# A Subjective Method to Estimate the Voice Quality for Speech Watermarking Based on Improved Spread Spectrum Technique

SHERVIN SHOKRI, MAHAMOD ISMAIL, NASHARUDDIN ZAINAL
Department of Electrical, Electronic, and Systems Engineering
University Kebangsaan Malaysia
Jalan Reko, 43600 Bangi, Selangor
MALAYSIA
shshokri,mahamod,nash@eng.ukm.my

***Abstract: -*** This paper investigates the measure of voice quality for a digital speech watermarking scheme using a subjectivity method. Data rate, inaudibility, and robustness are considered more than voice quality in speech watermarking. Since any technique in speech telecommunication should have acceptable quality, the perceptual measurement of voice quality is seen as an important topic in audio and speech watermarking. The proposed scheme is simulated and then evaluated on the basis of the perceptual quality of received voice (speech) using mean opinion score (MOS) of 40 participants. Experimental results indicate that the average MOS at the receiver end is 2.75 out of 5. The results show that the perceptual audible quality of the proposed algorithm is between poor and fair.

***Key-Words: -*** Improved spread spectrum, linear predictive (LP), MOS; BCH-code, speech watermarking

## 1 Introduction

In the early 1990s, watermarking as a modern technique was initiated for the hiding data. In this technique the private data as a guest signal (text or number) is embedded in the host signal (Image, video and audio) with a digital form. The main objectives of the digital watermark can be divided in three parts: undelectability (inaudibility), capacity (data rate), and robustness. Signal quality is one of the important factors that has not been fully considered in watermarking[1]. To avoid confusion between the concept of quality and acoustical invisibility or statistical un-detectability, we should define the meaning of these terms. Un-detectability is defined as an embedding process without sensitivity for human sensory system (HSS), while quality is the estimation of received signals from effects channel. In[2, 3] a speech watermarking scheme was proposed which is an interesting reference for new investigators in speech watermarking. Although voice quality was noted in these studies, but nothing was done for measure the voice quality. With this view in mind, the current study uses a practical approach to measure the voice quality of the proposed scheme. One of the most popular techniques for watermarking is spread spectrum (SS). SS can generate the conditions for implementing the embedded watermarks in any transform domain. SS is able to transform a narrow-band signal (a message that should be transmitted) to the wideband signal by modulating a broadband carrier signal. The watermark information can spread over a large set of samples with the utilization of a chip-rate parameter. Pseudo noise (PN) sequence is employed to spread the watermark spectrum in the frequency and time domains. The PN sequence can also be utilized as a secret key to protect the data of authorized users on both sides of the system[4, 5]. In the proposed scheme uses a new strategy for informed embedding has been defined to decrease the host interference in the watermark signal which is based on the traditional SS. Malvar and Florêncio[6] have shown a correlation between the host and the watermark signal make an improvement in watermark robustness, and since this method is based on the SS technology they called it improved spread spectrum (ISS).

In speech watermarking, the speech signal intended to cover the private data, and that's why the output speech should be have a good quality at the receiver. Voice quality is normally measured subjectively, where the most popular method is to interrupt users from listening and ask the users what they think. MOS was standardized by the International Telecommunications Union (ITU; P.800), a United Nations body responsible for telecommunications standardization[7]. In this measure technique the same series of sound files were played, and people were asked to rank the sound files on a scale of 1 to 5 in terms of voice

quality as seen in Table1. In our study, forty participants joined the practical test to estimate the speech quality at the receiver.

**Table 1**. Applications in each class

| Quality of the speech | Score |
|---|---|
| Excellent | 5 |
| Good | 4 |
| Fair | 3 |
| Poor | 2 |
| Bad | 1 |

This study mainly focused on the measuring the voice quality for the proposed scheme in [2, 3] by utilizing the mean MOS technique. The rest of this paper is organized as follows. Section 2 proposes the speech watermarking scheme and describes an overview of speech watermarking and simulations over the system. Section 3 presents some information about data and speech signals, which were utilized in the system input. In this section, the estimate of the voice quality is also presented in each part of the system. Conclusion and future studies are given in Section 4.

## 2 system overview

The proposed scheme is based on the SS technology because many studies have shown that this technology is one of the robust techniques in watermarking. Fig. 1 shows the block diagram for the transmitter side (encoder). The emitter can be divided into three major sections. First, the error-control coding is employed by channel coding to increase the reliability of the system. The next level involves spreading the watermark signal over the available frequency band. Finally, the watermark is embedded in the speech signal by utilizing perceptual methods. The ISS and linear prediction coefficients (LPC) filters are utilized for spectral shaping of the watermark spectrum. The shaped watermark is then embedded in the speech signal.

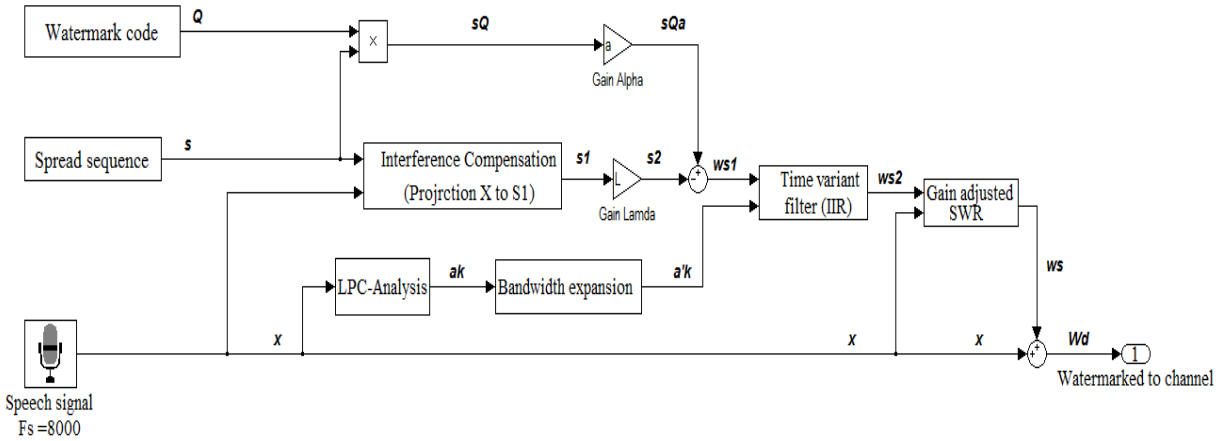


**Fig. 1.** Transmitter blocks diagram.

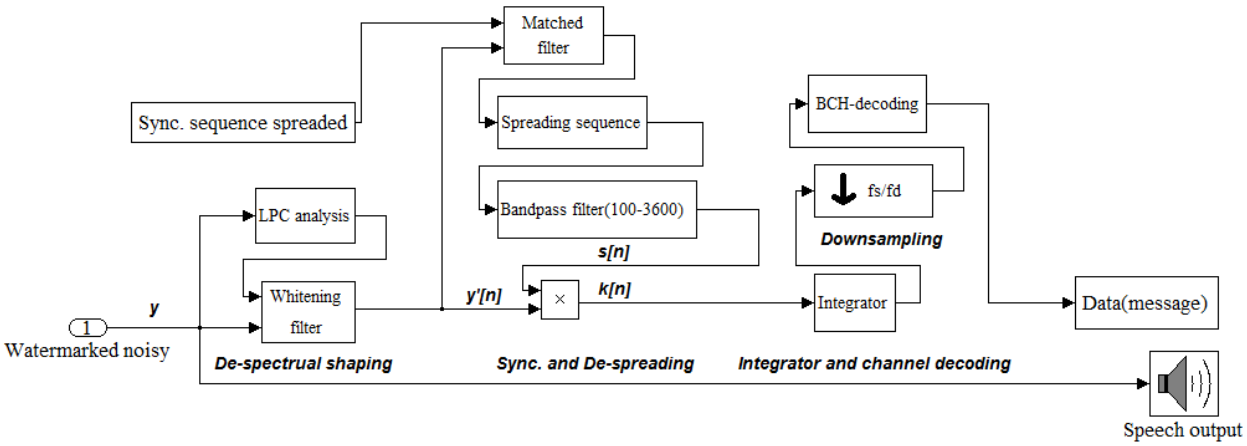


**Fig. 2**. Receiver blocks diagram.

The resulting signal, which is called the watermarked signal, will be transferred to the channel.
Fig. 2 shows the process of extracting the embedded watermark from the speech signal at the receiver side (decoder). A whitening filter is utilized to undo spectral masking. The signal has to be synchronized to perform watermark extraction and de-spreading can provide the payload data. The error correction block is utilized to correct errors that occurred because of watermark process and channel conditions [4].

**2.1 Error Control Coding (BCH-Code)**
The watermark code in the transmitter side follows the digital signal process to provide the spreading signal. In this box, the payload data have been encoded with the BCH code. BCH codes are a type of cyclic linear block code that is permitted to encode a large selection of block data and is widely used in channel coding. The BCH code checks the error numbers at the receiver side. If the number of errors is within the ability range, then the errors will be corrected, otherwise, the errors will only be detected [4]. In the next step, a synchronization sequence is added to the watermark signal for system synchronization. The message coding output ($Q$) is called watermark code.

**2.2 Data Spreading (spreading spectrum)**
SS is one of the most popular techniques in the watermarking. SS can generate the conditions to embed the watermarks in any frequency or time domain [8]. Whit this technique, a spread watermark signal $v(n)$ is achieved by spreading the bits of $Q_m$ (in $\{0,1\}$) over a set samples $N_b$ of sequence $s(n)$ (as the vector $s = [s(0),s(1),\ldots\ldots s(N_b\text{-}1)]^T$). The $s(n)$ signal is presented by the vector $s$, which is made by the $PN$ sequence (in $\{-1,+1\}$). The $v(n)$ signal is given by

$$v(n) = Q(n)s(n) \tag{1}$$

$$v(n) = \sum_{m=0}^{M-1} a_m s(n - mN_b) \tag{2}$$

where the $a_m$ symbol in $\{-1, 1\}$ is given by $a_m = 2k\text{-}1$. Therefore, the vector direction will adjust by the $a_m$ values.

$$v_m = \begin{cases} +s & if \quad a_m = +1 \\ -s & if \quad a_m = -1 \end{cases} \tag{3}$$

**2.3 Data Embedding**
Embedding the watermark signal with maximum energy and minimum perceptual distortion are the main aim of data embedding. The watermarked signal is achieved by embedding the spread watermark sequence $v(n)$ into the speech signal. Adding the watermark in the speech signal without any terms will create a large interference in watermark signal, therefore the embedding can be controlled better by utilizing ISS technique in terms of temporal energy of speech signals [9]. This can achieve by projection the speech signal over the spreading watermark [6, 10, 11]. The linear form of ISS embedding can be formulated as follows:

$$w[n] = x[n] + \mu(\tilde{x}, Q)s[n] \tag{4}$$

where:

$$\tilde{x} \triangleq \frac{\langle x, s \rangle}{\|s\|} \tag{5}$$

Here, $\tilde{x}$ is the projection of vector $x$ on vector $S$.

$$\langle x, s \rangle = \frac{1}{N}\sum_{I=0}^{N-1} x_i s_i \text{ , and } \quad \|x\| \triangleq \langle x, x \rangle \tag{6}$$

The function $\mu(\tilde{x},Q)$ is a linear function of speech signal ($x$). Vector $s$ has the $N$ signal sample and the bit rate is 1/N/bits/sample. The following equations were derived on the basis of Fig. 1

$$s_1 = \tilde{x}s \tag{7}$$

$$s_2 = \lambda s_1 \tag{8}$$

$$w_{s1} = \alpha Qs - \lambda\tilde{x}_k s = (\alpha Q - \lambda\tilde{x})s \tag{9}$$

The parameters $a$ and $\lambda$ are used to control the distortion level and removal of the carrier distortion on the detection statistics. In the traditional SS, these parameters are set to $a=1$ and $\lambda = 1$. In order to decrease the perceptual distortion, a linear prediction coefficients (LPC) is utilized to estimate the spectrum of the speech signal by vocal formants coefficients ($a_k$) of the speech signal $x(n)$ (as shown in Figs. 3 and 4). The spectral of the spread watermark $v(n)$ is made to be similar to the speech signal by being passing through a time variant filter ($IIR$), which is created from coefficients $a_k$ (Fig. 1) [12, 13]. The vocal transfer function and LPC transfer functions are defined as follows:

$$H(z)=\frac{g}{A(z)}=\frac{g}{1+\sum_{k=1}^{p}a_k z^{-k}} \quad (10)$$

$$A(z)=1+\sum_{k=1}^{p}a_k z^{-k} \quad (11)$$

The LPC order $p$ is expressed as

$$p=2+\frac{F_s}{1000}\ ,\ F_s=\text{sample frequency} \quad (12)$$

Thus, for telephone frequency sampling (8000 kHz), the LPC order is 10.

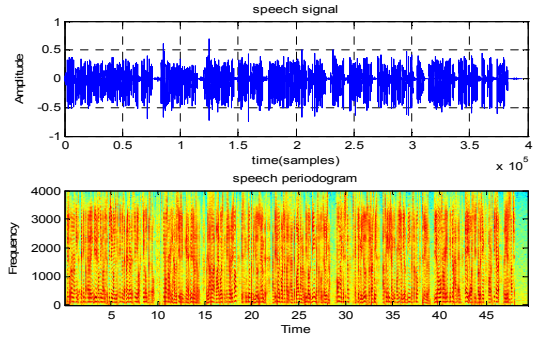


**Fig. 3**. Speech signal and periodogram.

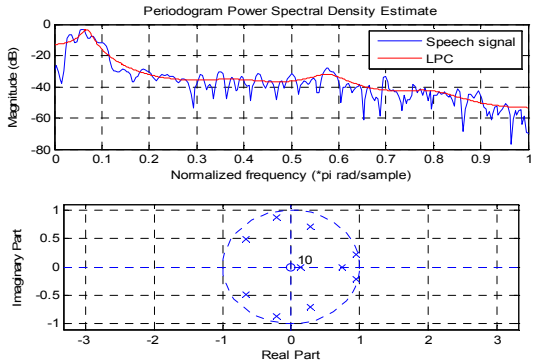


**Fig. 4.** Top: speech signal and LPC cover; Bottom: z plane of the polls.

The bandwidth expansion technique is utilized to avoid interference between the watermark and speech signal. In this technique, the filter coefficients ($a_k$) are adjusted by the $\gamma$ factor. This factor can create a small gap between two signals to protect the watermark signal from speech formants [3, 14].

$$a'_k = a_k \gamma^k \quad 0 \le k \le order \quad (13)$$

The best value for the $\gamma$ factor usually ranges from 0.90 to 0.97. In this case, by adjusting the $\gamma$ factor in 0.9, a bandwidth expansion appears at the spectral peaks by moving all the poles to the centre of the unit circle (see Figs. 5 and 6) [14].

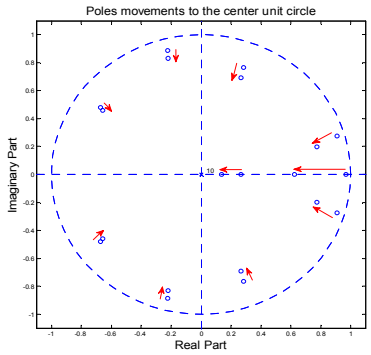


**Fig. 5**. Bandwidth expansion.

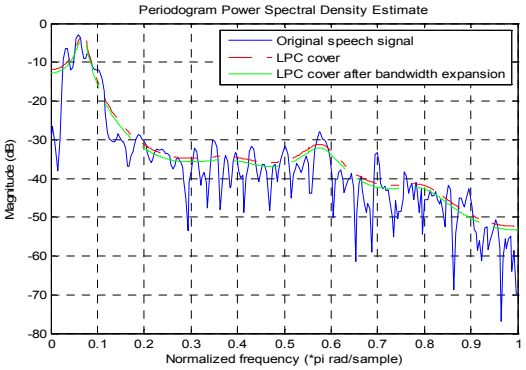


**Fig. 6**. Bandwidth expansion by moving the poles to the centre of the unit circle.

In the final step to embed the watermark in speech a variable gain ($\lambda_G$) is utilized to obtain the desired signal-to-watermark ratio (SWR) [15].

$$w_s = w_{s2}\lambda_G \quad (14)$$

After spreading and shaping, the watermark signal can be embedded in a speech signal by simple adding.

$$w[n]=x[n]+w_s[n] \quad (15)$$

Fig. 7 shows the group delay between $a_k$ and $a'_k$ coefficients during the LPC filter. This delay is typically considered in the embedding process. The simulations show that the delay is reduced after bandwidth expansion. This delay is shown in Equation (16) [2].

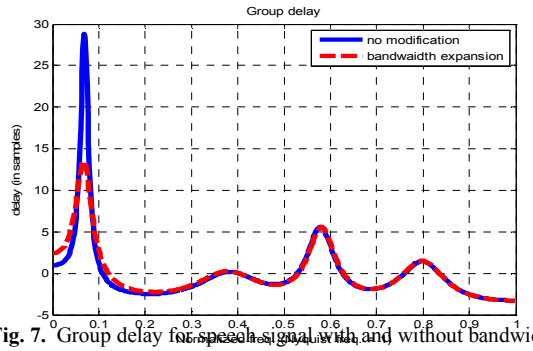


**Fig. 7.** Group delay for speech signal with and without bandwidth expansion.

$$w[n] = x[n-M] + w_s[n] \tag{16}$$

Here, $M$ is a delay that is practically set to 100 ms. Figs. 8 and 9 illustrate the embedding process. Fig. 8 shows the power spectral density (PSD) of the watermark, speech, and watermarked signal. Fig. 9 shows the waveforms of the original signal and the watermarked signals with and without spectral shaping.

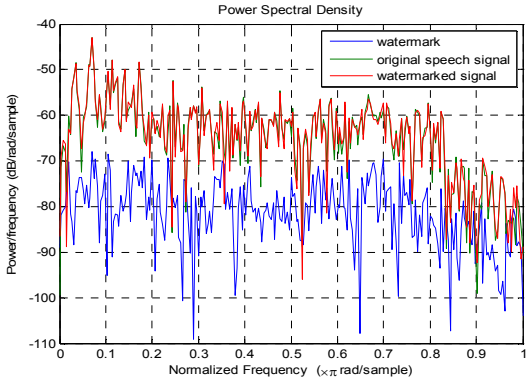


**Fig.8.** Power spectrum density for watermark, speech and watermarked signal.

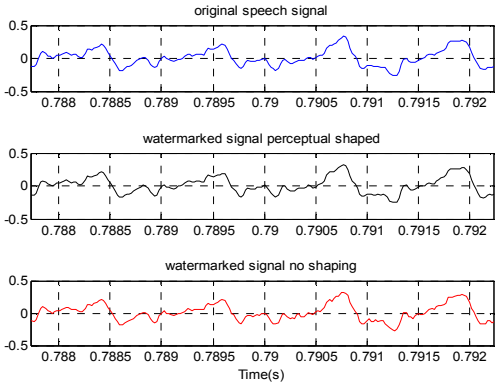


**Fig. 9.** Original Speech, watermarked shaped and watermarked without shaping signal.

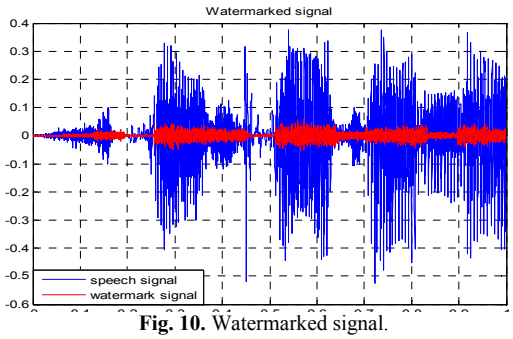


**Fig. 10.** Watermarked signal.

Fig.10 shows the watermarked signal ($w[n]$), which is made by the speech signal ($x[n]$) and the watermark signal ($w_s[n]$) by utilizing perceptual shaping and ISS with a different colour for better visualization.

## 2.4 Speech Channel Communication

The real dynamic channel model is not considered in this study. Therefore, the static channel model was simulated by adding white Gaussian noise channel (AWGN) to the emitted watermarked signal. Most researchers are looking at an SNR range from 0 dB to 40 dB for an AWGN channel. Fig. 11 shows the simple schematic of the AWGN channel by adding noise to the watermarked signal.

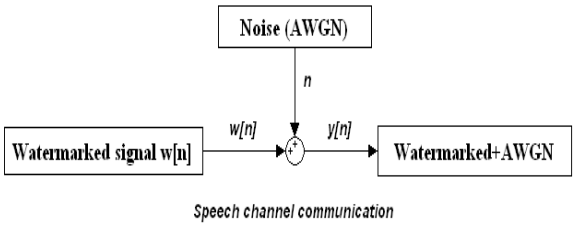


**Fig. 11**. Communication channel.

$$y[n] = w[n] + n[n] \tag{17}$$

The noisy watermarked signal is shown in Fig. 12.

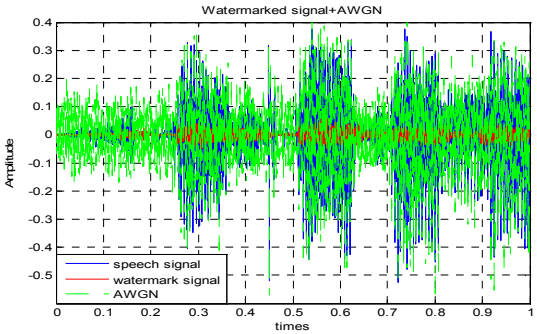


**Fig.12.** Watermarked signal in receiver side.

### 2.5 Whitening

At the next step, a whitening filter is used to undo the spectral shaping from the incoming signal at the receiver side. The whitening filter is an inverse filter $A(z)$, which calculates the prediction error. The incoming signal $y[n]$ is passed through the LPC filter again to extract the coefficients. $A(z)$ then utilizes the coefficients to undo spectral shaping. Note that for the bandwidth expansion, the zeroes are also returned to their place by the whitening filter. After inverse LPC filtering, the speech signal becomes the periodic pulse. Of course, the spectral shape of the signal at the receiver side is not equal to the original speech signal, but the spectral shape of the signal is expected to be similar. The whitening filter output shown in Fig. 13 [4].

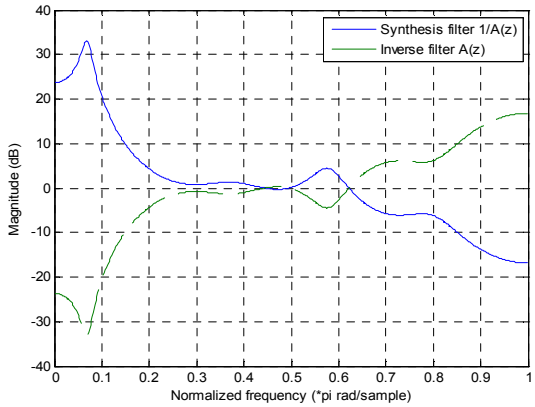


**Fig. 13.** Whitening filters output.

### 2.6 Synchronization

Because, the receiver does not know the source of the message, a blind detector is utilized at the receiver side. For synchronization, a special synchronization sequence that is known by the transmitter and receiver has been added to the payload data (in the encoder). De-spreading will start when the impulse response from the reverse of the spread synchronization sequence is sensed by the matched filter.

### 2.7 Data De-spreading

De-spreading is performed after synchronization to yield the payload data by multiplying the signal at the spread sequence. The PN sequence ($S(n)$) is passed through the BPF (100–3400) filter and then multiplied to the $y'[n]$ signal.

$$k_r[n] = y'[n]s[n] \tag{18}$$

with consideration to the above channel noise model, sufficient statistic is calculated as

$$k_r = \langle y', s \rangle / \|s\| = \alpha Q + (1-\lambda)\tilde{x} + n \tag{19}$$

where $n_r$ is a very low value of the noise that remainder in the signal. Therefore, for $\lambda \approx 1$ in Equation (19), more influence of the $\tilde{x}$ is reduced or removed from $k_r$.

### 2.8 Decoding and Detection

A simple integrator is utilized for detection by knowing the data bits and length of them. Bit detection is conducted by integrating over the period of one data bit and quantizing the result to 1 or -1. For one received data bit, $k'$ is

$$k' = sign\left( \sum_{n=0}^{bit-length} k_r[n]_i \right) \tag{20}$$

where $i$ is the current bit interval. At this step, the process to reduce the sampling rate to the binary symbol rate is performed via down-sampling $[x\downarrow]$. Finally, the BCH decoder is used for error correction as much as possible [4].

## 3 Simulation and Experimental Results

In this section, practical test is considered to evaluate the voice quality. For test preparation, Wavepad Sound Editor Masters edition version 5.33 is used for speech recording. The software is set to 8 kHz sample rate in mono channel. A Philips SHM3100U/97 In-ear Earphone microphone is also used for voice recording. The technical data for the microphone is as follows:

- Wired
- In-the-ear
- Ear Bud Design
- 50 mW of Max Power Input
- Built-in Microphone
- 12 Hz – 20000 Hz Headset Frequency Response
- 80 Hz – 15000 Hz Microphone Frequency Response
- 106 dB/mW Headset Sensitivity at 1 kHz

The pre-emphasized for speech signal is done by a first-order filter whose transfer function is $H(z) = 1 - 0.95z^{-1}$. A 30 ms Hamming window in LPC filter is used with the 2/3 (20 ms) overlapped frames, and a message signal (watermark) is converted into binary with a length of 520 bits. During The process of receiving the watermarked signal, an AWGN in 20 dB is used as a channel

noise. The MOS technique as a subjective technique, is used to measure voice quality in the proposed scheme. Forty participants in two different groups participated in the practical test, and their averaged votes are utilized to measure the quality of the processed signal. The Perceptual quality of the signal is estimated in four positions: the speech signal input, the watermarked signal with and without shaping, and the signal at the receiver. The experimental results are shown in Tables 2, 3 and Fig. 14.

**Table 2**. MOS score for group 1

| MOS score | 1 | 2 | 3 | 4 | 5 | Mean |
|---|---|---|---|---|---|---|
| Original Signal | - | - | - | - | 20 | 5 |
| Watermarked No Shaping | - | - | 3 | 17 | - | 3.85 |
| Watermarked Shaping | - | - | - | 3 | 17 | 4.85 |
| Receiving Signal | - | 5 | 15 | - | - | 2.75 |

**Table 3**. MOS score for group 2

| MOS score | 1 | 2 | 3 | 4 | 5 | Mean |
|---|---|---|---|---|---|---|
| Original Signal | - | - | - | - | 20 | 5 |
| Watermarked No Shaping | - | - | 5 | 15 | - | 3.75 |
| Watermarked Shaping | - | - | - | 4 | 16 | 4.8 |
| Receiving Signal | - | 6 | 14 | - | - | 2.7 |

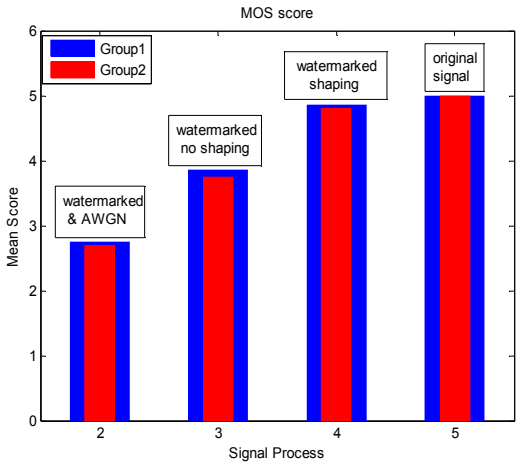


**Fig. 14**. Mean opinion score for speech quality.

The quality analysis results underwent T-test. T-test assesses whether the means of two groups are statistically different from each other. This analysis is appropriate whenever you want to compare the means of two groups, particularly for the post-test-only two groups randomized experimental design [16]. A statistical software which is called *G*power* (version 3.0.10) helped to estimate the quality of results. *G*power* can calculate the *t* value and draw the distribution plot for the sample results. In this case, the result for each position is calculated. For example, the watermarked shaping results of the practical test in Tables 2 and 3 are estimated with the output *G*power*. The distribution plot is shown in Fig. 15.

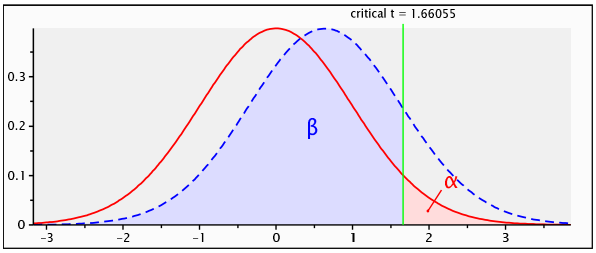


**Fig. 15**. PDF probability distribution for watermarked shaping by *G*power*.

The plotting results for the other stages are explained the same as above:

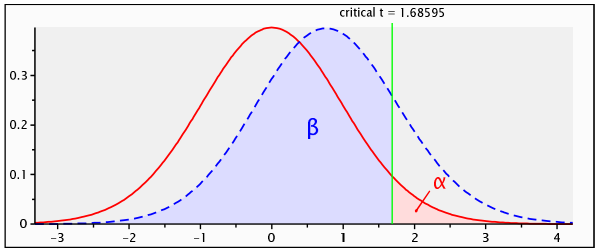


**Fig. 16**. PDF probability distribution for watermarked no shaping by *G*power*.

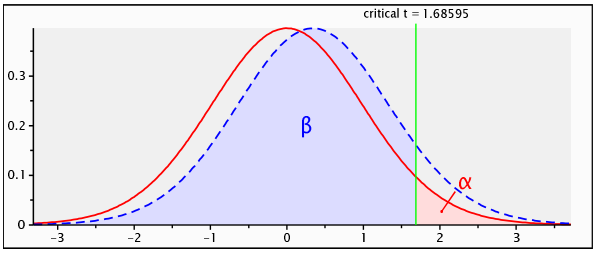


**Fig. 17**. PDF probability distribution for watermarked and AWGN by *G*power*.

The distribution graph in Figs. 15 to 17 shows the big overlap in two groups of samples, which means that a small difference exists between the two groups in the results. All results and figures show that the quality of the results estimation is the best range; thus, the practical MOS results from the 40 respondents are claimed reliable.

## 4 Conclusion

The quality of watermarking is not regarded as a core topic in previous works on speech watermarking. Therefore, the current study aims to measure the received signal voice quality with a

subjective method, which is proposed in [2, 3] as one of the robust watermarking schemes. A watermark signal was evaluated in the process of embedding and extraction in the real speech signal recorded. As a subjective method, the P. 800 (MOS) standardization by the international telecommunications union (ITU) was used as a practical technique to measure voice quality. The practical test was performed by employing 40 participants, organized into two different groups. The quality analysis results were estimated by using the statistical method called T-test. The experimental results show that the inaudibility of this algorithm after watermarked shaping is near excellent $(\approx 4.85)$, which is extraordinary. However, the average score at the receiver side shows that the MOS score is so close to the fair mean score $(\approx 2.75)$. We strongly believe that the speech quality in the receiver side needs to be studied for further improvement. Therefore, future studies will focus on the noise removal method of improving the voice quality.

*References:*


[1] N. Cvejic and T. Seppanen, "Channel capacity of high bit rate audio data hiding algorithms in diverse transform domains," in *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on*, 2004, pp. 84-88.

[2] M. Faundez-Zanuy, *et al.*, "Speaker verification security improvement by means of speech watermarking," *Speech communication,* vol. 48, pp. 1608-1619, 2006.

[3] M. Faundez-Zanuy, *et al.*, "Speaker identification security improvement by means of speech watermarking," *Pattern Recognition,* vol. 40, pp. 3027-3034, 2007.

[4] S. Shokri, *et al.*, "Voice quality in speech watermarking using spread spectrum technique," in *Computer and Communication Engineering (ICCCE), 2012 International Conference on*, 2012, pp. 169-173.

[5] S. Shokri, *et al.*, "Error probability in spread spectrum (SS) audio watermarking," in *Space Science and Communication (IconSpace), 2013 IEEE International Conference on*, 2013, pp. 169-173.

[6] H. S. Malvar and D. A. Florêncio, "Improved spread spectrum: a new modulation technique for robust watermarking," *Signal Processing, IEEE Transactions on,* vol. 51, pp. 898-905, 2003.

[7] A. W. Rix, "Perceptual speech quality assessment-a review," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, 2004, pp. iii-1056-9 vol. 3.

[8] M. Davarynejad, *et al.*, "Evolutionary hidden information detection by granulation-based fitness approximation," *Applied Soft Computing,* vol. 10, pp. 719-729, 2010.

[9] H. Hering, *et al.*, "Safety and security increase for air traffic management through unnoticeable watermark aircraft identification tag transmitted with the VHF voice communication," in *Digital Avionics Systems Conference, 2003. DASC'03. The 22nd*, 2003, pp. 4. E. 2-41-10 vol. 1.

[10] P. Zhang, *et al.*, "Robust Audio Watermarking Based on Extended Improved Spread Spectrum with Perceptual Masking," *International Journal of Fuzzy Systems,* vol. 14, pp. 289-295, 2012.

[11] M. Hagmüller, *et al.*, "Speech watermarking for air traffic control," *Watermark,* vol. 8, p. 10, 2004.

[12] B. Kotnik, *et al.*, "Data transmission over GSM voice channel using digital modulation technique based on autoregressive modeling of speech production," *Digital Signal Processing,* vol. 19, pp. 612-627, 2009.

[13] U. Zolzer, *DAFX: Digital Audio Effects*: Wiley Publishing, 2011.

[14] K. N. Ramamurthy and A. S. Spanias, "MATLAB® Software for the Code Excited Linear Prediction Algorithm: The Federal Standard-1016," *Synthesis Lectures on Algorithms and Software in Engineering,* vol. 2, pp. 1-109, 2010.

[15] A. Deshpande and K. Prabhu, "A substitution-by-interpolation algorithm for watermarking audio," *Signal Processing,* vol. 89, pp. 218-225, 2009.

[16] S. Wallenstein, *et al.*, "Some statistical methods useful in circulation research," *Circulation Research,* vol. 47, pp. 1-9, 1980.

**Shervin Shokri** (shshokri@eng.ukm.my) He received the BSc degree in Electrical and Electronics from Azad university, of Iran in 2000, the MSc degree in Communication and Computer from University Kebangsaan Malaysia (UKM), Malaysia in 2010. Now he is a PhD candidate in the Department of Electrical, Electronic and System Engineering, Faculty of Engineering and Built Environment, UKM. His research interests include the Digital Signal Processing, Audio and Speech processing in wireless communication and networking. He is also a member of the Institute of Electrical and Electronics Engineers (IEEE) USA.

**Mahamod Ismail** (mahamod@eng.ukm.my) joined the Department of Electrical, Electronic and System Engineering, Faculty of Engineering and Built Environment, UKM in 1985 and currently he is a Professor in Communication Engineering. He received the BSc degree in Electrical and Electronics from University of Strathclyde, U.K. in 1985, the MSc degree in Communication Engineering and Digital Electronics from University of Manchester Institute of Science and Technology (UMIST), U.K. in 1987, and the PhD from University of Bradford, U.K. in 1996. His research interests include mobile and satellite communication, and wireless networking particularly on the radio resource management for the next generation wireless communication network. He is Senior Member of the Institute of Electrical and Electronics Engineers (IEEE) USA, and the chair of IEEE Malaysia Section (2011-2012).

**Nasharuddin Zainal** (nash@eng.ukm.my) He received the B.E. degree from Tokyo Institute of Technology in 1998, M.E. degree from The National University of Malaysia in 2003 and the Ph.D. degree from Tokyo Institute of Technology in 2010. He is also a member of the Institute of Electrical and Electronics Engineers IEEE (USA), corporate member of The Institution of Engineers Malaysia and certified Professional Engineer of Board of Engineers Malaysia and his researches are on image and video processing, pattern recognition and robotics.