WATERMARKING OF SPEECH SIGNALS BASED ON FORMANT ENHANCEMENT

Shengbei Wang and Masashi Unoki

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
{wangshengbei, unoki}@jaist.ac.jp

ABSTRACT

This paper proposes a speech watermarking method based on formant enhancement. The line spectral frequencies (LSFs) which can stably represent the formants were firstly derived from the speech signal by linear prediction (LP) analysis. A pair of LSFs were then symmetrically controlled to enhance formants for watermark embedding. Two kinds of objective experiments regarding inaudibility and robustness were carried out to evaluate the proposed method in comparison with other typical methods. The results indicated that the proposed method could not only satisfy inaudibility but also provide good robustness against different speech codecs and general processing, while the other methods encountered problems.

Index Terms— Speech watermarking, formant enhancement, line spectral frequencies, inaudibility, robustness

1. INTRODUCTION

Modern technologies have enabled audio and speech to be easily reduplicated and edited at high fidelity. Illegal use of these technologies, however, results in serious problems in copyright protection and unauthorized tampering. Watermarking can identify copyright and detect tampering by embedding information such as copyright notice or serial number (also referred as watermarks) into the host signal. A desired watermarking should be inaudible to human perception and robust against general signal processing. For copyright protection of audio signals, robustness is controlled as the top priority. For speech signals, more attention has been paid to tampering detection to confirm the originality of speech signals. In this case, watermarking should be fragile to identify where tampering has occurred. Nonetheless, fragile watermarking should firstly be robust against general processing to confirm that the failed detection of watermarks is only caused by tampering [1]. Thus this work focuses on the basic inaudible and robust watermarking for speech signals.

Many watermarking methods have been proposed in recent years. The least significant bit-replacement (LSB) [2] and direct spread spectrum (DSS) [3] methods have separately exhibited good performance in inaudibility and robustness. Watermarking based on modifying the fundamental frequency [4] has been applied for authentication, although its robustness against malicious attacks has not been designed. A speech watermarking [5] has taken advantage that human is insensitive to absolute phase to realize inaudibility. Unoki and Hamada have realized a watermarking [6, 7] based on cochlear delay (CD). Swanson *et al.* [8] have suggested to embed watermarks by considering the perceptual masking. However, since inaudibility and robustness usually conflict with each other, it is difficult for most methods to satisfy them simultaneously.

We have previously proposed a speech watermarking [9] based on modifying line spectral frequencies (LSFs) with quantization index modulation (QIM) [10]. A problem inherent to this method is that the modified values of LSFs are just based on their initial values and the quantization step. These unintentional modifications to LSFs randomly disrupt the formant structure of speech signal and distort the sound quality. Additionally, it is difficult for this method to achieve a trade-off between inaudibility and robustness due to the nature of QIM. Nonetheless, previous work shows how formant can be produced and controlled. From related studies in the field of speech synthesis, where formants can be enhanced to improve the sound quality of synthesized speech [11–13], we found if watermarks could be embedded through formant enhancement, it would be more reasonable to achieve both improved inaudibility and robustness. Thus we propose a speech watermarking method based on formant enhancement in this paper.

2. CONCEPT UNDERLYING WATERMARKING

In speech synthesis, formants are enhanced with complicated methods so that the dynamics between formant peaks and spectral valleys can be increased. As to inherited formant enhancement for watermarking, in this paper, we investigate a direct but effective formant enhancement method. The following subsections talk about how formants can be estimated, enhanced, and applied for watermarking.

2.1. Formant estimation and enhancement

Formant estimation: The linear prediction (LP) analysis can predict current signal with its past samples and LP coefficients. Based on the source-filter model, the set of LP coefficients is an all-pole model that can provide accurate estimate of formants. In practice, LP coefficients are usually substituted with LSFs, reflection coefficients (RCs), etc. to ensure the stability of predictor. Among these, LSFs have several excellent properties: (*i*) they are less sensitive to noise; (*ii*) the influences caused by deviation of LSFs can be limited to the local spectral, which suggests that if LSFs are used to enhance formant for watermark embedding, the distortion introduced by watermarks in both spectral and sound quality can be minimized; (*iii*) LSFs are universal features in different speech codecs, which indicates that watermarks in LSFs are possible to survive from coding/decoding to provide the robustness against speech codecs. Hence we employ LSFs to enhance formant. The LSFs converted from LP coefficients satisfy the ordering property from 0 to π as follows:

$$0 < \phi_1 < \phi_2 < \phi_3 < \cdots < \phi_p < \pi, \tag{1}$$

where p indicates the LP order, $\phi_i, 1 \leq i \leq p$, are the LSFs.

This work was supported by a Grant-in-Aid for Scientific Research (B) (No. 23300070), an A3 foresight program made available by the Japan Society for the Promotion of Science, the telecommunication advancement foundation, and funding by China Scholarship Council.

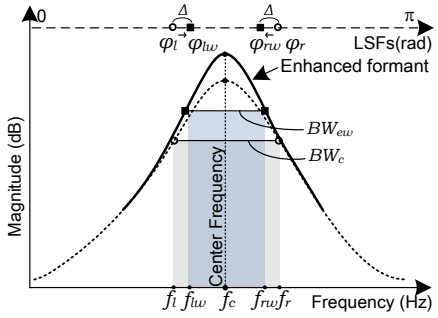

Fig. 1. Formant enhancement by controlling LSFs.

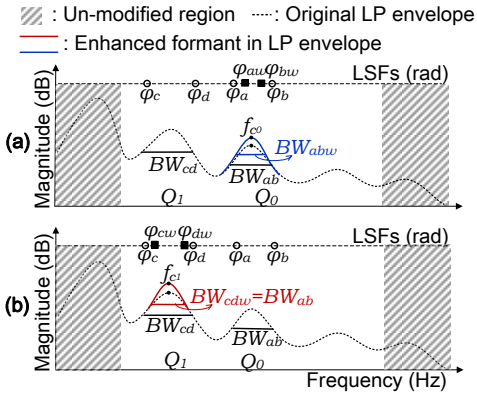

Fig. 2. Concept of watermark embedding based on formant enhancement: (a) embedding '0' and (b) embedding '1'.

Formant enhancement: The relationship between formant and LSFs is that each formant can be produced by two adjacent LSFs, the closer two LSFs are, the sharper the formant is. For a fixed formant, sharpness can be mathematically measured by tuning level, that is the Q-value defined in (2), where f is the center frequency of formant, BW is the bandwidth. For different applications, BW has different definitions. In our method, BW is defined as the bandwidth between two LSFs after transferring them into frequency domain.

$$Q = \frac{f}{BW} \tag{2}$$

For a fixed formant, when its Q-value is increased, formant will become sharper. Therefore, we enhance a formant (increase Q-value) by directly closing up two LSFs to produce a narrower bandwidth while keeping its center frequency unchanged (center frequency is maintained to preserve sound quality). As seen in Fig. 1, the formant (dotted curve) that is produced by two LSFs, ϕ_l and ϕ_r, has a tuning level Q_c defined in (3), where f_c is the center frequency, BW_c is the bandwidth between f_l and f_r that converted from ϕ_l and ϕ_r with (4), where F_s is the sampling frequency of signal. According to (3), the given formant can be enhanced by reducing BW_c without shifting f_c. Hence, ϕ_l and ϕ_r are symmetrically shifted to ϕ_{lw} and ϕ_{rw} with the same modification degree Δ in (5), after which f_{lw} and f_{rw} in (6) can produce a narrower bandwidth BW_{ew} in (7). The tuning level of enhanced formant is increased to Q_{ew}.

$$Q_c = \frac{f_c}{BW_c} = \frac{f_c}{f_r - f_l} \tag{3}$$

$$f_r = \frac{\phi_r}{2\pi} \times F_s \text{ and } f_l = \frac{\phi_l}{2\pi} \times F_s \tag{4}$$

$$\phi_{lw} = \phi_l + \Delta \text{ and } \phi_{rw} = \phi_r - \Delta, \quad 0 < \Delta < (\phi_r - \phi_l)/2 \tag{5}$$

$$f_{rw} = \frac{\phi_{rw}}{2\pi} \times F_s \text{ and } f_{lw} = \frac{\phi_{lw}}{2\pi} \times F_s \tag{6}$$

$$Q_{ew} = \frac{f_c}{BW_{ew}} = \frac{f_c}{f_{rw} - f_{lw}} \tag{7}$$

2.2. Watermarking based on formant enhancement

Preliminary analysis: Watermarks can be embedded into the host signal when LSFs are shifted for formant enhancement. Before embedding, we should clarify several points to make our method effective. (*i*) Several formants can be extracted from each speech frame. As we have surveyed, the distortion caused by enhancing formants in the low and high frequencies can be easily perceived by human, we thus leave the first formant and last formant un-modified. Only one formant in the middle region will be enhanced for watermark embedding. (*ii*) Since formant structures vary widely with different speech frames, it is preferable to enhance formants according to their original tuning characteristics to achieve inaudibility. However, such self-adaptive enhancement results in a serious problem for blind watermark detection because it is so difficult to detect watermarks without any prior knowledge of how the formant has been enhanced. As we have considered, one solution for blind detection is enhancing one formant and thus to establish an internal relationship between the enhanced formant and another formants in current frame. In detection process, two formants can make a cross-reference with each other, watermarks can be extracted by identifying the relationship.

Embedding concept: In our method, each speech frame will be embedded with one bit, '0' or '1'. For each frame, firstly, we use LP analysis to estimate the formants. Secondly, we check the bandwidths (indicated by two LSFs) of each formant in the middle region. The smaller the bandwidth is, the sharper the formant is. Thirdly, we separately calculate and label the tuning level of each formant as Q_0, Q_1, $\cdots$ with increased bandwidth. As seen in Figs. 2(a) and 2(b), the sharpest formant (labelled as Q_0, produced by ϕ_a and ϕ_b) has the smallest bandwidth BW_{ab}, and the second sharpest formant (labelled as Q_1, produced by ϕ_c and ϕ_d) has the second smallest bandwidth BW_{cd}. That is $BW_{cd} > BW_{ab}$. Relationships for '0' and '1' will be established between two sharpest formants, i.e., the Q_0, Q_1 labelled formants. Embedding rule will be selected from the following two cases according to the watermark '0' or '1'.

Rule of embedding '0': If '0' will be embedded, as seen in Fig. 2(a), we will enhance the sharpest formant by Ω_{e0} (Ω_{e0}>1) times by reducing BW_{ab}, where Ω_{e0} is called as enhancing factor. According to (8), BW_{ab} will be reduced to its $1/\Omega_{e0}$, i.e. $BW_{abw} = BW_{ab}/\Omega_{e0}$. To achieve this, original LSFs ϕ_a and ϕ_b in (9) will be symmetrically shifted to ϕ_{aw} and ϕ_{bw} with respect to the center frequency f_{c0}, where the modification degree Δ_{e0} is calculated by ϕ_a, ϕ_b, and Ω_{e0} with (10). Since BW_{cd} is originally larger than BW_{ab}, after embedding '0', a updated relationship, i.e., $BW_{cd} > BW_{abw} \times \Omega_{e0}$ has been established in current frame.

$$Q_0 \times \Omega_{e0} = \frac{f_{c0}}{BW_{ab}} \times \Omega_{e0} = \frac{f_{c0}}{BW_{ab}/\Omega_{e0}} = \frac{f_{c0}}{BW_{abw}} \tag{8}$$

$$\phi_{aw} = \phi_a + \Delta_{e0} \text{ and } \phi_{bw} = \phi_b - \Delta_{e0} \tag{9}$$

$$\Delta_{e0} = \frac{1}{2}[(\phi_b - \phi_a) \times (1 - \frac{1}{\Omega_{e0}})] \tag{10}$$

Rule of embedding '1': If '1' will be embedded, as seen in Fig. 2(b), we will enhance the second sharpest formant with (11) by using the enhancing factor $\Omega_{e1} = BW_{cd}/BW_{ab}$. With this factor, BW_{cd} can be reduced to the same as BW_{ab}. This is achieved by shifting ϕ_c and ϕ_d to ϕ_{cw} and ϕ_{dw} with (12), where Δ_{e1} is calculated by ϕ_c, ϕ_d and Ω_{e1} with (13). Therefore, '1' can be embedded by establishing the relationship that the second sharpest formant has the same bandwidth as the sharpest formant.

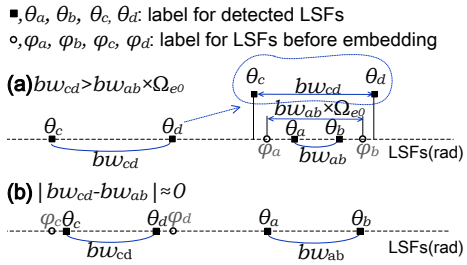

Fig. 3. Concept of watermark detection based on formant enhancement: (a) '0' is detected and (b) '1' is detected.

$$Q_1 \times \Omega_{e1} = \frac{f_{c1}}{BW_{cd}} \times \Omega_{e1} = \frac{f_{c1}}{BW_{cd}/\Omega_{e1}} = \frac{f_{c1}}{BW_{ab}} = \frac{f_{c1}}{BW_{cdw}} \quad (11)$$

$$\phi_{cw} = \phi_c + \Delta_{e1} \text{ and } \phi_{dw} = \phi_d - \Delta_{e1} \quad (12)$$

$$\Delta_{e1} = \frac{1}{2}[(\phi_d - \phi_c) \times (1 - \frac{1}{\Omega_{e1}})] \quad (13)$$

In summary, different watermarks are embedded with formant enhancement to establish different bandwidth relationships between the sharpest and the second sharpest formants. When '0' is embedded, bandwidth difference between two formants is increased since the smaller bandwidth is reduced; while for '1', bandwidth difference is reduced to 0. This opposite mechanism enables a blind detection of watermarks.

Detection concept: As we can see, bandwidth relationships always exist in the sharpest and the second sharpest formants no matter for embedding '0' or '1'. Therefore, for a received frame in detection process, as seen in Fig. 3, we extract two smallest bandwidths, named as bw_{ab} (smallest, produced by θ_a and θ_b) and bw_{cd} (the second smallest, produced by θ_c and θ_d). According to Fig. 3(a), if '0' has been embedded, we have $bw_{cd} > bw_{ab} \times \Omega_{e0}$, an equivalent representation is given in (14); if '1' has been embedded, bw_{cd} in Fig. 3(b) should be equal to bw_{ab} (note that the LSFs before embedding, ϕ_a, ϕ_b, ϕ_c, ϕ_d, are not available in the detection, they are just illustrated for understanding). However, since LP analysis calculates LP coefficients (or LSFs) with the criterion that the mean-squared error is always minimized, the LP coefficients (or LSFs) that are derived from watermarked frame are not exactly the same as those after embedding process even there is no modifications. Therefore, we set a threshold as expressed in (15) to discriminate two cases of embedding '0' or '1', and enable the method to be error-tolerant.

$$bw_{cd} - bw_{ab} > bw_{ab} \times (\Omega_{e0} - 1) \quad (14)$$

$$\hat{s}(m) = \begin{cases} 0, & bw_{cd} - bw_{ab} > bw_{ab} \times (\Omega_{e0} - 1)/2 \\ 1, & \text{otherwise} \end{cases} \quad (15)$$

3. SPEECH WATERMARKING SCHEME

The proposed watermarking method is based on the speech analysis/synthesis technique. LP analyses the speech signal by extracting LP coefficients and residue signal. Watermarks are embedded into LSFs by enhancing one formant. The watermarked speech can be resynthesized by the residue signal and the LP coefficients that converted from the modified LSFs and the other unmodified LSFs.

Figure 4(a) has a block diagram of embedding process. Watermark signal $s(m)$ is embedded into the original signal $x(n)$ as follows. First, $x(n)$ is segmented into non-overlapping frames. For a single frame, LP analysis is applied to obtain LP coefficients and LP residue. Then LP coefficients are converted to LSFs. The current watermark is embedded according to the concept that introduced

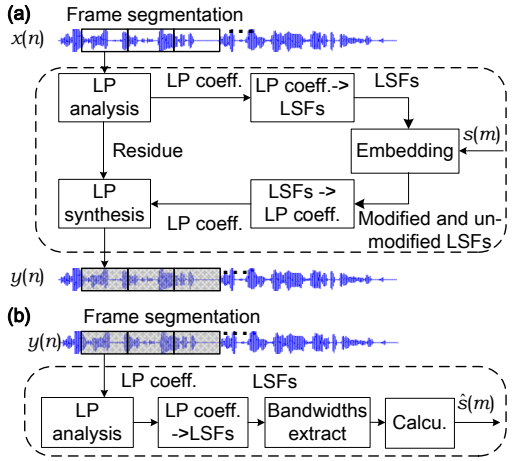

Fig. 4. Block diagram of proposed watermarking method: (a) embedding process and (b) detection process.

in Section 2.2, after which two modified LSFs are generated. All LSFs including the modified LSFs and the other unmodified LSFs are converted back to LP coefficients. The current frame is then synthesized by the residue and the LP coefficients. The watermarked signal, $y(n)$, is finally reconstructed with all watermarked frames using non-overlapping and adding function. Detection process is shown in Fig. 4(b). We apply the same procedures as those in the embedding process to watermarked signal $y(n)$ to obtain the LSFs. The watermark in current frame is detected with the method in Section 2.2. Each frame can be extracted with one bit, all extracted bits can construct the whole watermark signal, $\hat{s}(m)$.

4. EVALUATIONS

Several experiments were conducted to evaluate the proposed method. All twelve speech stimuli (male/female Japanese sentences) in the ATR speech database (B set) [14], which were clipped into 8.1-sec durations, sampled at 20 kHz, and quantized with 16 bits were used as the database. For extended use of watermarking as information hiding method, we evaluated the performance of the proposed method as a function of the bit rate. Since the watermarked signals in the proposed method were reconstructed based on speech analysis/synthesis techniques, the frame size for embedding was fixed at 25 ms (40 frames in 1.0 sec) to attain better sound quality. To construct bit rate, all frames within 1.0 sec speech segment were separately divided into 4, 8, 20, to 40 groups. All frames within the same group were embedded with the same watermark and then the watermark was detected with a majority decision. Thus, the bit rates for proposed method were 4, 8, 20, and 40 bps. The LP order for speech analysis/synthesis was adopt as 10 to balance inaudibility and robustness. As seen in Fig. 3, bigger Ω_{e0} is beneficial to discriminate '0' or '1' for robustness but worse for inaudibility. We thus chose $\Omega_{e0} = 2$ to balance the inaudibility and robustness. Ω_{e1} was fixed according to bandwidth characteristics of each frame. The embedded watermarks was "JAIST-IS-Acoustic".

Evaluations were also done for three typical methods, i.e., LSB [2], DSS [3], and CD [7], to enable a comparative study. A quick review of these methods is as follows: LSB replaces the least significant bits with watermarks at the quantization level so that the replacement does not cause severe distortion; DSS spreads watermarks over many (possibly all) frequency bands so that the watermarks cannot be easily destroyed; CD embeds watermarks by enhancing

the phase information of host signal with respect to two kinds of cochlear delay. The bit rates for LSB, DSS, and CD were 4, 8, 16, 32, and 64 bps according to their original implementations. All evaluation results were calculated on the average of twelve stimuli.

4.1. Evaluations of inaudibility

Two tests of log-spectrum distortion (LSD) [15] and perceptual evaluation of speech quality (PESQ) [16] were applied to check inaudibility. LSD in decibel (dB) was the distance measure between two spectra of the original and watermarked signals. 1.0 dB was chosen as the criterion, and a lower value indicated less distortion. PESQ in the objective difference grades (ODGs) that cover from -0.5 (very annoying) to 4.5 (imperceptible) was used to evaluate subjective quality. We set 3.0 (slightly annoying) as the criterion, and a higher value indicated better sound quality. Evaluation results of the proposed method, LSB, DSS, and CD are plotted in Fig. 5. As we can see, LSB had the best performance among all the four methods. CD could satisfy inaudibility when the bit rate was no more than 16 bps. DSS could not satisfy the criteria for either LSD or PESQ. The proposed method could satisfy criteria for both LSD and PESQ, which indicated it could satisfy the inaudibility requirement.

4.2. Evaluations of robustness

Robustness indicated whether watermarks could be detected from the watermarked signals although they have been processed with other signal processing. We chose bit detection rate of 90% as the criterion. Higher bit detection rate indicated stronger robustness.

Robustness against speech codecs: Speech codecs are usually applied to speech for transmission. Therefore, robustness against speech codecs is very important to guarantee the effectiveness of watermarks. We chose three typical speech codecs of G.711 (pulse code modulation (PCM)), G.726 (adaptive differential PCM (ADPCM)), and G.729 (Code-excited linear prediction (CELP)) to evaluate the robustness. The results after normal detection (no codec) and other three codecs are plotted in Fig. 6. As we can see, LSB was not robust against any speech codec except for normal detection; CD was not robust against G.726 and G.729; DSS was robust against G.711 and G.726, while for G.729 in Fig. 6(d), its bit detection rate drastically deteriorated. The proposed method had good robustness against normal detection, G.711, and G.726. It could also survived from G.729 since it provided a satisfactory bit detection rate of 90% at 4 bps. These results indicated that the proposed method was very robust against speech codecs.

Robustness against general processing: The proposed method were first evaluated with four processing: (a) re-sampling at 24 kHz and (b) at 12 kHz, and (c) re-quantization with 24 bits and (d) 8 bits. Figure 7 plots all the results. DSS obviously performed the best. The proposed method and CD provided better performance than LSB. For re-quantization with 8 bits, the proposed method was not so good because quantization at lower rate introduced distortions to watermarked signal and thus destroyed the bandwidth relationship for watermark detection. We then evaluated the proposed method with realistic speech processing, such as (a) signal amplifying by 2.0, speech analysis/synthesis by (b) short-time Fourier transform (STFT), and (c) gammatone filterbank (GTFB). We also took a series of standard processing that recommended by information hiding and its criteria (IHC) committee [17] as reference, although these processing were used to evaluate the robustness of audio watermarking. These were (d) Gaussian noise addition with an overall average SNR of 36 dB and (e) a single 100-ms echo addition of -6 dB (slight processing done by (d) and (e) can be viewed as general processing to evaluate robustness). The bit detection results at 4 bps are listed in Table. 1. It is clear that DSS had the best performance. The proposed method was also robust against all processing.

4.3. Discussion and future work

We give a performance analysis of all evaluated methods as follows. LSB method embeds watermarks in the least significant bits so that distortion to original signal is negligible and thus makes LSB to perfectly satisfy inaudibility. However, watermarks in the least significant bits can be easily reset by any operations that related to amplitude modifications or lossy processing, which makes the LSB method fragile. DSS is relatively robust (except for G.729) since watermarks are spread over a wide frequency range, only all possible frequencies are destroyed with considerable strength can eliminate the watermarks. Therefore DSS exhibits strong robustness for most processing. However, on the other hand, watermarks in a wide frequencies make them perceptually significant. Watermarks in CD are embedded as phase information by modelling the cochlear delay. According to the characteristics of cochlear delay, detection of different watermarks '0' and '1' strongly depends on the cue in low frequency phase. Correspondingly, once phase information in low frequency is destroyed or erased by other processing, such as GTFB and G.729 codec, watermarks cannot be detected.

The proposed method can basically satisfy both inaudibility and robustness compared with other methods (LSB: not robust but inaudible; DSS: robust but not inaudible; CD: conditionally satisfy inaudible and robust). In the proposed method, watermarks are embedded by enhancing formant without shifting center frequency, thus inaudibility can be achieved. Besides, watermark detection by identifying bandwidth relationship is able to tolerate slight changes of frequency components that are caused by general processing. Moreover, since each frame has its own frequency characteristics, the enhanced formant (the sharpest formant or the second sharpest formant) is possible to exist in any frequency range. When small proportion of frequency components that do not contain watermark information are changed, watermarks are able to survive. It also important to note that the proposed method is different from QIM-based watermarking, since most QIM-based watermarking methods modify the embedding parameter without physical meaning, while in our method, the modification to LSFs is motivated by formant enhancement.

While there are some remaining issues need be considered in the current work. As the proposed method is a frame-based watermarking, an automatic scheme for frame synchronization should be implemented in the future. We will also try to develop the proposed method for tampering detection in the future.

5. CONCLUSION

This paper proposed a novel speech watermarking method. The concept of formant enhancement was introduced to watermarking for the first time after we considered its superiority in improving the sound quality for synthesized speech. We investigated the principles of how the formants could be produced, controlled, and then enhanced to make the method effective. Watermarks, therefore, were embedded as formant enhancement with a straight-forward way by symmetrically controlling two LSFs, which made the proposed method to be implemented with less computation complexity. Several evaluations were carried out on the proposed method. The results from evaluations suggested that the proposed method could not only satisfy inaudibility but also provide good robustness, especially the robustness against speech codecs.

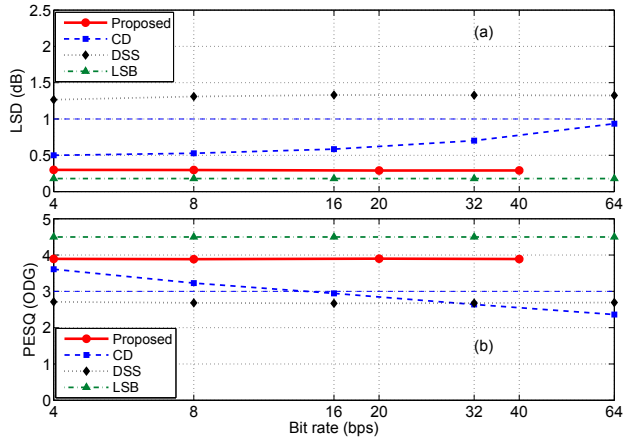

Fig. 5. Evaluation of inaudibility: (a) LSD and (b) PESQ.

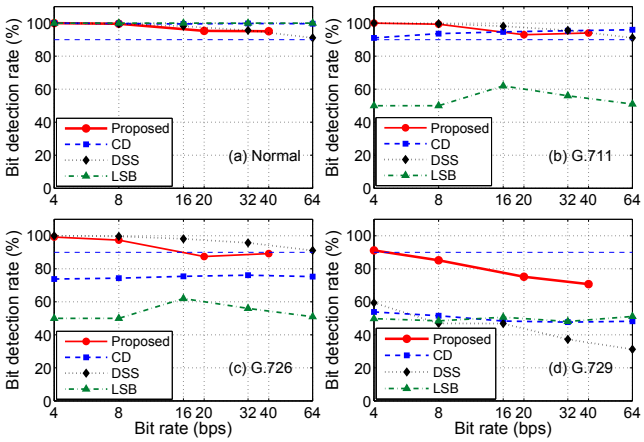

Fig. 6. Evaluation of robustness against speech codecs: (a) normal detection, (b) G.711, (c) G.726, and (d) G.729.

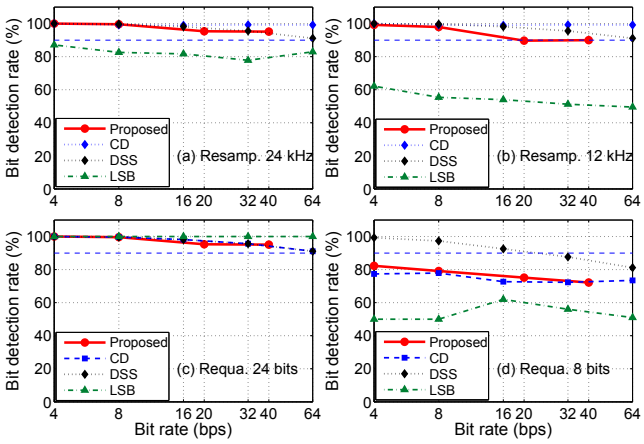

Fig. 7. Evaluation of robustness against re-sampling at (a) 24 kHz and (b) 12 kHz and re-quantization with (c) 24 bits and (d) 8 bits.

Table 1. Evaluations of robustness against realistic processing

Processing	CD	DSS	LSB	**Proposed**
Amplifying	96.43	100.00	50.00	**99.22**
STFT	96.43	100.00	100.00	**99.22**
GTFB	66.07	100.00	53.42	**98.96**
Noise addition	96.43	100.00	100.00	**100.00**
Echo addition	64.29	100.00	42.83	**92.19**

REFERENCES

[1] C. Wu and C. Jay Kuo, "Fragile speech watermarking based on exponential scale quantization for tamper detection," in *Proc. ICASSP*, vol. IV, pp. 3305–3308, 2002.

[2] P. Bassia and I. P. Pitas, "Robust audio watermarking in the time domain," in *Proc. EUSIPCO*, pp. 25-28, 1998.

[3] L. Boney, H. H. Tewfik, and K. H. Hamdy, "Digital watermarks for audio signals," in *Proc. ICMCS*, pp. 473-480, 1996.

[4] M. Celik, G. Sharma, and A. M. Tekalp, "Pitch and duration modification for speech watermarking," in *Proc. ICASSP*, vol. II, pp. 17–20, 2005.

[5] M. Narimannejad and A. S. Mohammad, "Watermarking of speech signal through phase quantization of sinusoidal model," in *Proc. ICEE*, pp. 1–4, 2011.

[6] M. Unoki and D. Hamada, "Method of digital-audio watermarking based on cochlear delay characteristics," in *J. Inn. Com. Inf., and Cont.*, vol. 6, no.(3(B)), pp. 1325–1346, 2010.

[7] M. Unoki and R. Miyauchi, "Reversible watermarking for digital audio based on cochlear delay characteristics," in *Proc. IIHMSP*, pp. 314–317, 2011.

[8] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," in *Signal Processing*, vol. 66, no. 3, pp. 337-355, 1998.

[9] S. Wang and M. Unoki, "Watermarking method for speech signals based on modifications to LSFs," in *Proc. IIHMSP*, pp. 283–286, 2013.

[10] B. Chen and G. W. Wornel, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," in *IEEE Trans. Information Theory*, vol. 47, no. 4, pp. 1423–1443, 2001.

[11] T. Raitio, A. Suni, H. Pulakka1, M. Vainio, and P. Alku, "Comparison of formant enhancement methods for HMM-based speech synthesis," in *Proc. ISCA Speech Synthesis Workshop*, 2010.

[12] HTS, "HMM-based speech synthesis system," *http://hts.sp.nitech.ac.jp*, 2009.

[13] Recommendation ITU-T P.800, "Methods for subjective determination of transmission quality," in *International Telecommunication Union*, 1996.

[14] K. Takeda et al, "Speech database user's manual," in *ATR Technical Report TR-I-0028*, 2010.

[15] A. Gray, Jr., and J. Markel, "Distance measures for speech processing," in *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[16] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," in *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[17] Information hiding and its criteria for evaluation. *http://www.ieice.org/iss/emm/ihc/en/index.php*.