the classification of speech quality levels for objective evaluation.

Keywords: Feature fusion; Subjective and objective comprehensive evaluation; Objective score; Speech quality levels.

1. Introduction

With the rapid development of electronic information technology, modern warfare has entered an era of informatization, intelligence, and unmanned operation. As the core of modern warfare, communication serves as the “neural hub” of the joint combat system. Speech quality is a core performance indicator of communication systems^[1], reflecting the acoustic clarity and intelligibility of speech^[2]. The speech quality directly affects the fulfillment of mission tasks by combat personnel, including effective acquisition of combat commands and completion of the combat operation^[3]. The communication equipment speech quality is an assessment means for the network transmission capability and speech codec capability of communication equipment, but traditional speech quality evaluation methods mostly rely on subjective listening tests, which have defects such as time consumption, poor consistency, and difficulty in automation^[4]. At present, speech quality testing for military communication equipment remains dominated by subjective testing, which relies on human hearing for subjective judgment. This imposes high requirements on listeners, and it is necessary to maintain extremely high concentration and sufficient objectivity and impartiality during testing due to the influence of subjective factors of humans.

Based on speech signal feature extraction for objective testing combined with subjective evaluation results, this paper presents a set of objective evaluation methods and evaluation criteria for speech quality testing, providing a reference for future speech testing of communication equipment.

2. Test Environment Construction

2.1 Establishment of Customized Speech Corpus

Speech quality testing for military communication equipment differs from end-to-end testing carried out by industrial departments in accordance with standards such as ITU-T P.862^[5]. It mainly assesses the speech quality of communication equipment under actual working conditions, so constructing test materials required for testing is the foundation of testing. According to actual application requirements and characteristics, common speech utterances are sorted out and the construction of a test speech database is studied.

When conducting speech quality testing for communication equipment, the use of a unified speech test database can ensure the comparability of test results under different times, locations, and equipment conditions. Taking it as the “reference speech”, it is played to the tested communication system through a microphone, and the system output is obtained and compared with the reference speech, so as to realize the subjective and objective combined test evaluation.

2.2 Construction of Artificial Head Test Environment

To accurately collect binaural acoustic signals, parts with great influence, such as the head and auricle, are made into a human head model, and traditional microphones (or sound level meters) are placed in the left and right ears for speech signal collection, i.e. the artificial head structure^{[6][7]}.

A standard military speech test database is constructed using high-fidelity recording equipment, and a speech quality test method based on the “artificial head+condenser microphone+sound level meter+human” framework is designed. Specifically, the “artificial head+condenser microphone” is used to capture speech signals after simulating the human ear, so as to observe the real-time frequency response of the speech. The sound level meter (placed in the ear canal) is employed to record the intensity of the speech signal during communication. Finally, human evaluators determine the speech quality grade of the communication equipment based on the signal parameters obtained from the sound level meter and subjective evaluation.

3. Subjective-Objective Speech Quality Evaluation Method Driven by Feature Fusion

Features such as range, variance, skewness, and kurtosis from the data collected by the sound level meter are extracted. Combined with the specific requirements and priorities of speech quality testing for communication equipment and the current subjective evaluation methods, a subjective-objective combined evaluation method is adopted to evaluate speech quality.

3.1 Subjective Test and Evaluation Method

Whether speech signals can be clearly distinguished is an important assessment indicator for communication distance testing, usually dominated by subjective testing. Listeners give speech evaluations based on subjective auditory perception. In actual operation, it is easily affected by multiple factors such as listening environment, emotion, attention, and listening level, and has high requirements for listeners.

For subjective testing, GJB 2763-96, namely Standard and Evaluation Method for Speech Quality Level of Communication Equipment, is mainly adopted^[8]. Its standards and methods for assessing speech quality include monosyllable intelligibility, sentence intelligibility, speaker recognition time, speech signal quality level, etc.

3.2 Objective Test and Evaluation Method

In objective evaluation, humans are not taken as the subject, but sound signals are mainly obtained through sensors, analyzed and processed, and conclusions are drawn. The direct object of objective evaluation of speech quality is the communication speech sending and receiving equipment, which is a comprehensive evaluation of the speech transmission network (including wired, wireless, and other transmission networks) and the speech transmission capability of communication terminals.

1) Build an artificial head test environment: Use wireless communication equipment (including sending and receiving terminals) to build a speech communication environment. Collect the sent and received speech, process it to generate statistical data about sound level, and generate an Excel file using the statistical data of the sending end and receiving end.

2) Data reading: Read the time and A-weighted sound level data of the reference end and receiving end in the Excel file (Column A is reference time, Column B is reference A-weighted value, Column C is receiving time, and Column D is receiving A-weighted value).

3) Add serial numbers of sampling points.

4) Feature extraction: Extract range, variance, skewness, and kurtosis.

① Range (reflecting the dynamic variation range of signal loudness):

$$R = \max(x) - \min(x) \quad (1)$$

where R : range; $\max(x)$: maximum value of the signal sequence; and $\min(x)$: minimum value of the signal sequence.

② Mean value (overall average intensity of the signal):

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (2)$$

where μ : signal mean; N : total number of sampling points; x_i : amplitude of the i th sampling point; and $\sum_{i=1}^N x_i$: sum of amplitudes of all sampling points.

③ Variance (reflecting signal energy fluctuation and dispersion degree):

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2 \quad (3)$$

where σ^2 : sample variance; $N-1$: correction of degree of freedom of sample; and $x_i - \mu$: deviation of a single point from the mean.

④ Standard deviation:

$$\sigma = \sqrt{\sigma^2} \quad (4)$$

⑤ Skewness (describing the symmetry of signal distribution and reflecting whether the waveform is biased in the positive or negative direction):

$$S = \frac{N}{(N-1)(N-2)} \sum_{i=1}^N \left(\frac{x_i - \mu}{\sigma} \right)^3 \quad (5)$$

where N : number of sampling points; $N-1$, $N-2$: statistical correction coefficients of sample; and $x_i - \mu$: deviation of amplitude from the mean.

⑥ Kurtosis (describing the steepness of distribution and impact characteristics and reflecting the peak prominence of speech):

$$K = \frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum_{i=1}^N \left(\frac{x_i - \mu}{\sigma} \right)^4 - \frac{3(N-1)^2}{(N-2)(N-3)} \quad (6)$$

where K : excess kurtosis (peakedness); N , $N+1$, $N-1$, $N-2$, $N-3$: statistical correction coefficients of sample; and $\left(\frac{x_i - \mu}{\sigma} \right)^4$: the fourth power of standardized deviation.

5) Difference calculation: Calculate the absolute difference of sending-receiving features.

Absolute feature difference (absolute deviation of a certain feature between sending and receiving signals):

$$\Delta_f = |f_{\text{rec}} - f_{\text{ref}}| \quad (7)$$

where Δ_f : absolute difference of the feature; f_{rec} : feature value at the receiving end; and f_{ref} : feature value at the reference end.

6) Normalization and scoring: Perform weighted summation according to auditory perception of humans (i.e. high sensitivity to loudness and energy, and insensitivity to waveform details) to achieve a comprehensive speech quality score consistent with subjective auditory perception.

Normalized difference (conforming to human ear) (normalize with the larger of the two values as the reference, and stabilize the difference in the [0,1] interval, more in line with auditory perception):

$$\hat{\Delta}_f = \frac{\Delta_f}{\max(|f_{\text{ref}}|, |f_{\text{rec}}|) + \varepsilon} \quad (8)$$

where Δ_f : absolute difference of the feature; f_{rec} : absolute value of the reference feature; f_{ref} : absolute value of the receiving feature; and ε : extremely small amount to prevent division by zero.

Weighted scoring formula of auditory perception of humans:

$$\text{score} = 0.4 \cdot \hat{\Delta}_R + 0.4 \cdot \hat{\Delta}_{\sigma^2} + 0.1 \cdot \hat{\Delta}_s + 0.1 \cdot \hat{\Delta}_K \quad (9)$$

Weight description:

Range $\hat{\Delta}_R$: Weight of 0.4 (loudness, highly sensitive to human ear)

Variance $\hat{\Delta}_{\sigma^2}$: Weight of 0.4 (energy, highly sensitive to human ear)

Skewness $\hat{\Delta}_s$: Weight of 0.1 (waveform shape, insensitive to human ear)

Kurtosis $\hat{\Delta}_K$: Weight of 0.1 (impact characteristics, insensitive to human ear)

3.3 Comparison of Subjective and Objective Evaluation Results

The speech quality of communication equipment is evaluated subjectively (speech quality level in GJB 2763-96,) and verified with objective results. The comparison between subjective evaluation and objective data processing results is shown in Table 1.

Table 1. Comparison of Subjective Evaluation and Objective Data Processing Results

No.	1	2	3	4	5	6	7	8	9
SEL	Level 3	Level 3	Level 3	Level 3	Level 3	Level 3	Level 4	Level 4	Level 4
OER	0.484	0.4629	0.4956	0.5103	0.5256	0.4534	0.4281	0.4204	0.4173
No.	10	11	12	13	14	15	16	17	18
SEL	Level 4	Level 4	Level 4	Level 4	Level 4	Level 3	Level 3	Level 3	Level 3
OER	0.4279	0.432	0.4225	0.4107	0.4456	0.4597	0.4917	0.4893	0.4552
No.	19	20	21	22	23	24	25	26	27
SEL	Level 3	Level 3	Level 4	Level 4	Level 5	Level 4	Level 4	Level 4	Level 4
OER	0.4927	0.4505	0.3649	0.2704	0.2289	0.3347	0.3546	0.3221	0.3728
No.	28	29	30	31	32	33	34	35	36
SEL	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4
OER	0.297	0.3072	0.314	0.3052	0.2641	0.3674	0.3088	0.3074	0.3512
No.	37	38	39	40	41	42	43	44	45
SEL	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4	Level 5	Level 5	Level 4
OER	0.3098	0.2945	0.3452	0.288	0.2537	0.2897	0.2476	0.2027	0.2885
No.	46	47	48	49	50	51	52	53	54
SEL	Level 5	Level 5	Level 4	Level 4	Level 5	Level 4	Level 4	Level 5	Level 5
OER	0.1914	0.238	0.2643	0.2572	0.2378	0.2724	0.2568	0.2278	0.2458
No.	55	56	57	58	59	60	61	62	63
SEL	Level 5	Level 5	Level 5	Level 4	Level 5	Level 5	Level 3	Level 3	Level 3
OER	0.2097	0.2173	0.2271	0.2691	0.1717	0.1657	0.5074	0.5104	0.5188
No.	64	65	66	67	68	69	70	71	72
SEL	Level 3	Level 3	Level 3	Level 3	Level 3	Level 3	Level 3	Level 3	Level 3
OER	0.5148	0.4628	0.4656	0.5088	0.5052	0.5256	0.5263	0.5127	0.5438
No.	73	74	75	76	77	78	79	80	81
SEL	Level 3	Level 3	Level 3	Level 3	Level 2	Level 3	Level 3	Level 2	Level 1
OER	0.4936	0.4523	0.5031	0.5265	0.6406	0.5354	0.5462	0.5537	0.8097
No.	82	83	84	85	86	87	88	89	90
SEL	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4
OER	0.2694	0.3092	0.3112	0.3303	0.3079	0.2956	0.3146	0.268	0.3016
No.	91	92	93	94	95	96	97	98	99
SEL	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4
OER	0.2924	0.331	0.3174	0.3183	0.3193	0.2987	0.3108	0.2921	0.3057
No.	100	101	102	103	104	105	100	101	102

SEL	Level 4	Level 4	Level 5	Level 5	Level 5	Level 4	Level 4	Level 4	Level 5
OER	0.2992	0.3414	0.145	0.1741	0.2008	0.4364	0.2992	0.3414	0.145
No.	103	104	105	106	107	108	109	110	111
SEL	Level 5	Level 5	Level 4	Level 5	Level 5	Level 5	Level 5	Level 5	Level 5
OER	0.1741	0.2008	0.4364	0.2203	0.1894	0.1804	0.1599	0.1779	0.1578
No.	112	113	114	115	116	117	118	119	120
SEL	Level 5	Level 3	Level 5	Level 5	Level 5	Level 5	Level 5	Level 3	Level 5
OER	0.1857	0.4818	0.1341	0.1069	0.1038	0.1285	0.082	0.4963	0.1309
No.	121	122	123	124	125	126	127	128	129
SEL	Level 5	Level 5	Level 3	Level 5	Level 4	Level 5	Level 5	Level 5	Level 5
OER	0.1416	0.2473	0.4694	0.1025	0.2958	0.2146	0.169	0.2327	0.1546
No.	130	131	132	133	134	135	136	137	138
SEL	Level 5	Level 5	Level 5	Level 5	Level 5	Level 3	Level 4	Level 4	Level 5
OER	0.2321	0.16	0.1463	0.1816	0.1704	0.5429	0.2772	0.3059	0.1861
No.	139	140	141	142	143	144	145	146	147
SEL	Level 4	Level 4	Level 4	Level 5	Level 5	Level 4	Level 5	Level 5	Level 5
OER	0.257	0.261	0.283	0.1205	0.1078	0.2677	0.0152	0.2059	0.2207
No.	148	149	150	151	152	153	154	155	156
SEL	Level 5	Level 5	Level 5	Level 5	Level 5	Level 5	Level 5	Level 5	Level 2
OER	0.13	0.2314	0.0546	0.2453	0.0367	0.1493	0.0494	0.1455	0.5512
No.	157	158	159	160	161	162	163	164	165
SEL	Level 5	Level 5	Level 4	Level 5	Level 5	Level 5	Level 5	Level 5	Level 5
OER	0.1119	0.0381	0.28	0.2361	0.1266	0.2333	0.2115	0.0368	0.1177
No.	166	167	168	169	170	171	172	173	174
SEL	Level 5	Level 5	Level 5	Level 5	Level 5	Level 5	Level 4	Level 4	Level 4
OER	0.1032	0.1262	0.1599	0.0845	0.1556	0.0941	0.3232	0.3901	0.3619
No.	175	176	177	178	179	180	181	182	183
SEL	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4	Level 4	Level 2	Level 1
OER	0.4129	0.345	0.2887	0.2959	0.2835	0.3269	0.3476	0.5606	0.6517

Note: SEL stands for subjective evaluation level; OER stands for objective evaluation result.

Typical evaluation levels of 5 and 3 in subjective speech quality evaluation are selected for graphical display, respectively. Figure 1 shows the waveform comparison with an objective evaluation of Level 5. Figure 2 shows the waveform comparison with an objective evaluation of Level 3.

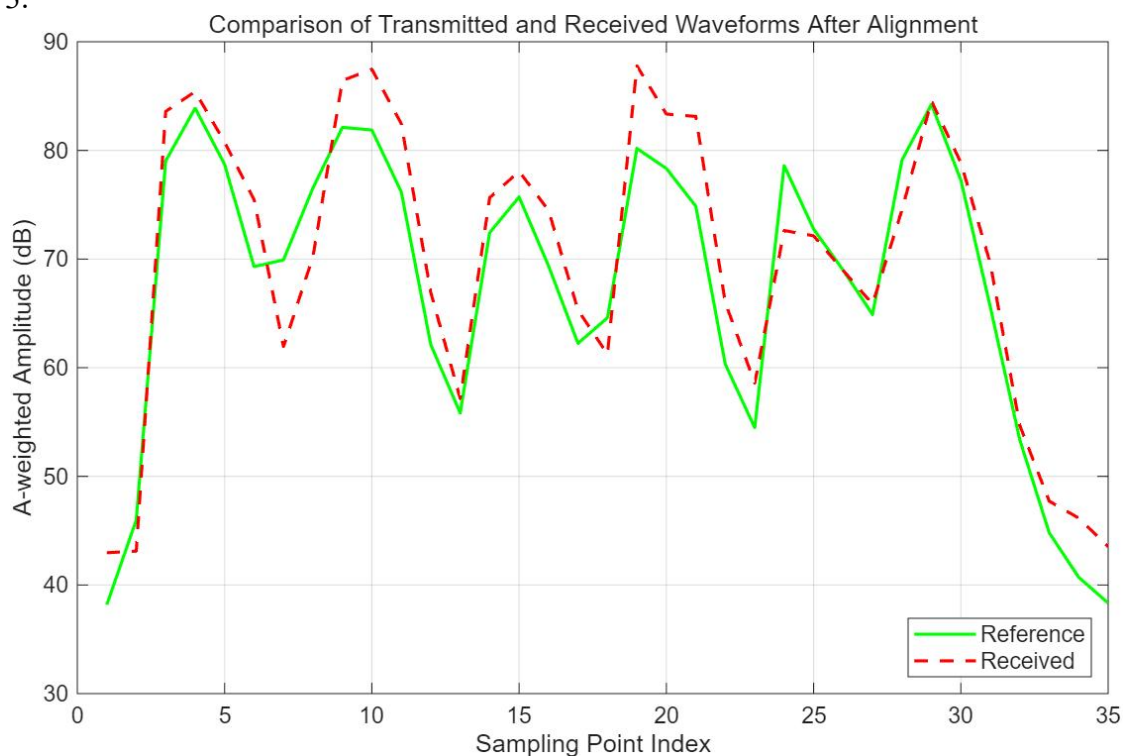


Fig. 1 Comparison I of waveforms of transmitted and received after alignment

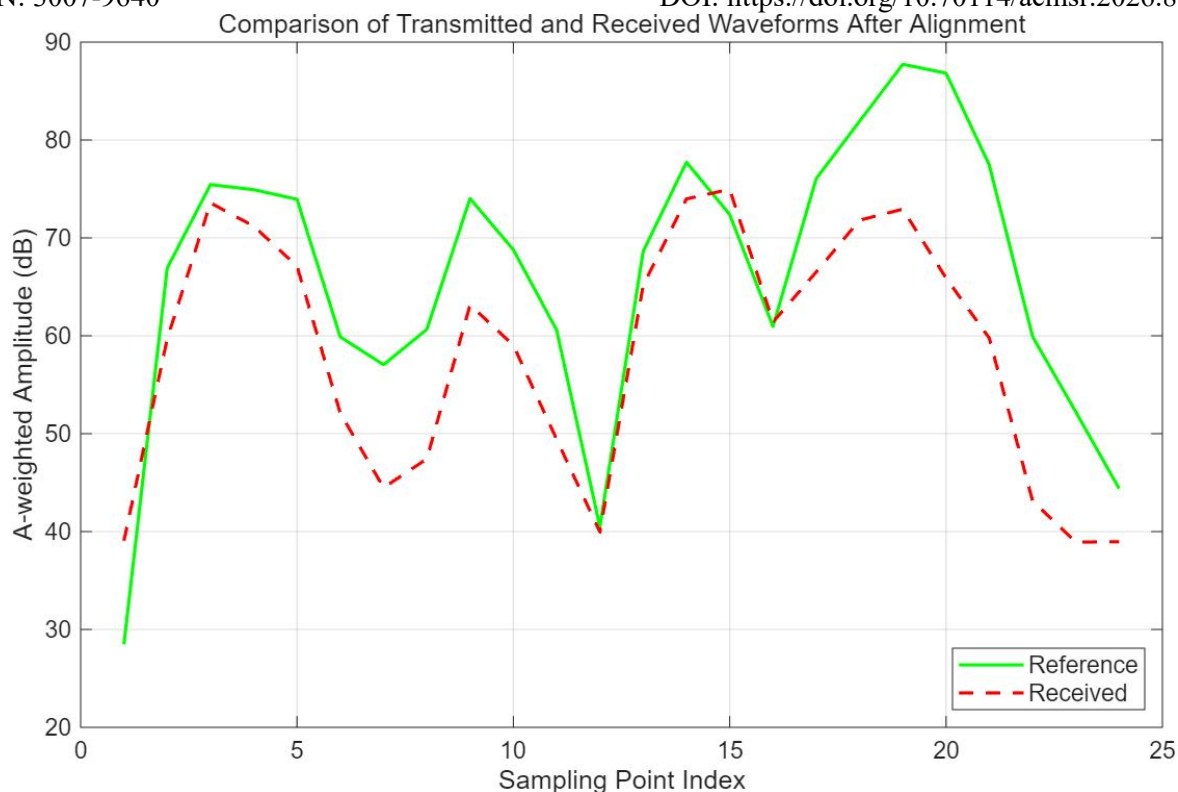


Fig. 2 Comparison II of transmitted and received waveforms after alignment

Figure 1 above shows that the waveforms of the sent and received speech signals are highly consistent, indicating normal loudness and volume transmission. The reference speech feature values: range: 46.0672, variance: 183.658, skewness: -0.8701, and kurtosis: -0.2539; and the received speech feature values: range: 44.8198, variance: 186.9301, skewness: -0.6044, and kurtosis: -0.6123. The corresponding comprehensive score is 0.1069.

Figure 2 above shows that the waveforms of the sent and received speech signals are basically consistent, indicating basically normal loudness and volume transmission. The reference speech feature values: range: 59.263, variance: 210.2089, skewness: -0.8633, and kurtosis: 0.7164; and the received speech feature values: range: 36.0392, variance: 161.7406, skewness: -0.3325, and kurtosis: -1.3654. The corresponding comprehensive score is 0.4629.

3.4 Speech Quality Level Judgment for Objective Evaluation

Scoring S (score) level standard:

$S < 0.25$: Excellent, no distortion or missing word, corresponding to speech signal quality of Level 5

$0.25 \leq S < 0.45$: Good, clear and intelligible, no obvious missing word, corresponding to speech signal quality of Level 4

$0.45 \leq S < 0.55$: Fair, slight distortion, large noise, poor sound quality, corresponding to speech signal quality of Level 3

$0.55 \leq S < 0.65$: Poor, obvious distortion, weak signal, corresponding to speech signal quality of Level 2

$S \geq 0.65$: Very poor, almost all the speech with distortion and missing words, weak signal, corresponding to speech signal quality of Level 1

4. Summary

This paper proposes an objective evaluation method based on the fusion of multiple features such as range, variance, skewness, and kurtosis. The fused features are processed automatically, and

a comprehensive scoring of the objective evaluation is realized by combining objective evaluation with subjective evaluation, which further improves the efficiency and objectivity of speech testing.

References

- [1] M. Keller, et al, Non-intrusive speech quality assessment based on deep neural networks for speech communication, *IEEE/ACM Trans. Audio, Speech, Lang. Process*, 2024, 32: 1-12
- [2] B. Baik, et al, Vocal delivery quality in earnings conference calls, *J. Account. Econ*, 2025, 80: 101568.
- [3] Wu, C. L. *Communication and Command and Control Equipment Test*. Beijing, China: Military Science Press, 2009: 1-9.
- [4] Lin, Y., Xu, H. T., et al. Objective Evaluation of Communication Speech Interference Effects Based on Feature Fusion. *Journal of Communication*, 2023, 44(3): 102-113.
- [5] International Telecommunication Union-Telecommunication Standardization Sector (ITU-T). *Perceptual Objective Listening Quality Analysis*. Geneva, Switzerland: ITU-T, 2011.
- [6] L. Genz, M. Vorländer. Effect of HRTFs and head motion on auditory-visual localization in real and virtual studio environments, *Acta Acustica*, 2025, 9(1): 90.
- [7] F. Brinkmann, J. Fels, U. Svensson. Evaluation of head-tracked binaural auralizations of speech signals generated with a virtual artificial head in anechoic and classroom environments, *Acta Acustica*, 2021, 5(1), 37.
- [8] GJB 2763-96, *Grade Standard and Assessment Method of Speech Quality for Communication Equipments*. Beijing, China: Commission of Science, Technology and Industry for National Defense, 1996.