acoustics

MDPI

Article

Perceptual Evaluation of Speech Quality for Inexpensive Recording Equipment

Anas Hashmi

Department of Electrical and Electronic Engineering, University of Jeddah, Jeddah, P.O. Box 13151, Jeddah 21493, Saudi Arabia; ahashmi@uj.edu.sa

Abstract: This research studies the perceptual evaluation of speech signals using an inexpensive recording device. Different types of noise-reduction and electronic enhancement filters viz. Hamming window, high-pass filter (HPF), Wiener-filter and no-speech activity-cancelling were applied in compliance with the testing conditions such as P.835. In total, 41 volunteers participated in the study for identifying the effects of those filters following a repeatable approach. Performance was assessed in terms of advanced perceptual audio features. This study is believed to be beneficial for both users and device manufacturers as the suggested technique is relatively simple to embed in operational device algorithms or in the master GPU.

Keywords: audio signal processing; speech enhancement; speech processing; filters

Citation: Hashmi, A. Perceptual Evaluation of Speech Quality for Inexpensive Recording Equipment. *Acoustics* **2021**, *3*, 200–211. https://doi.org/10.3390/acoustics3010014

Academic Editor: Arianna Astolfi

Received: 11 February 2021
Accepted: 8 March 2021
Published: 10 March 2021

1. Introduction

Voice recording instruments are widely used in numerous applications and embedded in various types of technologies, ranging from gaming to military purposes. During the past fifty years or so, there has been a massive growth in communication technologies facilitating broader transmission bands and enabling to transmit a huge amount of data. As a result, the data rates were increased from 2 kbps for 1G in the early 1970s up to several Gbps for 5 G by now [1]. Further, the sensors with higher data rates that perform closer to the human level of perception were developed.

Microphones and recording instruments have become more popular during the COVID-19 pandemic as professionals have moved into online meeting environments. Even during the post pandemic, people may choose to communicate through social media platforms or via VoIP services. Further, it was noticed that there was a significant increase in the use of audio multimedia, including podcasts. Generally, microphones of the phones and inexpensive voice recording devices are used for audio recording and streaming. However, such inexpensive devices result in a degradation of voice quality and make them susceptible to noise.

Only a few research studies were conducted to improve the commercially available inexpensive recording instruments. For example, use of FFT and windowing for enhanced sonogram of an inexpensive microphone is explained in [2]. However, no other no other digital signal processing techniques were explored. In some research studies, the performance of the audio recording systems was assessed only in terms of the SNR [3].

Various subjective and objective speech assessment methods are available to analyze, improve, or compare the speech quality. Though noise can be removed via conventional noise filtering techniques, it was found that filtered speech signals do not result in a more realistic sound. Unlike image or video processing, distinguishing performances of different filters using existing algorithms such as Perceptual Objective Listening Quality Analysis (POLQA) or Perceptual Evaluation of Speech Quality (PESQ) is a highly challenging task. Moreover, the Signal-to-Noise Ratio (SNR) does not adequately predict subjective quality for modern network equipment [4]. Based on these facts, it is proposed that more suitable models, which can be implemented easily, should be employed.

This research focuses on the subjective quality assessment of different types of filters, which can be applied on inexpensive microphone speech output. The support of more than 40 volunteers in different ages and genders was obtained for this study to identify the optimum arrangement for processing of the speech signals.

2. Audio Quality Assessment Techniques

2.1. Subjective Speech Testing

Human hearing is accomplished through complex operations in the neural network of the brain after sounds are perceived using vibration detection in the middle and inner ear [5]. The subjective audio test is executed by gathering subjective opinions of a number of individuals following procedures such as the ITU-T Recommendation P.800–P.835 [6]. This five-point-scale-based procedure focuses on assessing the speech quality with different levels of noise components. It is commonly used for the analysis of noise mitigation techniques in audio or speech signal processing, making it possible to derive speech as well as noise levels [7].

2.2. Objective Speech Testing

Objective testing methods were performed to compensate differences that might occur in the subjective tests, which apply psycho-acoustic models. The primary concept behind them is to compare a "clean" signal with a distorted sample using multiple algorithms and procedures as listed in the international standards. Among the six major international voice quality objective assessments viz. PSQM (1996), PEAQ (1999), PESQ (2000), 3SQM (2004), PEVQ (2008), and POLQA (2010), the most commonly used methods are PESQ and POLQA [8].

Currently, conducting both subjective and objective tests is more popular than digital signal processing methods. Subjective methods generate more precise opinions but require more time, expenses, and data analysis. They are usually performed where users are placed in an anechoic or semi-anechoic environment and are subjected to different conditions and types of noise. A typical PESQ test includes measuring the features of both objective and subjective quality assessment in terms of $RMSE$ as given in Equation (1) [9–13]. The $RMSE$ is a measure of the differences between values predicted by a model and the subjective values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{L} \left(Q_i - \hat{Q}_i\right)^2}{L}} \tag{1}$$

where Q_i is the Measured Subjective Listening Quality, $\hat{Q}_i$ is the Measured Objective Listening Quality and L is the number of coefficients in the database.

3. Practical Experiment

3.1. Introduction

The experimental tests were applied in compliance with the ITU-T Recommendation P.835 to assess the quality of speech recorded via an inexpensive microphone. The speech signals were processed using several audio enhancement techniques and filters. As the experiments involved human interactions, all necessary approvals were obtained from the Dean of the College of Engineering at the University of Jeddah in Saudi Arabia.

3.2. Procedure

The speech segments were recorded using an inexpensive general-purpose microphone from a low-quality brand with a cost of less than $2 per unit, enclosed within housings for noise reduction (Figure 1). An evaluation was conducted to compare the operation of this microphone with a professional Rode NT1-Kit, which costs around $300 per unit. Recordings were made by placing the microphones closer to each other around 0.15 m away from the speaker's mouth. Table 1 shows a comparison between the two types of microphones.

Figure 1. The recording setup with two inexpensive microphones, where one of them is used as the backup.

Table 1. A comparison between the microphones used in the experiments.

	Inexpensive Microphone	**Rode NT1-Kit**
Max value	1.049	0.8594
Min value	−1.0003	−0.24954
Mean value	-3.7901×10^{-5}	-2.3820×10^{-5}
RMS value	0.22281	0.123398
Dynamic range D (dB)	164.8592	158.0119
Crest factor Q	13.4568	22.3666
Autocorrelation time (s)	1.4091	0.0092063
Mean noise	0.25	0.15

The subjective testing was performed online in April 2020. Due to the COVID-19 pandemic lockdown, the audio samples were distributed and volunteers were asked to playback the audio using HD noise-cancelling headphones of the given standards owned by them. The users were asked to ensure that they pay their full attention on the played audio and would not be distracted. Collecting observations by following a similar procedure under different focusing arrangements (e.g., dual tasks and listening efforts) is proposed in [7]. The development of audio-only or audio—visual corpora for speech enhancement is reported in the literature. They are summarized in Table 2. However, there are only a limited number of audio-only datasets and some of them were recorded in a noisy environment.

Table 2. Comparison of online benchmark datasets.

Dataset	**Modality**	**Speakers**	**Environment**
COSINE [14]	Audio-only	133	Noisy
VOICES [15]	Audio-only	300	Not Noisy
GRID [16]	Audio—Visual	34	-
Mandarin Sentences [17]	Audio—Visual	1	-
AVSPEECH [18]	Audio—Visual	-	-
BANCA [19]	Audio—Visual	208	Noisy
AVICAR [20]	Audio—Visual	100	Noisy
ASPIRE [21]	Audio—Visual	3	Noisy
VISION [22]	Audio—Visual	209	Noisy
Proposed	Audio-only	1	Not Noisy

The 41 volunteers were reached via the institutional mailing system, social media, as well as personal contacts. The volunteers comprised of 23 males (56%) and 18 females (44%). They were from a wide range of ages between 18 to 65 years having a median of about 33 years. This number was found sufficient as it was comparable to similar research studies [4,10]. As all the participants were Arabic speakers and the samples were recorded in Arabic Language with Saudi accent maintaining clarity, correct accent, and consistency. In the literature, it was found that some tests were performed using subjects of different nationalities [7]. Still, some studies suggest that the imperfections in the languages were not an issue when assessing the speech quality [23].

Speech signals of duration of 30 s were used as test samples, where only the first 8 s were played [4]. A total of 12 speech audio files were played. The chosen phrases were parts of a modern Arabic poem, which was clearly recognizable. Responses were collected via online forms or via questionnaires (from the physically presented volunteers). Volunteers were given the opportunity to send their comments and observations, and they were also considered. Data were recorded and analyzed using speech quality test (S-MOS), as the scope of this research is to distinguish between filtered signals with no added noise involved [24]. The assessments were conducted as per the ITU-T Recommendation P.800–P.835 using 5-rating scores [6]. The opinion score of 1–5 are given for the status of "Very distorted", "Fairly distorted", "Somewhat distorted", "Slightly distorted", and "Not distorted".

4. Signal Processing

4.1. Test Samples

For temporal analysis, the outputs obtained from the two microphones were plotted. A recorded sample of around 30 s duration that contained 16 speech segments with short, medium, and long pauses in between is shown in Figure 2. It was observed that the test signal went into saturation more often with noticeably increased edges in the non-silence zones. This technique had a positive impact on decreasing SNR compared to expensive options. However, it resulted in an excessive audio sharpness, which was uncomfortable to listen, especially for recordings of longer durations.

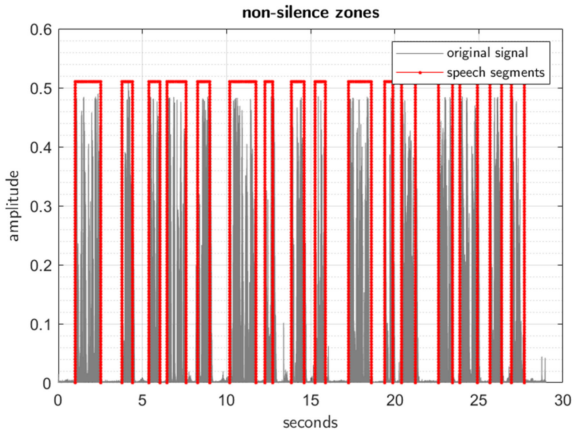
4.2. Speech Activity Detection (SAD)

When the first non-silence zone in the audio signals was eliminated [25], speech quality was enhanced by around 8%. As setting a threshold is not an efficient method and can lead to misclassification of speech components, the neighboring point option was used. It was tested for noise identification in an ultrasonic oscillating sensing loop [26].

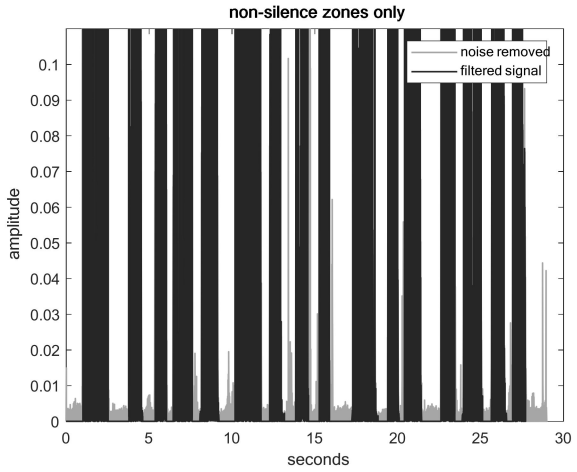
As speech activity zones do not change rapidly from one sample to another, occurrence of signal jumps can be addressed and eliminated by applying an appropriate signal processing technique. The boundaries were selected by trial and error to keep the original speech signal while removing any white noise. Based on Equation (2), the accepted speech signal that formed the filter was implemented.

$$T_k = s_0 \pm \alpha k \;\; k \in \mathbf{Z} \tag{2}$$

where T_k is the horizontal threshold relative to the neighboring k number of samples in speech activity zones, s_0 are the examined samples of the speech activity zone, and α is the constant of examined amplitude intervals. Equation (2) was applied to detect speech activities accurately, given that normal speech usually contains short (0.15 s), medium (0.50 s), and long (1.50 s) pauses [27,28]. The identified regions were highlighted in Figure 2a. The areas outside the highlighted boxes were removed, as shown in Figure 2b for analysis.



(**a**)



(**b**)

Figure 2. Temporal analysis of a test signal (**a**) speech activity zones (**b**) filtering the non-speech activity zones.

4.3. High-Pass Filtering

Low-pass filters (LPFs) are commonly used for speech quality enhancement, especially for medical purposes such as diagnosis of auditory processing disorders. Due to the mixed noises present in the lower spectrum closer to the pitch (Figure 3), it was found that using a LPF in speech enhancement has a negligible effect in improving the speech quality for such microphones. Given that typical human pitch can range from 100–120 kHz with slight variations depending on the age and gender [29], designing an LPF that produces promising results was highly challenging.

Use of different arrangements of high-pass filters (HPFs) for audio signal enhancements is reported in [30,31]. This research focuses producing an output audio signal much closer to real signal, mitigating the effect of the imperfections of the inexpensive microphone. Therefore, a finite impulse response (FIR) high-pass filtering was implemented. The filter has a band-stop frequency of 1.9 kHz and a passband frequency of 2 kHz [28]. Further, its stop band attenuation is 65 dB and passband ripple is 1 dB.

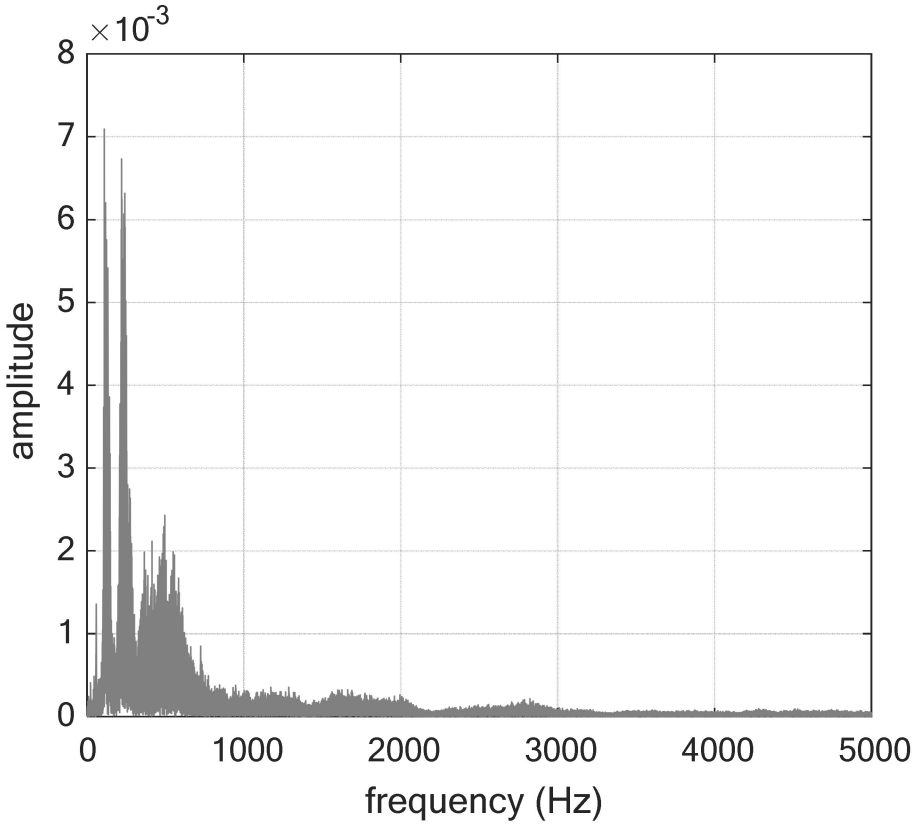


Figure 3. Frequency spectrum of a test signal.

4.4. Wiener Filtering

Wiener filtering is a signal processing tool applied on noisy signals that utilizes Linear-Time Invariant (LTI) filtering to generate an estimated random process. They can be operated by assuming that both the clean signal and the associated additive noise are known and defined. Though the Wiener technique is widely known for the effective processing of noisy images [32], it is employed in recent research for audio signal processing for both single-channel speech enhancement [33] and PESQ-based speech enhancement [24].

However, employing Wiener filters in a speech signal of this nature and specifications is rather complex as it is required to distinguish the signal from noise in a narrow bandwidth. One method employed is to track a priori SNR estimation using the decision-directed method [34] and enhance using Harmonic Regeneration Noise Reduction (HRNR) [35].

4.5. Windowing

It was observed that the microphone operates in the saturation region, as the diaphragm was pushed slightly toward the microphone's back plate [36]. Manufacturers usually do not employ electronic inter-stages in inexpensive devices for impedance reduction. Due to the non-linearity, signal distortions such as a Total Harmonic Distortion (*THD*) occur [37]. *THD* tends to clip the signal, which in turn generates more unwanted harmonics. The *THD* for saturated signals can be as large as 48% when calculated by applying Equation (3) but is significantly less for sinusoidal signals [38].

$$THD_F = \sqrt{\frac{\pi}{8} - 1} \approx 48.3\% \tag{3}$$

One method to maintain a sinusoidal waveform in longitudinal waves is to implement an Automatic Gain Control (AGC), which requires additional cost and a method of control [26].

Another method to mitigate the distortion effect for the audio signal is to employ a windowing technique for decreasing the sharpness of the raw signal and hence to reduce the *THD*. The window should soften the signal edges transforming the wave to a more sinusoidal shape (Figure 4). There is a trade-off between the signal loss and the operation of such a technique. It was found that the hamming window modulates the envelope of the source signal where the peak reaches the maximum range of the speech activity zone, as shown in Figure 4. The hamming window was shifted to preserve 75% of the original signal.

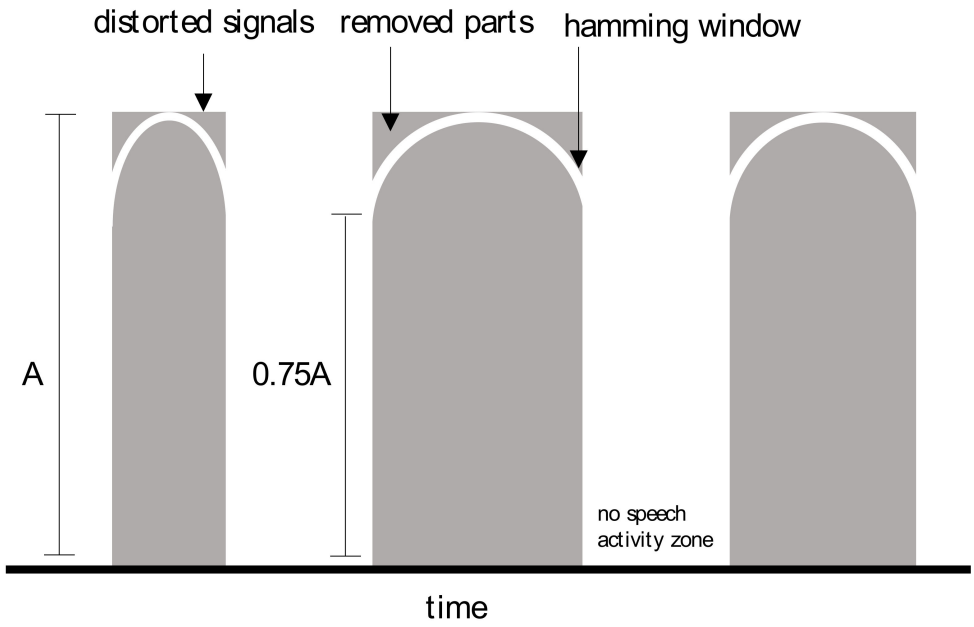


Figure 4. Application of the Hamming window on the test signal.

5. Results

5.1. Experimental Results

The abovementioned experiments were simulated and tested individually using MATLAB scripts, where their functional parameters were adjusted to produce the optimum audible results. The results were analyzed using a spectrogram to illustrate the frequency components in the time domain (Figure 5).

5.2. Results of the Speech Quality Test

Results from subjective tests carried out in compliance with the ITU-T Recommendation P.835 are given in Table 3, where each run summarizes an average of four audio clips. Each of the six test methods ([a]–[f]) was run 3 times and each of them contained 4 audio clips generating 72 tests (=6 × 3 × 4). The average score as well as the standard deviation of each test was determined to maintain consistency. A high quality test of the high-precision Rode NT1-Kit microphone was also conducted and treated as a reference signal (Figure 6).

Table 3. Summary of the speech quality test results.

Method	Run 1	Run 2	Run 3	AVG	STD	Δ from Ref. [a]
[a]	2.76	2.65	2.34	2.58	0.35	0
[b]	2.76	2.93	2.83	2.84	0.16	0.26
[c]	3.29	3.37	3.15	3.27	0.22	0.92
[d]	3.03	2.96	3.03	3.01	0.24	0.43
[e]	3.41	3.49	3.59	3.50	0.30	0.69
[f]	4.14	4.02	4.09	4.08	0.59	1.50

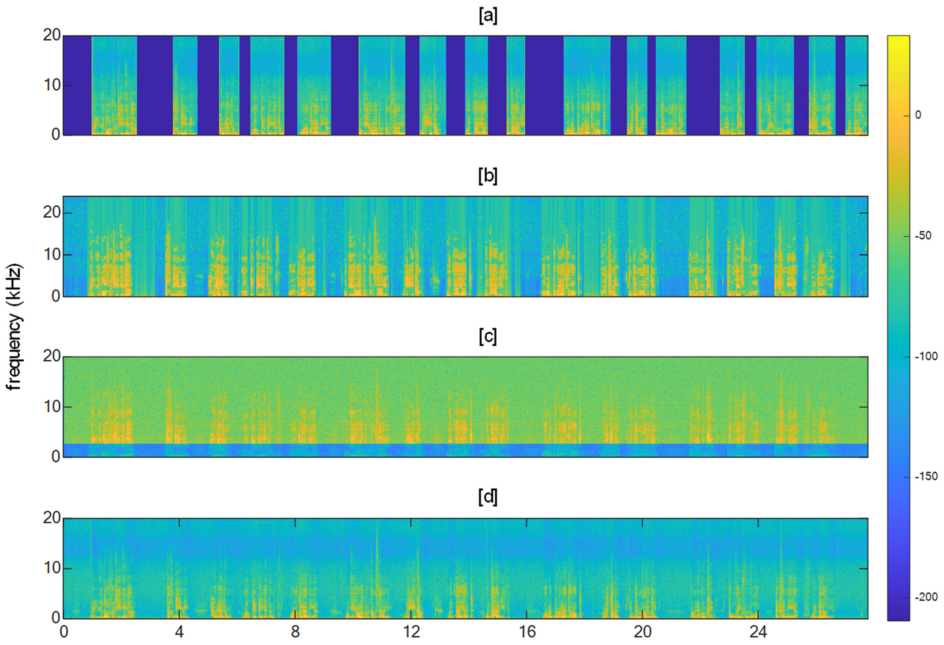


Figure 5. Experimental results of speech quality enhancement: (**a**) non-silence zones; (**b**) Wiener-filtered signal; (**c**) HPF signal; and (**d**) hamming segments.

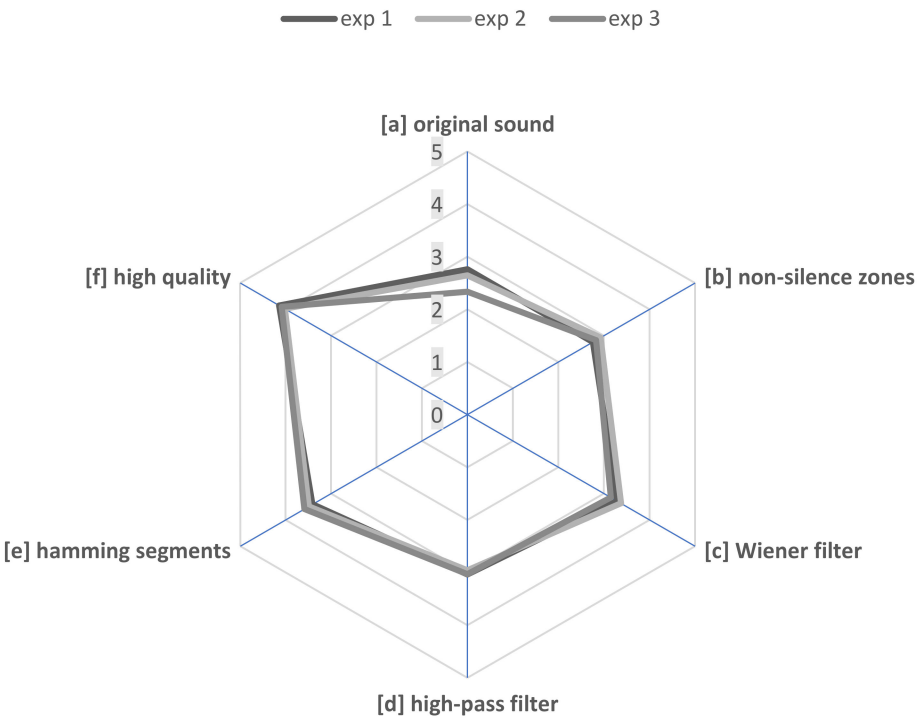


Figure 6. A summary of the speech quality test results.

5.3. Results of the Modified Assessment

During the experiments, some common comments were reported giving an insight for further assessment and analysis. Therefore, the survey was modified by adding the following parameters, where the assessments follow the G-MOS scaling from 1–5. The results are given in Table 4 and Figure 7.

- Heavy, reversed: the pitch was higher than expected, or the voice turned into a 'robotic' sounding tone.

- Realistic: closer to the real (reference) voice of the reader.
- Sharp, reversed: the speech was at a sharper pitch where the frequency was lowered.
- Annoying, reversed: a measure of how annoyed the listener became with the voice, and it included the presence of unwanted noise.
- Clear: the words and sentences were clear and heard in an understandable manner.
- Convenient: the user could listen to the clip for longer durations without getting annoyed.

Table 4. Summary of modified subjective test results.

Method	Heavy, R	Realistic	Sharp, R	Annoying, R	Clear	Convenient
[a]	3.5	3.2	2.9	2.8	3.4	3.1
[b]	3.3	2.3	2.8	2.0	3.8	2.2
[c]	4.0	4.2	4.1	4.3	3.9	3.4
[d]	2.3	1.5	2.8	4.1	3.5	3.2
[e]	3.4	1.6	3.0	3.1	3.5	3.2
[f]	4.2	4.7	4.6	5.0	4.9	4.9

R: reversed

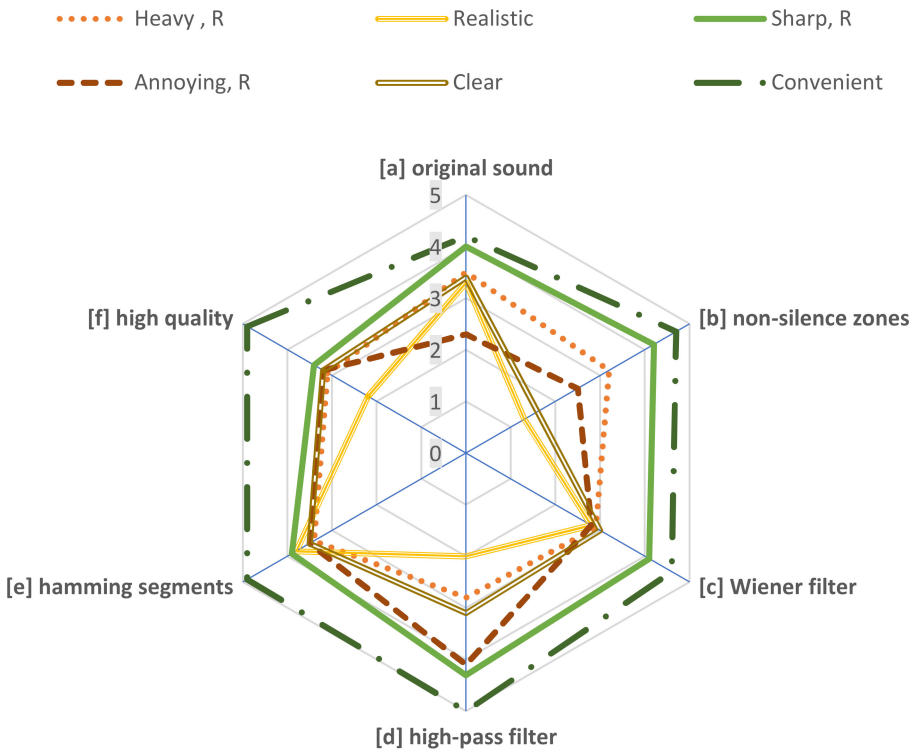


Figure 7. Results of the modified assessment.

6. Discussion

The experiments have produced a range of outcomes related to the subjective speech quality assessment. The recorded raw data were assessed by volunteers where it scored between "poor" and "fair" scales. Some users indicated that the inexpensive speakers are still considered acceptable for general purpose applications.

When the no-speech activity zones were removed entirely by the neighboring-points approach, the results improved slightly reaching a level closer to the "good" boundary. Wiener filtering provided a significant improvement of almost a full point on the marking scale, making the performance between "fair" and "good" on the scale. Introducing an HPF resulted in a moderate increase by almost half a point, improving the speech quality up to the border of "fair" on the scale.

The Hamming segmentation method results in a noticeable increase in quality by around 0.7 points, improving up to the "fair" level. Interestingly, speech recorded using professional microphones resulted in quality rated in the "good" level. Though the signal processing produced less realistic voice (e.g., computerized, robotics voice), the volunteers found that their quality is still high.

To ensure repeatability and to increase the reliability of results, tests were shuffled and conducted in a random manner. The standard deviation of voice filtering resulted in a difference of only ± 0.15 on the scale. Fortunately, this level is acceptable to clearly distinguish speech quality based on the levels of ITU-T Recommendation P.835. Interestingly, the standard deviation was almost doubled when high-quality speech was considered.

The signal processing results show that HPF and hamming segments produced a less realistic voice. HPF results in a heavier speech as it removes the lower pitches, therefore significantly changing the identity of the speaker. Playing "non-silence zones" for longer durations made the listeners uncomfortable due to the sudden silent zones; therefore, this method is not recommended to be used for speech that is longer than 30 s in duration.

In similar research studies conducted by using audio-only as well as audio—visual signals, the findings and conclusions were presented in numerous ways. For example, the Pearson correlation coefficients and pairwise comparisons between each pair of MOS tests, conducted following the ITU-T P.835 methodology, were presented in [7]. A multiple number of subjective speech quality tests were conducted with and without a parallel task. Though the test results were highly correlated, ten differences were found due to voting mistakes during the parallel tasks as the respondents lost concentration.

The results of the objective and subjective assessment were compared in [9]. The value of the percent correct words was 98.9% suggesting that the voice quality of the respondents is good. A close correlation with the Perceptual Evaluation for Speech Quality (PESQ) was reported as DMOS demonstrated a correlation coefficient of 0.82. Performance evaluation using objective and subjective tests showed that visual cues are more effective at low SNRs [21]. Therefore, the audio—visual models outperformed the audio-only model in silent speech regions. Though the performance of audio—visual models remained constant until 20% of the visuals were removed, and the performance decreased linearly thereafter. Demonstrating a high correlation between speech and lip shape, the importance of using correct lip shapes in speech quality enhancement of audio—visual signals is stressed in [28].

The possibility of using subjective tests performed at one location to predict the speech quality elsewhere was explored in [10]. Tape recordings of voice samples in seven languages were processed by 38 communications circuits and assessed by native listeners based on a five-point scale. The estimate of MOS that contains an additive correction demonstrated a root mean square error of 0.276 on the scale from 1 to 5.

7. Conclusions

Various filters were tested to evaluate speech signals recorded using an inexpensive voice recording device. As per the research findings, all the techniques employed resulted in an increase in performance compared to the original signal. Wiener filtering showed the most significant improvement. Employing the Wiener filtering resulted in a full point increase as per the ITU-T Recommendation P.835 evaluation criteria, with an error less than ± 0.15 scale points. The research findings are useful to perform both online and offline speech processing in telecommunications and multimedia environments. Guidance will be provided to the buyers for purchasing devices with an acceptable recording quality, which in turn reduces buying expenses. This research can be expanded in the field of speech quality assessment by using voice samples of different languages. Further, the participation of more volunteers can provide more reliable testing environments. A wider range of microphones in different qualities can be used in testing to provide recommendations to the buyer and feedback to manufacturers. The audio—visual data consisting of speech signals can also be used for evaluation of speed quality enhancement [39]. Furthermore, the research can be extended to process speech signals collected from noisy environments [29].

Funding: This research was not funded by any organization.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: The consent of participation and publication forms included the necessary information on the paper and some guidelines to follow. All participants approved the following option: "I have read and understand the above consent form, I certify that I am 18 years old or older, and, by selecting "Yes" below, I indicate my willingness to voluntarily take part in the study."

Data Availability Statement: Data available on request from the authors.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Net-Informations, "1G Vs. 2G Vs. 3G Vs. 4G Vs. 5G". Available online: http://net-informations.com/q/diff/generations (accessed on 4 June 2020).
2. Staderini, E.M. Inexpensive Microphone Enables Everyday Digital Recording of Deglutition Murmurs. In Proceedings of the 2014 8th International Symposium on Medical Information and Communication Technology (ISMICT), Florence, Italy, 2–4 April 2014; IEEE: Piscataway Township, NJ, USA, 2014; pp. 1–5.
3. Rempel, R.S.; Francis, C.M.; Robinson, J.N.; Campbell, M. Comparison of audio recording system performance for detecting and monitoring songbirds. *J. Field Ornithol.* **2013**, *84*, 86–97. [CrossRef]
4. International Telecommunication Union. *Rec. P.861, Objective Quality Measurement of Telephone-Band (300-3400 Hz) Speech Codecs*; International Telecommunication Union: Geneva, Switzerland, 1996.
5. Smith, S.W. Audio Processing. In *Digital Signal Processing*; California Technical Publishing: San Diego, CA, USA, 2003; pp. 351–372.
6. International Telecommunication Union. *P.800: Methods for Subjective Determination of Transmission Quality*; International Telecommunication Union: Geneva, Switzerland, 1996.
7. Avetisyan, H.; Holub, J. Subjective speech quality measurement with and without parallel task: Laboratory test results comparison. *PLoS ONE* **2018**, *13*, e0199787. [CrossRef] [PubMed]
8. Pomy, J. POLQA The Next Generation Mobile Voice Quality Testing Standard. 2011. Available online: https://www.itu.int/ITU-D/tech/events/2011/Moscow_ZNIIS_April11/Presentations/09-Pomy-POLQA.pdf (accessed on 1 June 2020).
9. Arifianto, D.; Sulistomo, T.R. Subjective evaluation of voice quality over GSM network for quality of experience (QoE) measurement. In Proceedings of the 2015 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Nusa Dua Bali, Indonesia, 9–12 November 2015; pp. 148–152. [CrossRef]
10. Goodman, D.; Nash, R. Subjective quality of the same speech transmission conditions in seven different countries. In Proceedings of the ICASSP 1982 IEEE International Conference on Acoustics, Speech, and Signal Processing, Paris, France, 3–5 May 1982; Volume 7, pp. 984–987. [CrossRef]
11. Damiani, E.; Howlett, R.J.; Jain, L.C.; Gallo, L.; de Pietro, G. Intelligent Interactive Multimedia Systems and Services. In *Smart Innovation, Systems and Technologies*; Springer International Publishing: Cham, Switzerland, 2015.
12. Ipsic, I. (Ed.) *Speech and Language Technologie*; InTech: London, UK, 2011.
13. Benesty, J.; Sondhi, M.M.; Huang, J. *Springer Handbook of Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2007.
14. Stupakov, A.; Hanusa, E.; Bilmes, J.; Fox, D. Cosine-a Corpus of Multi-Party Conversational Speech in Noisy Environments. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; IEEE: Piscataway Township, NJ, USA, 2009; pp. 4153–4156.
15. Richey, C.; Barrios, M.A.; Armstrong, Z.; Bartels, C.; Franco, H.; Graciarena, M.; Lawson, A.; Nandwana, M.K.; Stauffer, A.; van Hout, J.; et al. Voices obscured in complex environmental settings (voices) corpus. *arXiv* **2018**, arXiv:1804.05053.
16. Cooke, M.; Barker, J.; Cunningham, S.; Shao, X. An audiovisual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* **2006**, *120*, 2421–2424. [CrossRef] [PubMed]
17. Hou, J.-C.; Wang, S.-S.; Lai, Y.-H.; Tsao, Y.; Chang, H.-W.; Wang, H.-M. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Trans. Emerg. Top. Comput. Intell.* **2018**, *2*, 117–128. [CrossRef]
18. Ephrat, A.; Mosseri, I.; Lang, O.; Dekel, T.; Wilson, K.; Hassidim, A.; Freeman, W.T.; Rubinstein, M. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.* **2018**, *37*, 112. [CrossRef]
19. Bailly-Bailli´ere, E.; Bengio, S.; Bimbot, F.; Hamouz, M.; Kittler, J.; Mari´ethoz, J.; Matas, J.; Messer, K.; Popovici, V.; Por´ee, F.; et al. The banca database and evaluation protocol. In Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication, Guildford, UK, 9–11 June 2003; Springer: Berlin/Heidelberg, Germany, 2003; pp. 625–638.
20. Lee, B.; Hasegawa-Johnson, M.; Goudeseune, C.; Kamdar, S.; Borys, S.; Liu, M.; Huang, T. Avicar: Audio-visual speech corpus in a car environment. In Proceedings of the Eighth International Conference on Spoken Language Processing, Jeju Island, Korea, 4–8 October 2004.
21. Gogate, M.; Dashtipour, K.; Adeel, A.; Hussain, A. CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement. *Inf. Fusion* **2020**, *63*, 273–285. [CrossRef]
22. Gogate, M.; Dashtipour, K.; Hussain, A. Visual Speech in Real Noisy Environments (VISION): A Novel Benchmark Dataset and Deep Learning-based Baseline System. *Proc. Interspeech* **2020**, *2020*, 4521–4525.

23. Schinkel-Bielefeld, N.; Zhang, J.; Qin, Y.; Leschanowsky, A.K.; Fu, S. Perception of coding artifacts by nonnative speakers—A study with Mandarin Chinese and German speaking listeners. *AES J. Audio Eng. Soc.* **2018**, *66*, 60–70. [CrossRef]
24. Milner, B.; Almajai, I. Noisy audio speech enhancement using Wiener filters derived from visual speech. In Proceedings of the International Workshop on Auditory-Visual Speech Processing (AVSP), Hilvarenbeek, The Netherlands, 1–3 September 2007. ISCA, Archive.
25. Stiefelhagen, R.; Garofolo, J. *Multimodal Technologies for Perception of Humans*; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4122.
26. Hashmi, A.; Kalashnikov, A.N. Sensor data fusion for responsive high resolution ultrasonic temperature measurement using piezoelectric transducers. *Ultrasonics* **2019**, *99*. [CrossRef] [PubMed]
27. Zandan, N. The Power of Pause. Available online: https://www.quantifiedcommunications.com/blog/the-power-of-pause (accessed on 6 June 2020).
28. Fosler-Lussier, E.; Morgan, N. Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Commun.* **1999**, *29*, 137–158. [CrossRef]
29. Facts about Speech intelligibility: Human Voice Frequency Range. Available online: https://www.dpamicrophones.com/mic-university/facts-about-speech-intelligibility (accessed on 1 June 2020).
30. Bhagat, R.; Kaur, R. Improved Audio Filtering Using Extended High Pass Filters. *Int. J. Eng. Res. Technol.* **2013**, *2*, 81–84.
31. Kirubagari, B.; Selvaganesh, R. A Noval Approach in Speech Enhancement for Reducing Noise Using Bandpass Filter and Spectral Subtraction. *Bonfring Int. J. Res. Commun. Eng.* **2012**, *2*, 5–8.
32. Hansen, M. Assessment and Prediction of Speech Transmission Quality with an Auditory Processing Model. Ph.D. Thesis, University of Oldenburg, Oldenburg, Germany, 1998.
33. Lawrie, J.B.; Abrahams, I.D. A brief historical perspective of the Wiener-Hopf technique. *J. Eng. Math.* **2007**. [CrossRef]
34. Upadhyay, N.; Jaiswal, R.K. Single Channel Speech Enhancement: Using Wiener Filtering with Recursive Noise Estimation. *Procedia Comput. Sci.* **2016**. [CrossRef]
35. Plapous, C.; Marro, C.; Scalart, P. Improved Signal-to-Noise Ratio Estimation for Speech Enhancement. *IEEE Trans. Audio Speech, Lang. Process.* **2006**, *14*, 2098–2108. [CrossRef]
36. Scalart, P. Wiener Filter for Noise Reduction and Speech Enhancement. MATLAB Central File Exchange. 2020. Available online: https://www.mathworks.com/matlabcentral/fileexchange/24462-wiener-filter-for-noise-reduction-and-speech-enhancement (accessed on 10 June 2020).
37. Mic University. The Basic about Distortion in Mics. DPA Microphones. 2018. Available online: https://www.dpamicrophones.com/mic-university/the-basics-about-distortion-in-mics (accessed on 2 June 2020).
38. International Electrotechnical Commission. *IEC 60.268 Sound System Equipment, Part 2: Explanation of General Terms and Calculation Methods*; IEC: Geneva, Switzerland, 1987.
39. Blagouchine, I.V.; Moreau, E. Analytic method for the computation of the total harmonic distortion by the cauchy method of residues. *IEEE Trans. Commun.* **2011**. [CrossRef]