An Efficient Tampering Detection and Localization Method for Speech Signals

Sharr Wint Yee Myint, Twe Ta Oo
University of Computer Studies, Yangon, Myanmar
pyaylay963@gmail.com

Abstract

The development in digital technologies and advanced speech analysis/synthesis tools these days enable speech signals to be tampered without leaving any perceptual clues. As an important information carrier, the integrity of speech signals should be strictly confirmed and digital watermarking can be used for this purpose. In this paper, a self-embedding speech watermarking scheme is proposed for detection of unauthorized manipulation (tampering) and localization. Firstly, hash code is generated from original speech signal and then this hash code (as a watermark) is embedded directly into the signal itself without affecting the original quality. Once tampering occurred, the watermark in the tampered segment should be destroyed and thus the receiver can detect the tampered position. Experimental results show that the proposed method provides not only inaudibility and blindness but also fragility against tampering and detects the position of tampering.

Keywords: *Speech watermarking, inaudibility, fragility, authentication, tampering detection and localization.*

1. Introduction

Digital technologies have greatly facilitated the transmission of speech signals, but they have also increased the need for protection of signals from any misuse. Speech watermarking is the art of embedding watermark information (e.g. bits, logo etc.) into the host speech signal and using it to detect tampering or copyright violation.

Based on application areas, speech watermarking is usually categorized into robust and fragile watermarking. Robust watermarking is mainly used for copyright protection and fragile watermarking stands for authentication and tampering detection of speech signals. A watermark is said to be fragile if it is destroyed as soon as the watermarked signal undergoes any manipulation. The destroyed watermark could provide an evidence that the signal has been tampered. Basically, speech tampering detection schemes come down to two main categories: i) schemes just verify the originality of speech without localizing the tampering and ii) schemes that can localize the tampering regions.

To effectively detect tampering, speech watermarking should generally satisfy inaudibility to human auditory system (HAS) (watermarked speech sounds similar to the original un-watermarked speech), blindness (extract watermarks without referring to the host signal), and fragility against malicious tampering.

The rest of the paper is organized as follows. It discusses related work for tampering detection system in section 2, generalized tampering detection scheme in section 3, the proposed system in section 4, experimental results in section 5, and conclusion in section 6, respectively.

2. Related Work

In the last few years, several fragile watermarking techniques have been proposed for authentication and tampering detection of speech signals. Sarreshtedari et al. proposed a self-recovery watermarking method for tampering detection. Original speech signal, which is compressed with a speech codec and protected against tampering by proper channel coding, is embedded into itself. The embedded watermark (channel coded output) helps the receiver to detect tampering and localize it, and to recover the lost content with a certain quality as well [1].

Wang et al. proposed a tampering detection scheme for speech signals based on formant enhancement-based watermarking. Watermarks are embedded as slight enhancement of formant by symmetrically controlling a pair of linear spectral frequencies (LSFs) of corresponding formant. The core idea is to provide inaudibility, fragility against tampering, and robustness against meaningful processing [2].

This paper focuses on an efficient tampering detection and localization method for speech signals based on a combined approach of fragile watermarking with hash function. By using fragile watermarking, the proposed method not only can detect tampering but also can localize it. In addition, the proposed method uses hash generated from original speech as watermark and thus it provides inaudibility and blindness.

3. Generalized Tampering Detection Scheme

Speech watermarking has already been applied in different application areas, especially for copyright protection and tampering detection. Examples of copyright protection applications are air traffic control application [3], audio watermarking [4] [5] and pirate recorder detection [6]. In addition, speech authentication has always been one of the most important applications of speech watermarking. In such methods, the integrity of the received speech signal is determined by examining an embedded fragile watermark vulnerable against malicious attacks [7] and [8].

A watermarking system can be modeled as a communication system. Watermark embedding process is considered as the signal transmitter. The watermark can be seen as the signal to be transmitted, the host can be seen as the noise, any operation to the watermarking (e.g. compression, noising, etc.) can be modelled as the communication channel, and the detection of the embedded watermark corresponds to the detection of signal with the presence of noise at the receiver in the communication scenario. The communication channel may be any kind of wire or wireless transmission channel such as radio broadcasting, Internet and mobile communication channels.

General framework for checking whether tampering has occurred to speech signals during transmission is shown Fig. 1. This scheme consists of watermark embedding, extraction, and tamper detection processes. On the embedding side, the watermarks h_o is embedded into the speech signal $x(n)$ to construct the watermarked signal $y(n)$. Then $y(n)$ is transmitted. At the receiver side, the watermarks are blindly extracted from $\hat{y}(n)$ (may be tampered or original $y(n)$). The extracted watermarks $\hat{h}_o$ are then compared with h_o to check whether tampering has occurred. If a speech watermarking method satisfies fragility, once tampering occurred, the watermarks in tampered regions will be destroyed. Therefore, tampering could be detected by the mismatched bits between h_o and $\hat{h}_o$. If there is no mismatch, it confirms the integrity of the received signal.

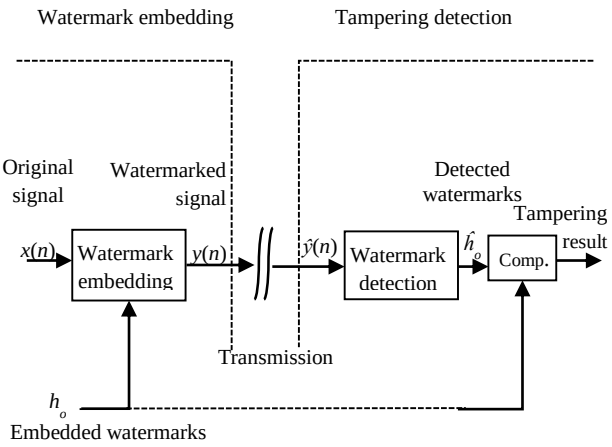

Figure 1. Generalized tampering detection scheme

4. The Proposed System

In this system, a self-embedding speech watermarking scheme is proposed for detecting and localizing tampering. Least significant bit (LSB) replacement method is used for watermark embedding because the effect of replacing the LSB bits with watermarks is less perceptible to HAS and thus provides better inaudibility.

To generate watermark, secure hash algorithm (SHA) is used for the sake of authentication and tampering detection. Hash function takes a message of arbitrary length and creates a message digest of fixed length. There are many well-known hashing algorithms. Among them, SHA-512 is considered very secure and no attacks are known presently. Maximum input size of SHA-512 is $(2^{64}-1)$ bits and output size is 512 bits. Unlike previous versions of SHA, SHA-512 uses different shift amounts and additive constants, and 80 number of rounds for stronger security [9].

The followings are the steps of SHA-512:

1) Preprocessing
   - ⇒ As shown in Fig. 2, the message is padded with 0's so that the length of the message to be an exact multiple of 1024 (N ×1024).
   - ⇒ The message is divided into 1024-bits chunks.

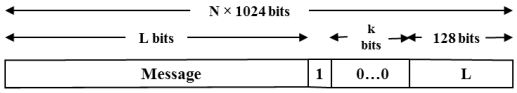

Figure 2. Format of padded message

2) Extension
   ⇒ To achieve better security, sum and sigma functions are used to extend the length of each chunk from 1024-bit to 5120-bit.
3) Compression
   ⇒ Majority and choice functions are used to compress 5120-bit into 512-bit hash code.

4.1. Watermark Generation and Embedding

This section discusses watermark generation and embedding processes of the proposed system.

The self-embedding speech generation framework at the transmitter side is sketched in Fig. 3. Consider a 16-bit 8-kHz sampled speech signal S.

Step 1: The S is divided into N frames, each with size of 128 samples.

$$S = \{f_1, f_2, f_3, \dots, f_N\} \quad (1)$$

$$f_i = \{s_{(1,i)}, s_{(2,i)}, s_{(3,i)}, \dots, s_{(128,i)}\} \quad (2)$$

where $i = \{1, \dots, N\}$.

Step 2: Generate bit pattern from sample values.

$$s_j = \{b_{(15,j)}, b_{(14,j)}, b_{(13,j)}, \dots, b_{(0,j)}\} \quad (3)$$

where $j = \{1, \dots, 128\}$.

Step 3: Retrieve MSB and LSB from each sample. Out of the 16 bits, n_w = 4LSB of each sample are dedicated to the watermark embedding, while the rest n_m= (16 - n_w)MSB are left intact during the embedding process.

$$b_m = \begin{bmatrix} s_1(b_{15}, \dots, b_4) \\ \vdots \\ s_{128}(b_{15} \dots, b_4) \end{bmatrix}, \; b_w = \begin{bmatrix} s_1(b_3, \dots, b_0) \\ \vdots \\ s_{128}(b_3, \dots, b_0) \end{bmatrix} \quad (4)$$

where b_m = MSB bits for watermark generation and b_w = LSB bits to carry watermark.

Step 4: The SHA-512 hash algorithm is used to generate hash bits from $b_m = n_m \times 128 = 12 \times 128 = 1536$ MSB bits.

$$h_o = \text{SHA512}(b_m) \quad (5)$$

where h_o is the original hash data (512 bits).

Step 5: The $b_w = 4 \times 128 = 512$ LSB bits of each frame is replaced with the hash bits h_o of that frame. Finally, the watermarked speech signal S' is produced.

$$S' = \text{Embed}(b_w, h_o) \quad (6)$$

The above steps (steps 2 to 5) are repeated for all frames.

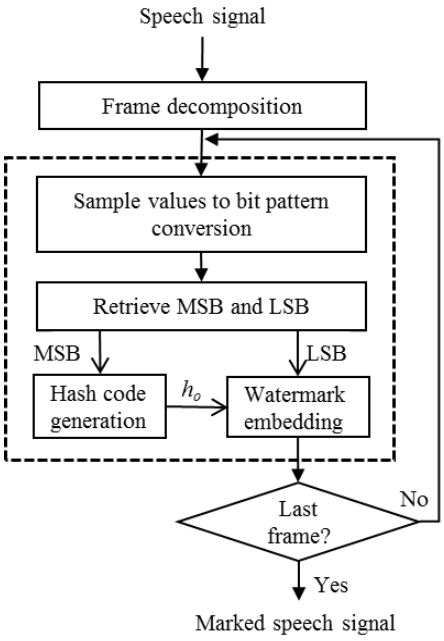

Figure 3. Self-embedding speech generation framework

4.2. Watermark Extraction and Tampering Detection

At the receiver, the detection algorithm is performed blindly on the received signal S' (may be tampered or not tampered). Block diagram of the watermark extraction and tampering detection is shown in Fig. 4.

Step 1 to *Step 3* are the same as the watermark embedding process.

Step 4: For each frame, hash bits are generated using the same algorithm as the transmitter.

$$\hat{h_o} = \text{SHA512}(b'_m) \quad (7)$$

where b'_m = received MSB bits for watermark generation and $\hat{h_o}$ is the generated watermark.

Step 5: Extract the watermarking bits from b'_w LSB bits.

$$\hat{h_e} = \text{Extract}(b'_w) \quad (8)$$

The extracted hash data $\hat{h_e}$ are compared to the generated hash data $\hat{h_o}$ of the same frame. The speech frames are marked as healthy for matching hash data, and tampered otherwise. When a frame is marked as tampered, tampered regions are shown in graph with zero and one stand for the healthy and tampered frames, respectively. The above steps are repeated for all frames.

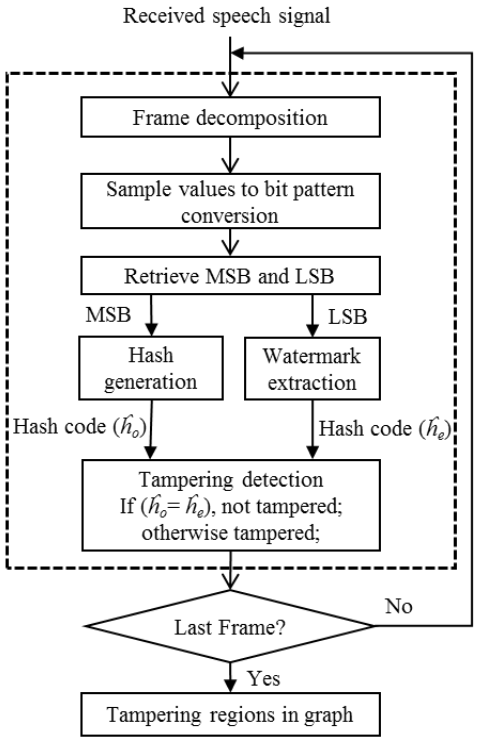

Figure 4. Watermark extraction and tampering detection framework

5. Experimental Results

This section discusses the performance of the proposed system evaluated by using 40 speech files (male/female Burmese and English read speech). Each is a 16-bit, 8 kHz sampled WAVE file. Performance is evaluated based on the quality of watermarked signals (inaudibility) and fragility against tampering. The following section discusses the evaluation results for five speech signals: news 1(f,B), news 2(f,B), news 3(m,B), news 4(f,E), and news 5(m,E) where m = male speaker, f = female speaker, B = Burmese read speech, and E = English read speech.

5.1. Performance Evaluation for Inaudibility

In order to verify inaudibility, signal-to-noise ratio (SNR) and log spectrum distortion (LSD) measures are used. These measures can estimate the degradation between the original and the watermarked speech signals.

5.1.1. Signal to Noise Ratio (SNR)

SNR is used to know the amount by which the signal is corrupted by the noise. It is defined as the ratio of the summed squared magnitude of the clean signal $s(n)$ to the summed squared magnitude of the noise signal (difference between $s(n)$ and watermarked speech signal $\hat{s}(n)$. The SNR in dB is calculated according to Eq. (9).

$$\text{SNR} = 10*\log_{10} \frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N} \{s(n)-\hat{s}(n)\}^2} \; [dB] \qquad (9)$$

The resulting SNR values are given in Table 1. According to the International Federation of Phonographic Industry [10], SNR≥20 dB means good inaudibility and thus the noise introduced by watermark embedding does not affect the speech quality. The SNR values in Table 1 are greater than 47 dB for all files and thus confirms the inaudibility of the watermarked speech signal.

Table 1. SNR results for inaudibility

Signal (8 kHz, 16-bit)	Length (sec)	Total Samples	Watermark bits	SNR (dB)
news 1 (f,B)	7	60416	4LSB	48.52
news 2 (f,B)	8	64512	4LSB	48.62
news 3 (m,B)	9	75904	4LSB	48.94
news 4 (f,E)	10	86016	4LSB	47.11
news 5 (m,E)	11	92032	4LSB	48.84

5.1.2. Log Spectrum Distortion (LSD)

LSD defined in Eq. (10) is used to measure the spectral distance between the original signal and the watermarked signal.

$$LSD = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \left(10 * log_{10} \frac{|Y(w,m)|^2}{|W(w,m)|^2}\right)^2} \; (dB) \qquad (10)$$

where m indicates the frame index, M is the total numbers of frames, and $Y(w,m)$ and $W(w,m)$ are the spectra of m-th frame in the original and watermarked signals. The resulting LSD values are given in Table 2. The LSD≤1.0 dB is considered to be good for inaudibility of the watermarks and a lower value indicates a less distortion [2]. As shown in the table, the values are very close to the criterion (less than 0.01 for all files) and ensure the inaudibility of the watermarked speech signal.

Table 2. LSD results for inaudibility

Signal (8 kHz, 16-bit)	Length (sec)	Total Samples	Watermark bits	LSD (dB)
news 1 (f,B)	7	60416	4LSB	0.01
news 2 (f,B)	8	64512	4LSB	0.02
news 3 (m,B)	9	75904	4LSB	0.02
news 4 (f,E)	10	86016	4LSB	0.01
news 5 (m,E)	11	92032	4LSB	0.01

5.2. Performance Evaluation for Fragility

Fragility means that watermarks are sensitive to tampering and easy to be destroyed once tampering has been made. In general, tampering are performed based on the motivation of the attackers. The following tampering types are applied to test the fragility of the proposed watermarking method.

5.2.1. Tampering Types

1. Zeroing: Replacing the samples of a speech signal by zeros introduces silence in the speech signal (similar to the lack of audible sounds or presence of sounds with very low intensity).
2. Adding Noise: Sound heard when there's no signal on TV or radio and sound of fun.
3. Reverberation: Reverberation is the persistence of sound after a sound is produced when a person sings, talks, or plays an instrument acoustically in a hall with sound reflective surfaces.
4. Time Scaling: The duration and tempo of a speech signal is modified without affecting its pitch.
5. Filtering: Filtering (high-pass/low-pass) is regarded as removing specific frequency information of the speech.
6. Concatenation: Concatenating with un-watermarked speech signal can be considered as content replacement. e.g., a word replacement from "YES" to "NO".
7. Compression: A speech signal is compressed in order to reduce file size for efficient transmission.

5.2.2. Bit Detection Rate (BDR)

Fragility can be indicated by bit detection rate, i.e., the percentage of the ratio between correctly extracted watermark bits in the transmission of digital information and total amount of embedded watermark bits. The BDR can be calculated with Eq. (11).

$$\text{BDR} = \frac{\text{M} - \sum_{m=0}^{M-1} \text{s(m)} \oplus \hat{\text{s}}\text{(m)}}{\text{M}} * 100\ (\%) \qquad (11)$$

where $s(m)$ is the embedded watermark, $\hat{s}(m)$ is the extracted watermark, and M is the total length of $s(m)$. In order to be fragile against tampering, a lower BDR indicates strong confirmation of tampering [2].

Table 3 shows the average BDR results for the following attack parameters on five tested signals.

(zeroing) 21% of the watermarked signals are replaced by zeros; (adding noise) additive white Gaussian noise is added to the watermarked signals by keeping the SNR=20dB; (reverberation) an echo is added to the watermarked signals with delay of 500 ms; (time scaling) the speed of the watermarked signals are turned up and down twice; (filtering) the watermarked signals are low-pass and high-pass filtered by Butterworth filter with cutoff frequency of 2 kHz; (concatenation) 21% of the watermarked signals are replaced with samples from the un-watermarked signals; and (compression) G.711 codec is used to compress the watermarked signals by reducing the bit rate from 128 kbps to 64 kbps. The SNR values in Table 3 show how severe the tampering attacks are.

From Table 3, it can be seen that the average BDR values are 100% for no tampering, around 91% for zeroing and concatenation attacks, and around 49-53% for other attacks. The more severe the attacks, the lower the BDR values. These obtained results suggest that the proposed method is fragile against tampering and the destroyed watermarks can provide an evidence that a signal has been tampered.

Table 3. Average BDR results for fragility

No	Tampering type	BDR (%)	SNR (dB)
1	No tampering	100	Inf
2	Add zeroing	91.02	5.3
3	Add white Gaussian noise	50.02	21.6
4	Reverberation	52.86	-0.27
5	Concatenation	91.05	56.04
6	Low-pass filtering	50.03	-1.72
7	High-pass filtering	50.03	-3.66
8	Speed up	49.93	-1.07
9	Speed down	49.98	0.21
10	Compression	50.1	-19.12

Fig. 5 shows how the severity of tampering attacks affects the BDR. For zeroing, concatenation, and reverberation attacks, the effect of tampering depends on the amount of tampering.

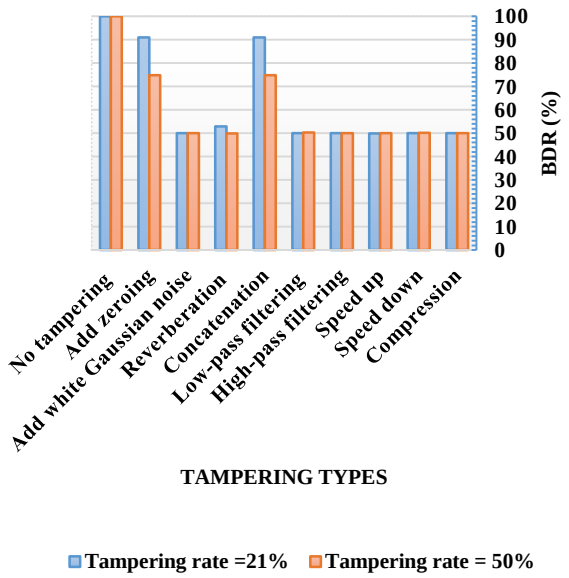

Figure 5. The amount of tampering on BDR

For example, 50% replacement by zeros will destroy the signal more than 21% replacement by zeros, and leads to lower BDR. However, for filtering, time scaling, noise addition, and compression attacks, tampering affects equally on all samples. Thus, no matter how severe the attack is, the BDR results will be absolutely the same.

5.3 Tampering Localization

For clear illustration, tampering regions can be shown on a graph with tampered and not-tampered frames denoted by one and zero, respectively.

Fig. 6 (a) shows the waveform of a 7-sec long un-watermarked speech signal. Fig. 6 (b) shows the waveform of the watermarked speech signal. It can be observed that the waveform of the watermarked speech signal looks similar to its respective original speech signal, and the differences are not perceivable. Therefore, they do not attract the attention of attackers. Fig. 6 (c) shows the waveform of the tampered signal by a malicious attacker. In this attack, 21% of the watermarked speech signal is replaced by silence (zero). Fig. 6(d) shows the results of the hash bit examination procedure in determining the tampered frames, in which 0's shows the reserved frames and 1's shows the tampered identified frames. In this way, the tampered regions can be easily localized.

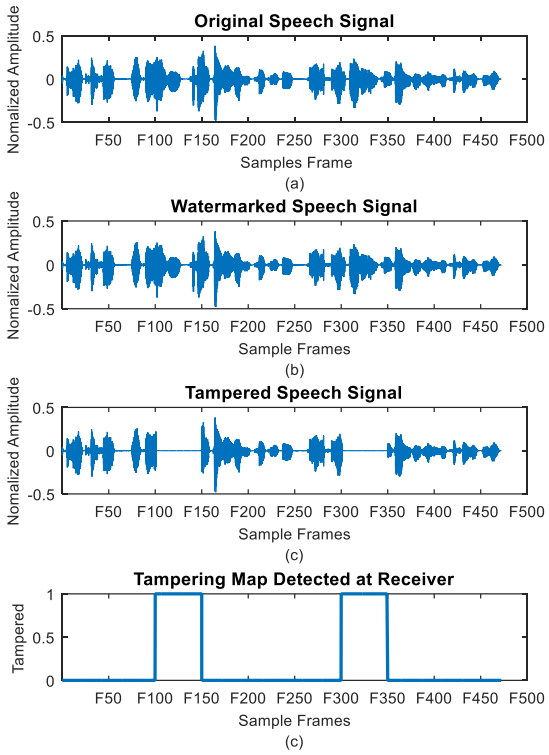

Figure 6. (a) Original speech signal; (b) Watermarked speech signal; (c) Tampered speech signal; and (d) Tampering localization

6. Conclusion

In this paper, an efficient tampering detection and localization method for speech signals was proposed. A self-embedding speech signal was produced by inserting a watermark that consists of a representation of the original signal into itself to show fragility against tampering. Experimental results showed that the proposed method maintained high quality of the watermarked speech signals and provided fragility against tampering. Moreover, it could detect tampering position precisely.

References

[1] S. Sarreshtedari, M. A. Akhaee, and A. Abbasfar, "A watermarking method for digital speech self-recovery," IEEE/ACM Trans. Audio, Speech, Lang. Process., pp. 1917-1925, vol. 23, Nov 2015.

[2] S. Wang, N. S. Kim, and M. Unoki, "Formant enhancement based speech watermarking for tampering detection," School of Information Science, JAIST, vol. 6, Sep 2014.

[3] H. Hering, Hagmuller M Hagmuller M and G. Kubin, "Safety and security increase for air traffic management through unnoticeable watermark aircraft identification tag transmitted with the VHF voice communication," Digital Avionics

Syst. Conf. (DASC'03), pp. 4.E.2–41-10, vol. 1, Oct 2003.

[4] M. Parvaix, L. Girin, and J. Brossier, "A watermarking-based method for informed source separation of audio signals with a single sensor," IEEE Trans. Audio, Speech, Lang. Process., pp. 1464–1475, vol. 18, no. 6, Aug 2010.

[5] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," IEEE Trans. Audio, Speech, Lang. Process., pp. 1721–1733, vol. 19, no. 6, Aug 2011.

[6] Y. Nakashima, R. Tachibana, and N. Babaguchi, "Watermarked movie soundtrack finds the position of the camcorder in a theater," IEEE Trans. Multimedia, pp. 443–454, vol. 11, no. 3, Apr 2009.

[7] C.-P. Wu and C.-C. Kuo, "Fragile speech watermarking based on exponential scale quantization for tamper detection," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. IV-3305–IV-3308, vol. 4, May 2002.

[8] M. Unoki and R. Miyauchi, "Detection of tampering in speech signals with inaudible watermarking technique," in Proc. 8th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process. (IIH-MSP), pp. 118–121, Jual 2012.

[9] https://en.wikipedia.org/wiki/Secure_Hash_Algorithms

[10] H. Yi and C. L. Philipos, "Evaluation of objective measures for speech enhancement," Interspeech2006, pp. 1447-1450, Pittsburgh, Pennsylvania, Sep 2006.