PAPER Special Section on Enriched Multimedia

Speech Watermarking Method Based on Formant Tuning

Shengbei WANG[†a)], ***Student Member and*** **Masashi UNOKI**[†], ***Senior Member***

SUMMARY This paper proposes a speech watermarking method based on the concept of formant tuning. The characteristic that formant tuning can improve the sound quality of synthesized speech was employed to achieve inaudibility for watermarking. In the proposed method, formants were firstly extracted with linear prediction (LP) analysis and then embedded with watermarks by symmetrically controlling a pair of line spectral frequencies (LSFs) as formant tuning. We evaluated the proposed method by two kinds of experiments regarding inaudibility and robustness compared with other methods. Inaudibility was evaluated with objective and subjective tests and robustness was evaluated with speech codecs and speech processing. The results revealed that the proposed method could satisfy both inaudibility and robustness that required for speech watermarking.

key words: *speech watermarking, formant tuning, line spectral frequencies, inaudibility, robustness*

1. Introduction

Due to the illegal use of digital techniques, the problems of unauthorized tampering in speech signals have arisen. For digital forensics, the originality of recorded speech that used in the court should be strictly confirmed. Speech watermarking [1], [2] can detect tampering as well as check the originality of speech by embedding information (referred as watermarks) into the host signal. The embedded information should be inaudible to human auditory system and fail to be detected once a slight tampering has been made to the host signal. This kind of watermarking is referred as fragile watermarking. Nonetheless, to guarantee a reliable identification of tampering, fragile watermarking should first and foremost be robust against speech processing to confirm that the failed detection of watermarks could only be caused by tampering. Therefore, robustness is extremely important for speech watermarking. As a fundamental work, this paper focuses on the inaudible and robust speech watermarking.

In literature, many speech watermarking methods toward inaudibility and robustness have been proposed. These methods can be categorized according to the implemented domain. The time-domain methods, such as the least significant bit-replacement (LSB) [3] method and the echo hiding-based methods [4], [5], however, were prone to be not robust. Methods in [6]–[8] and the spread spectrum-based methods [9]–[12] tried to achieve stronger robustness in the transform-domain while inaudibility could not be always satisfied. Since the human auditory system (HAS) is particularly sensitive, several previous studies [13]–[15] exploited the properties of HAS and applied such knowledge for watermarking. In these methods, watermarks were embedded into the perceptually inaudible components while leaving the sensitive components intact to realize inaudibility.

Since the inaudibility and robustness conflict with each other, watermarking that can satisfy both inaudibility and robustness are difficult to realize. We previously proposed a speech watermarking [16] based on modifying the line spectral frequencies (LSFs) [17] with quantization index modulation (QIM). However, QIM-based modifications to LSFs could easily disrupt the formant structure of the host signal and distort the sound quality. Moreover, due to the nature of QIM, robustness of the previous method was improved at the expense of degraded inaudibility which made it difficult to get a trade-off between inaudibility and robustness. According to related studies in formant enhancement/tuning-based speech synthesis where formants can be tuned to improve the sound quality of synthesized speech, we found if watermarks could be embedded through formant tuning, it would be more reasonable to achieve both improved inaudibility and robustness. In this paper, we propose a speech watermarking method based on formant tuning.

The organization of this paper is as follows. Section 2 introduces the fundamentals of formant tuning and the concept of watermarking. Section 3 details the whole scheme of the formant tuning-based watermarking. In Sect. 4, the proposed method is evaluated with respect to inaudibility and robustness in comparison with other typical methods. In addition, a short discussion about the performance of the proposed method as well as other compared methods is also given out. Finally, we conclude our works in Sect. 5.

2. Concept Underlying Speech Watermarking

2.1 Related Studies on Formant Tuning

Formants correspond to concentrations of frequencies that are close to the resonance frequencies of the vocal tract. As a crucial acoustic feature for speech perception, formant needs to be enhanced or tuned when the quality or intelligibility of speech is impaired by various reasons. The method of re-shaping the formant to make it sharper is commonly referred as formant enhancement/tuning. This kind of method was originally developed in the adaptive post-filtering of speech codec to alleviate the perceptual effect

Manuscript received March 30, 2014.
Manuscript revised July 25, 2014.

[†]The authors are with the School of Information Science, Japan Advanced Institute of Science and Technology, Nomi-shi, 923–1292 Japan.
a) E-mail: wangshengbei@jaist.ac.jp

DOI: 10.1587/transinf.2014MUP0009

of quantization noise. Similar approaches that deal with formant to achieve better speech quality are widely found in the speech recognition system where the speech quality is reduced by noise [18], [19], and the hidden Markov model (HMM) based speech synthesis [20] where speech is muffled by the over-smoothed spectral envelope. In speech synthesis, the post-filtering technique for both mel-cepstrum based [21] and all-pole spectrum based [20], [22] spectra is applied to obtain a more prominent formant structure as well as improved sound quality.

Since formant tuning can improve the sound quality of speech, modifications introduced by formant tuning may not cause perceptual distortion to the original speech. Therefore, watermarking based on formant tuning is possible to be imperceptible to human to realize inaudibility. In this paper, we take advantage of formant tuning to achieve inaudibility for watermarking. However, in most speech synthesis methods, formants are tuned with complicated methods so that the dynamics between formant peaks and spectral valleys can be increased. As to accommodate watermarking with formant tuning, we investigate a direct but effective formant tuning method in this paper. The following subsections separately talk about how the formant can be estimated, tuned, and then applied for watermarking.

2.2 Formant Estimation and Formant Tuning

Formant estimation: The source-filter model of speech production is known as a linear model, in which the sound source, such as the glottis, and the filter that formed by the vocal tract, are assumed to be independent with each other. Based on the source-filter model, the set of linear prediction (LP) coefficients in Eq. (1) is an all-pole model that can provide accurate estimate of formants, where p indicates the LP order, a_i is the LP coefficient, $\hat{x}(n)$ is the prediction of $x(n)$, and $x(n-i)$ stands for the i-th previous sample. Removing the effect of formants with the inverse filter $A(z)$ in Eq. (2) is called inverse filtering. The signal $e(n)$ in Eq. (3) that left after inverse filtering is referred as residue.

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i) \tag{1}$$

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} \tag{2}$$

$$e(n) = x(n) - \sum_{i=1}^{p} a_i x(n-i) \tag{3}$$

In practice, the LP coefficients are often substituted with other representations such as LSFs and reflection coefficients (RCs) 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 spectra, which suggests that if LSFs are used to tune formant for watermark embedding, the distortion introduced by watermarks in both spectra and sound quality can be minimized; (*iii*) LSFs are universal features in different speech codecs, hence watermarks in LSFs are possible to survive from coding/decoding to provide the robustness against different speech codecs. According to these, we employ LSFs to tune formant for watermarking. The LSFs converted from LP coefficients satisfy the ordering property from 0 to π as follows, where p indicates the LP order, ϕ_i, $1 \le i \le p$, are the LSFs.

$$0 < \phi_1 < \phi_2 < \phi_3 < \cdots < \phi_p < \pi \tag{4}$$

Formant tuning: In general, 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 Eq. (5):

$$Q = \frac{f}{BW}, \tag{5}$$

where f stands for 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 of corresponding formant after converting them to frequency domain. For a fixed formant, when Q-value is increased, formant will be much sharper. Therefore, we tune a formant by increasing the Q-value. This can be realized by symmetrically closing up two LSFs to generate a narrower bandwidth. As shown in Fig. 1, the original formant (dotted curve) produced by two LSFs, ϕ_l and ϕ_r, has a tuning level Q_c defined in Eq. (6), where f_c is the center frequency, BW_c is the bandwidth between the frequencies f_l and f_r that converted from ϕ_l and ϕ_r with Eq. (7), where F_s is the sampling frequency of signal.

$$Q_c = \frac{f_c}{BW_c} = \frac{f_c}{f_r - f_l} \tag{6}$$

$$f_r = \frac{\phi_r}{2\pi} \times F_s \quad \text{and} \quad f_l = \frac{\phi_l}{2\pi} \times F_s \tag{7}$$

To tune this formant, as shown in Fig. 1, ϕ_l and ϕ_r are symmetrically shifted to close to each other, that is ϕ_l to ϕ_{lw} and ϕ_r to ϕ_{rw}. This process can be expressed as follows:

$$\phi_{lw} = \phi_l + \Delta \ \text{ and } \ \phi_{rw} = \phi_r - \Delta, \ \ 0 < \Delta < (\phi_r - \phi_l)/2, \tag{8}$$

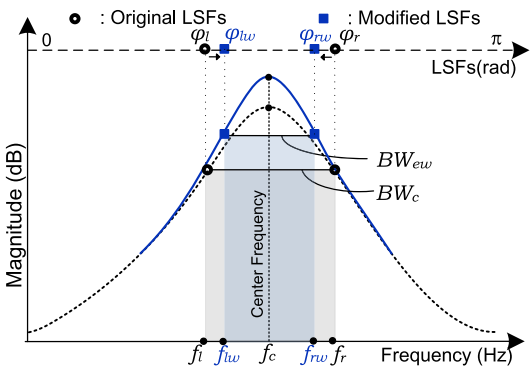


Fig. 1 Formant tuning by LSFs with respect to center frequency.

where Δ is the modification degree to control formant tuning, a bigger Δ indicates a more severe modification and a much tuned formant. After this, a narrower bandwidth BW_{ew} in Eq. (9) is produced. The tuning level of the newly obtained formant (solid curve in Fig. 1) is increased to Q_{ew}.

$$Q_{ew} = \frac{f_c}{BW_{ew}} = \frac{f_c}{f_{rw} - f_{lw}} \tag{9}$$

where f_{lw} and f_{rw} are calculated as follows:

$$f_{rw} = \frac{\phi_{rw}}{2\pi} \times F_s \quad \text{and} \quad f_{lw} = \frac{\phi_{lw}}{2\pi} \times F_s \tag{10}$$

Note that in the above method, two LSFs are symmetrically modified, there is no deviation in the center frequency which furthest maintains sound quality of original signal.

2.3 Watermarking Based on Formant Tuning

Preliminary analysis: We use the above formant tuning method for watermarking. Watermarks can be embedded into the host signal when LSFs are shifted for formant tuning. Before embedding process, several issues should be clarified to make the watermarking method effective.

(*i*) Selection of the suitable formant for tuning. Several formants can be estimated from the speech segment in each frame, we should select the suitable formant for tuning. As we have surveyed, the distortion caused by tuning formants in the lower and higher frequencies can be easily perceived by human, we thus leave the first formant and last formant unmodified. Only one formant in the middle region will be tuned for watermark embedding.

(*ii*) Embedding and blind detection mechanism. Formants in each frame can be consecutively indexed with F_1, F_2, F_3, $\cdots$, from the low frequency to high frequency. For different frames, if the watermarks are embed into the same indexed formants, it will be easy for the attackers to destroy them with simple rule. As to well hide watermarks, the formant for embedding will be randomly selected from each frame according to watermark '0' or '1' in our method. Moreover, since formant structures vary widely with different speech frames, it is preferable to tune the selected formants according to their original tuning characteristics (self-adaptive tuning) to achieve inaudibility.

However, the above embedding mechanism concerning random formant selection and self-adaptive tuning results in a serious problem for blind watermark detection since it is so difficult to detect watermarks just relying on the irregular formant structure extracted from the watermarked signal when any prior knowledge about which formant has been tuned and how it has been tuned is not available. As we have considered, one solution for both inaudibility and blind detection is we can tune the selected formant and hence to establish an internal relationship between the tuned formant and another formant in current frame, where the relationship is used to reflect the position of tuned formant and how the formant is tuned. In detection process, two formants can make a cross-reference. Watermarks can be extracted by identifying the relationship.

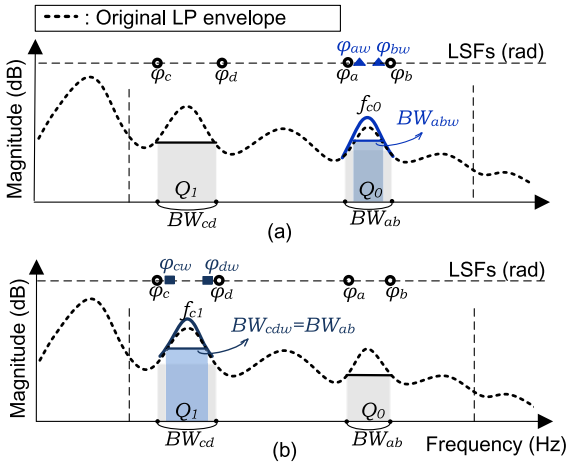


Fig. 2 Concept of watermark embedding: (a) embed '0' and (b) embed '1'.

Embedding concept: In our method, each speech frame will be embedded with one bit watermark, '0' or '1'. For each frame, firstly, we use LP analysis to estimate the formants. Secondly, we check the bandwidth (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 (produced by ϕ_a and ϕ_b, labelled as Q_0, $Q_0 = \frac{f_{c0}}{BW_{ab}}$) has the smallest bandwidth BW_{ab}, and the second sharpest formant (produced by ϕ_c and ϕ_d, labelled as Q_1, $Q_1 = \frac{f_{c1}}{BW_{cd}}$) has the second smallest bandwidth BW_{cd}. That is $BW_{cd} > BW_{ab}$. Relationships for embedding '0' and '1' will be established between two sharpest formants (the Q_0, Q_1 labelled formants) by tuning one of them. The reason why these two formants are selected to carry the relationships of watermarks will be explained later.

A. Rule of embedding '0': If '0' will be embedded, as seen in Fig. 2 (a), we will tune the sharpest formant with a tuning factor Ω_{e0} ($\Omega_{e0} > 1$). Therefore, the original bandwidth BW_{ab} will be reduced to its $1/\Omega_{e0}$. In Eq. (11), the newly obtained bandwidth BW_{abw} equals to $\frac{BW_{ab}}{\Omega_{e0}}$.

$$Q_0 \times \Omega_{e0} = \frac{f_{c0}}{BW_{ab}} \times \Omega_{e0} = \frac{f_{c0}}{BW_{abw}}, \quad \Omega_{e0} > 1 \tag{11}$$

As to reduce BW_{ab} to BW_{abw}, original LSFs ϕ_a and ϕ_b in Eq. (12) will be symmetrically shifted to ϕ_{aw} and ϕ_{bw} with respect to the center frequency f_{c0}.

$$\phi_{aw} = \phi_a + \Delta_{e0} \ \text{ and } \ \phi_{bw} = \phi_b - \Delta_{e0} \tag{12}$$

where the modification degree Δ_{e0} is calculated by ϕ_a, ϕ_b, and Ω_{e0} with Eq. (13).

$$\Delta_{e0} = \frac{1}{2}\left[(\phi_b - \phi_a) \times \left(1 - \frac{1}{\Omega_{e0}}\right)\right] \tag{13}$$

Since BW_{cd} is originally larger than BW_{ab}, after formant tuning, $BW_{cd} > BW_{abw} \times \Omega_{e0}$. And this relationship

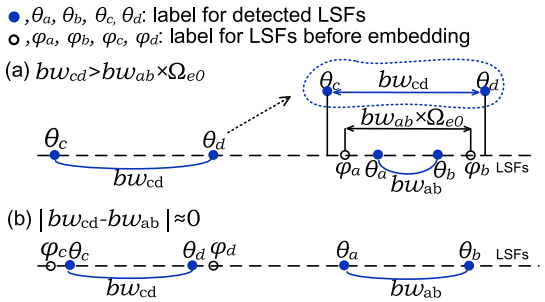


Fig. 3 Concept of watermark detection: (a) '0' is detected and (b) '1' is detected.

has been established in current frame for embedding '0'.

B. Rule of embedding '1': If '1' will be embedded, as seen in Fig. 2 (b), we will tune the second sharpest formant with a tuning factor $\Omega_{e1} = \frac{BW_{cd}}{BW_{ab}}$ in Eq. (14). With this factor, the newly obtained bandwidth BW_{cdw} of the second sharpest formant will be reduced to the same as BW_{ab}.

$$\begin{aligned} &Q_1 \times \Omega_{e1} = \frac{f_{c1}}{BW_{cd}} \times \Omega_{e1} = \frac{f_{c1}}{BW_{cdw}} = \frac{f_{c1}}{BW_{ab}}, \\ &\Omega_{e1} = \frac{BW_{cd}}{BW_{ab}} \end{aligned} \tag{14}$$

To achieve this, original LSFs ϕ_c and ϕ_d in Eq. (15) will be shifted to ϕ_{cw} and ϕ_{dw} as follows:

$$\phi_{cw} = \phi_c + \Delta_{e1} \text{ and } \phi_{dw} = \phi_d - \Delta_{e1} \tag{15}$$

where Δ_{e1} is calculated by ϕ_c, ϕ_d and Ω_{e1} with Eq. (16).

$$\Delta_{e1} = \frac{1}{2}\left[(\phi_d - \phi_c) \times \left(1 - \frac{1}{\Omega_{e1}}\right)\right] \tag{16}$$

Therefore, '1' can be embedded by establishing the relationship that the second sharpest formant has the same bandwidth as the sharpest formant, that is $BW_{cdw} = BW_{ab}$.

In summary, watermarks are embedded by tuning one formant and thus to establish different bandwidth relationships between the sharpest and the second sharpest formants. When '0' is embedded, bandwidth difference between the sharpest and the second sharpest formants is increased since the smaller bandwidth is reduced; while for '1', bandwidth difference is reduced to 0 since the larger bandwidth is reduced to the same as the smaller one. This opposite mechanism enables blind detection of watermarks.

Detection concept: According to the embedding concept, bandwidth relationships always exist in the sharpest and the second sharpest formants no matter for embedding '0' or '1'. In detection process, for each frame of the watermarked signal, we separately extract two smallest bandwidths from the sharpest and the second sharpest formants. As seen in Fig. 3, we name them as bw_{ab} (the 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 Eq. (17); if '1' has been embedded, according to Fig. 3 (b), bw_{cd} should be equal to bw_{ab}. Note

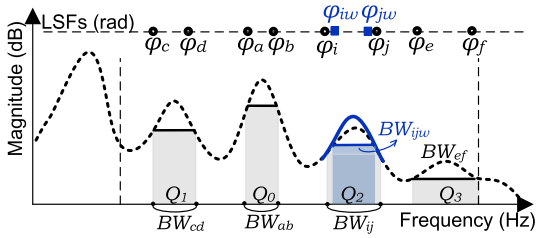


Fig. 4 Problem when tuning a smooth formant for watermarking.

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. 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 Eq. (18) to discriminate two cases of embedding '0' or '1', and enable the method to be error-tolerant.

$$bw_{cd} - bw_{ab} > bw_{ab} \times (\Omega_{e0} - 1) \tag{17}$$

$$\hat{s}(m) = \begin{cases} 0, & bw_{cd} - bw_{ab} > bw_{ab} \times (\Omega_{e0} - 1)/2 \\ 1, & \text{otherwise} \end{cases} \tag{18}$$

Embedding and detection analysis: Now we discuss why the sharpest and the second sharpest formants are selected to carry the relationship for watermarks. For the example in Fig. 4, three sharpest formants that labelled as Q_0 (the sharpest formant), Q_1 (the second sharpest formant), and Q_2 (the third sharpest formant) originally follow the bandwidth relationship that $BW_{ij} > BW_{cd} > BW_{ab}$. Consider one case that Q_0 and Q_2 labelled formants are selected for watermark embedding. To embed '1', BW_{ij} will be made to the same as BW_{ab} for formant tuning. Since $BW_{ij} > BW_{cd} > BW_{ab}$, the modification to BW_{ij} will be severer in comparison with tuning the Q_1 labelled formant. Therefore, sound quality will be much degraded. Alternatively, if we slightly reduce BW_{ij} to embed '1' and if BW_{ijw} in Fig. 4 is still larger than BW_{cd} after tuning, it will be difficult or even impossible to recognize bandwidth relationship for watermark detection. Although this phenomenon can be alleviated by setting bandwidth bounds for detection, formant tuning in embedding process, however, will be much hampered and complicated.

In comparison, establishing bandwidth relationships in the sharpest and the second sharpest formants can effectively avoid the above problem. This is because these two formants always possess two smallest bandwidths no matter before or after watermarking, so the bandwidth relationships in the detection process can be extracted for watermark detection without any ambiguity. Besides, the distortion introduced by formant tuning in this case can be minimized compared with tuning other formants.

3. Scheme of Formant Tuning-Based Watermarking

The proposed watermarking scheme is based on the speech

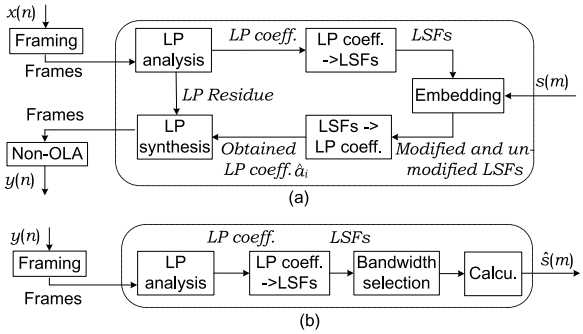


Fig. 5 Block diagram of watermarking: (a) embedding and (b) detection.

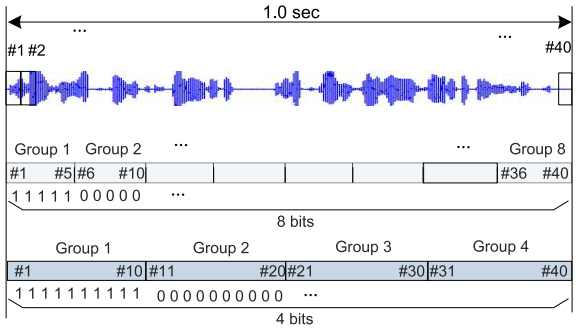


Fig. 6 Group separation for bit rates of 4 bps and 8 bps.

analysis/synthesis technique that explained in Sect. 2.2 (formant estimation). LP can analyze the speech signal by estimating the formants and extracting the residue. Watermarks are embedded by tuning one formant. Convolution of the residue signal (excitation signal) with the filter response that characterized by the tuned formant and the other formants can synthesize the watermarked signal.

Embedding process: Figure 5 (a) has a block diagram of the whole embedding process. The host signal $x(n)$ is segmented into non-overlapping frames. For each frame, LP analysis is applied to obtain LP coefficients and LP residue. The LP coefficients are converted to LSFs. Watermark is embedded into current frame according to the concept that introduced in Sect. 2.3, after which two modified LSFs are generated. All LSFs including the modified LSFs and the other un-modified LSFs are converted back to LP coefficients. The current frame is then synthesized by inputting the residue signal to the synthesis filter in Eq. (19) that characterized by the newly obtained LP coefficient $\hat{a}_i$. Watermarked signal, $y(n)$, is finally reconstructed with all watermarked frames using non-overlapping and adding function.

$$\frac{1}{\hat{A}(z)} = \frac{1}{1 - \sum_{i=1}^{p} \hat{a}_i z^{-i}} \tag{19}$$

Detection process: The detection process is illustrated in Fig. 5 (b). We apply the same procedures as those in embedding process to the watermarked signal $y(n)$ to obtain the LSFs of each frame. Two smallest bandwidths are then extracted. The watermark in each frame is detected with the method in Sect. 2.3. Each frame can be extracted with one bit. All extracted bits can construct the whole watermark signal, $\hat{s}(m)$.

4. Evaluations

4.1 Database and Conditions

We conducted several experiments with respect to inaudibility and robustness to evaluate the proposed method. Twelve speech stimuli (Japanese sentences, uttered by six males and six females) in the ATR speech database (B set) [23] were used as the host signals. All stimuli were clipped into 8.1-second duration, sampled at 20 kHz, and quantized with 16 bits. The embedded watermarks was "JAIST-IS-Acoustic". Since our method is based on speech analysis/synthesis, the frame size was fixed at 25 ms (40 frames in 1.0 second) to attain better sound quality. For extended use of watermarking as information hiding method, we evaluated the performance of the proposed method as a function of bit rate. To construct the bit rates, all frames within 1.0-second speech segment were separately divided into 4, 8, 20, and 40 groups. Frames within the same group were embedded with the same watermark and then detected the watermark with a majority decision. Thus, the bit rates for the proposed method were 4, 8, 20, and 40 bps. An example of frame separation at 4 bps and 8 bps is shown in Fig. 6.

Inaudibility could be checked by objective and subjective tests. The log spectrum distortion (LSD) [24] and perceptual evaluation of speech quality (PESQ) [25] were objective measures. They could estimate the degradation between the host signal and the watermarked signal. LSD in decibel (dB) was the spectral distance measure between the host signal and watermarked signal. LSD of 1.0 dB was the criterion, and a lower value indicated less distortion. PESQ recommended by ITU-T recommendation P.862 could map PESQ scores to objective difference grades (ODGs) that coverd from −0.5 (very annoying) to 4.5 (imperceptible). PESQ of 3.0 (slightly annoying) was set as the criterion.

Robustness could be evaluated by Bit Error Rate (BER) that defined as the percentage of mismatched bits between the embedded watermarks and the detected watermarks. A lower BER indicated stronger robustness. We chose BER of 10% as the criterion.

4.2 Parameter Analysis

In the proposed method, two adjustable parameters, i.e., LP order and Ω_{e0} for embedding '0' affect the performance of inaudibility and robustness (Ω_{e1} for embedding '1' is automatically fixed according to bandwidth characteristics of each frame). These two parameters should be optimized for the proposed method.

LP order: The order of LP analysis is important to determine the characteristics of formant structure. High LP order is beneficial to follow the details of spectrum contour, and more finer formants can be estimated under high LP order. Low LP order can just provide global frequency in-

formation, only a few global formants can be provided in this case. Under low order LP analysis, each estimated formant will carry more information in comparison with the formant that estimated by high order LP analysis. That is to say the sound distortion brought by tuning one formant that estimated with low order LP analysis will be more severe. Therefore, to achieve inaudibility, LP order should be as higher as possible. On the other hand, since most processing will bring distortion to the formant structure of watermarked signal, if LP order is so high to follow all the spectral details, any distortion will result in LSFs deviation, which will obstruct the watermark detection. In this case, LP order should be low to achieve robustness.

Modification degree Ω_{e0}: According to Eq. (17), bigger Ω_{e0} will increase the bandwidth difference between the sharpest formant and the second sharpest formant which makes it easier to discriminate '0' or '1'. However, bigger Ω_{e0} also means severe modification to the formant in the host signal which will degrade the sound quality severely.

The inaudibility and robustness are conflicting, and affected by LP order and Ω_{e0}. To select the optimal parameters, we tentatively checked the inaudibility and robustness performance (at 4 bps) as a function of LP order and Ω_{e0}. LP order was selected as 8, 10, 12, 14, 16, 18, and 20. Ω_{e0} was selected as 1.50, 1.65, 2.0, and 3.0. Since objective measures enable quick results, we evaluated inaudibility with LSD and PESQ. Robustness was checked by normal detection with BER results. We also checked the detection after G.729 (Code-excited linear prediction (CELP)) codec. This is because many watermarking methods [1] fail to extract watermarks after this codec. Therefore, robustness against G.729 is one of the most difficult criterion, which can typically check whether the method is robust or not.

According to the LSD and PESQ results in Fig. 7, we can find: (*i*) under the same Ω_{e0}, inaudibility was not obviously affected by different LP orders; (*ii*) under the same LP order, when Ω_{e0} was increased to 3.0, there was an obvious distortion in inaudibility. Therefore, Ω_{e0} should be less than 3.0 for inaudibility. The results in Fig. 8 shows the robustness results. We can see LP order and Ω_{e0} almost had nothing to do with normal detection, while it greatly influenced the robustness against G.729 since BER results drastically increased when LP order was increased. Hence, it would be prefer to choose lower LP order for robustness. According to these results, we finalized Ω_{e0} as 2.0 for inaudibility and LP order as 10 (where BER after G.729 at Ω_{e0} = 2.0 could be controlled below 10%) for robustness.

4.3 Comparative Evaluations

We follow the above parameters to evaluate the proposed method. Evaluations were also done to three typical methods: the least significant bit-replacement (LSB) method [3], direct spread spectrum (DSS) method [10], and cochlear delay (CD) method [26], which have separately exhibited excellent performance in inaudibility, robustness, and both inaudibility and robustness. A quick review of these methods is as follows: LSB replaces the least significant bits with watermarks at the quantization level so that replacement in less perceptible component does not cause distortion to human perception; DSS spreads watermarks over many (possibly all) frequency bands so that watermarks cannot be easily destroyed; CD embeds watermarks by enhancing the phase information of the host signal with respect to two kinds of cochlear delay (one is for '0' and the other one is for '1'). 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.

4.3.1 Evaluations of Inaudibility

Objective evaluations: LSD and PESQ results of the proposed method, CD, DSS, and LSB are plotted in Fig. 9. 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 objectively satisfy the inaudibility requirement.

Subjective evaluation: Inaudibility of the proposed method was also investigated via a listening test in which all twelve stimuli were involved. The following experiment conditions referred to those in [27]. For each stimulus, five test pairs were set up. Each test pair contained two tracks,

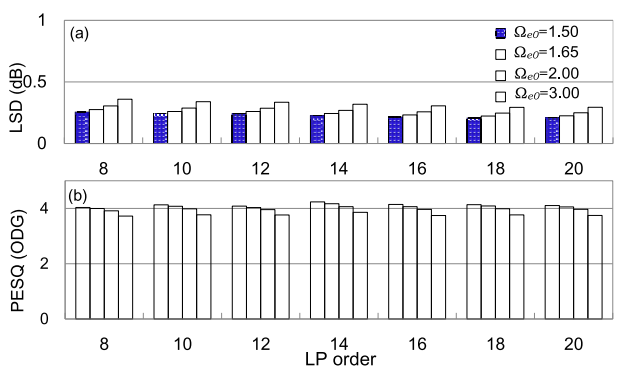


Fig. 7 Inaudibility affected by LP order and Ω_{e0}, (a) LSD and (b) PESQ.

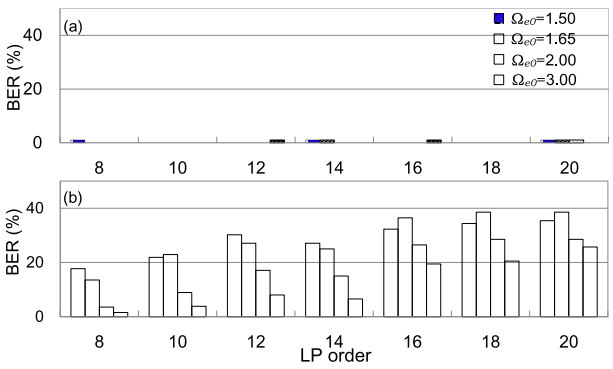


Fig. 8 Robustness affected by LP order and Ω_{e0}, (a) normal detection and (b) detection after G.729.

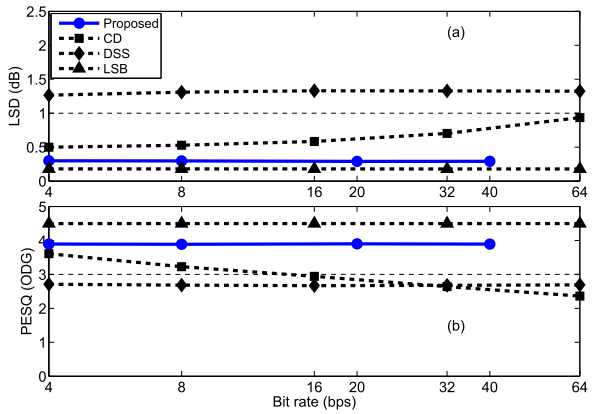


Fig. 9 Evaluation results of inaudibility: (a) LSD and (b) PESQ.

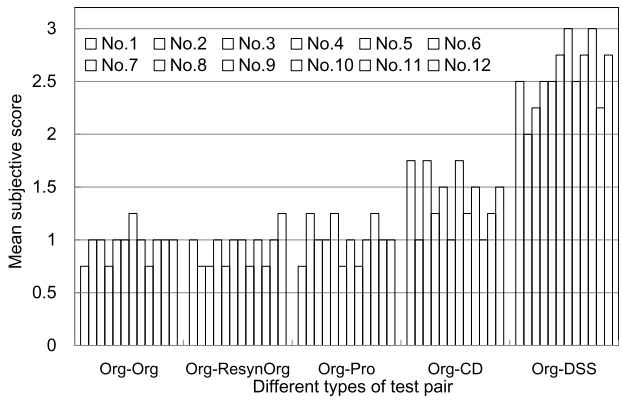


Fig. 10 Evaluation results of inaudibility with listening test.

one was the original (Org) stimulus and the other was the same original (Org) stimulus, or the resynthesized original (ResynOrg) stimulus, or the watermarked stimulus (at 4 bps) that was realized by the proposed method (Pro), CD, or DSS, where the test pair consisted of the original (Org) stimulus and resynthesized original (ResynOrg) stimulus was evaluated for the proposed method to check whether sound distortion could be caused by speech analysis/synthesis in spite of watermarks. Three male subjects and one female subject with normal hearing participated in the listening test. Each subject was presented with one test pair in a trial and then asked to report the similarity between two tracks by choosing a subjective score from 0 (completely the same), 1 (probably the same), 2 (probably different), and 3 (completely different). Each subject was totally presented with 60 test pairs (twelve stimuli × five pairs (Org-Org, Org-ResynOrg, Org-Pro, Org-CD, Org-DSS)).

The mean subjective scores on five test pairs for each stimuli are given out in Fig. 10. These results revealed that it was difficult for subjects to tell the difference between two tracks in the Org-Org, Org-ResynOrg, and Org-Pro test pairs, which suggested that the sound distortion caused by speech analysis/synthesis and watermarks embedding in proposed method was perceptually insignificant. In comparison, CD was slightly perceptible for a few stimuli, while DSS introduced obvious distortion to the host signals.

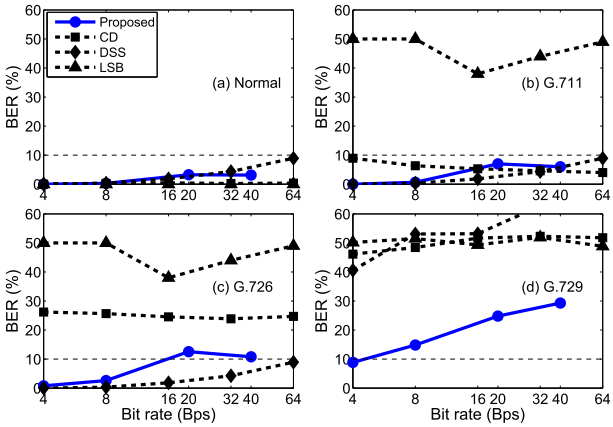


Fig. 11 Evaluation of robustness against speech codecs: (a) normal detection (no processing), (b) G.711, (c) G.726, and (d) G.729.

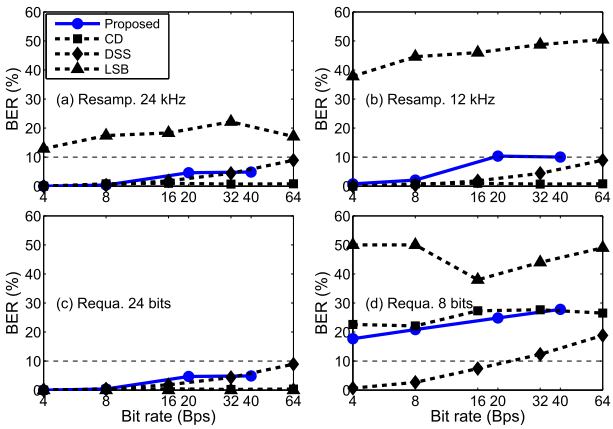


Fig. 12 Evaluation of robustness against re-sampling at (a) 24 kHz and (b) 12 kHz and re-quantization with (c) 24 bits and (d) 8 bits.

4.3.2 Evaluations of Robustness

Robustness against speech codecs: We applied three typical speech codecs of G.711 (pulse code modulation (PCM)), G.726 (adaptive differential PCM (ADPCM)), and G.729 to the watermarked signals. As shown in Fig. 11, none of the compared methods (CD, DSS, and LSB) could survive from all speech codecs, even for the robust DSS method. The proposed method was robust against normal detection, G.711, G.726, and G.729, although BER after G.729 was not so perfect. These results implied that the proposed method had good robustness against different speech codecs.

Robustness against speech processing: First, we evaluated the proposed method against general processing: (a) re-sampling at 12 kHz and 24 kHz, (b) re-quantization with 24 bits and 8 bits. Figure 12 plots all results. DSS obviously performed the best. LSB was only good for re-quantization with 24 bits. The proposed method and CD provided good performance except for re-quantization with 8 bits. The reason for this with the proposed method was re-quantization at lower rate compared with signal's orig-

Table 1 BER (%) results of robustness against practical processing.

Bit rate	Processing	CD	DSS	LSB	Proposed
4 bps	Ampl. by 2.0	3.57	0.00	50.00	0.78
	Ampl. by 0.5	3.57	0.00	52.58	0.78
	STFT	3.57	0.00	0.00	0.78
	GTFB	33.93	0.00	46.58	1.04
	Noise addition	3.57	0.00	0.00	0.00
	Echo addition	35.71	0.00	57.17	7.81
8 bps	Ampl. by 2.0	3.59	0.00	52.21	0.86
	Ampl. by 0.5	5.52	0.00	50.67	0.45
	STFT	6.91	0.00	0.00	0.78
	GTFB	39.47	0.00	48.73	1.25
	Noise addition	4.42	0.00	0.00	0.00
	Echo addition	36.81	0.00	55.36	6.36

inal parameter introduced some distortions to the watermarked signal, which destroyed the bandwidth relationship for watermark detection. Second, we evaluated the proposed method with other practical speech processing. These included (a) signal amplifying by 2.0 and 0.5, 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 the Information Hiding and its Criteria (IHC) committee [28] as reference, although these were designed to evaluate audio watermarking. These involved (d) Gaussian noise addition with an overall average SNR (signal to noise ratios) of 36 dB; (e) a single 100-ms echo addition of −6 dB. The BER results (in %) at 4 bps and 8 bps are listed in Table 1. The proposed method and DSS were robust against these processing.

4.4 Discussion and Future Work

Discussion on compared methods: We give a short discussion on the performance of all compared methods. LSB method embedded watermarks in the least significant bits so that the distortion was negligible which made the LSB perfectly inaudible. However, watermarks in the least significant bits could be easily reset by the operations related to amplitude modifications or lossy processing, which made this method fragile. DSS was relatively robust (except for G.729) since watermarks were spread over a wide frequency range, only all possible frequencies were destroyed with considerable strength could eliminate the watermarks. Therefore DSS exhibited strong robustness for most processing. However, watermarks in a wide frequencies made them perceptually significant. Watermarks in CD were embedded as phase information by modelling the cochlear delay. According to the characteristics of cochlear delay, detection of watermarks '0' and '1' strongly depended on the cue in low frequency phase. Correspondingly, once phase information in low frequency was destroyed or erased by other processing, such as GTFB and G.729 codec, watermarks could not be detected. In summary, LSB was not robust but inaudible, DSS was robust but not inaudible, and CD could conditionally satisfy inaudibility and robustness.

Discussion on the proposed method: The proposed method had several advantages. (*i*) It can basically satisfy both inaudibility and robustness. Formant tuning is capable to improve the sound quality of synthesized speech, watermarks embedded as formant tuning was almost inaudible to HAS. Watermark detection by identifying bandwidth relationship was able to tolerate small change of frequency components caused by other processing. Besides, each frame had its own frequency characteristic, the tuned formant (the sharpest formant or the second sharpest formant) was possible to exist in any frequency range. When small proportion of frequency components that did not contain watermarks were changed, watermarks were able to survive. Moreover, (*ii*) from the point of security, embedding watermarks into the intrinsically irregular formant structures made the watermarks confidential. This was because various formant structures made it difficult for the attackers or the third party to confirm whether the formant structure was formed by artificial manipulation or not, since embedded bandwidth relationship was also possible in a rough speech. Especially when the LP order for estimating formants was unknown, bandwidth relationship was unable to discover. Furthermore, as mentioned before, the tuned formant was possible to exist in any frequency, which made it difficult to eliminate watermarks by just destroying a narrow frequency range.

It is also important to note that although LSFs in the proposed method were shifted so that watermarks could be embedded, the proposed method was essentially different from QIM-based watermarking. This is because QIM-based watermarking modify embedding parameter without physical meaning, while the modification to LSFs in our method was motivated by formant tuning.

Nonetheless, our work left something to be desired. (*i*) The proposed method is frame-based watermarking, a frame synchronization scheme will be implemented in the future. (*ii*) We have investigated the inaudibility and robustness of the proposed method. In the next step, we will develop the proposed method for tampering detection.

5. Conclusion

This paper proposed a novel speech watermarking based on formant tuning. The property that formant can be tuned to improve the sound quality for synthesized speech was introduced to the proposed method to achieve inaudibility. To make the method effective, we investigated how the formant could be tuned with a pair of LSFs. Considering the desired performance of inaudibility and blind detection, watermarks were embedded as bandwidth relationships between the sharpest and the second sharpest formant by tuning one of them. We conducted several experiments to evaluated the proposed method. The evaluation results showed the proposed method possessed good performance in both inaudibility and robustness, which established a good foundation for tampering detection of speech signals in the next step.

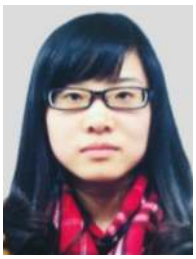
Acknowledgements

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. We would also like to express our great appreciation to Dr. Ryota Miyauchi who helped us in the subjective evaluations.

References

[1] C.-P. Wu and C.-C.J. Kuo, "Fragile speech watermarking based on exponential scale quantization for tamper detection," Proc. ICASSP, vol.IV, pp.3305–3308, 2002.
[2] M.D. Swanson, M. Kobayashi, and A.H. Tewfik, "Multimedia data-embedding and watermarking technologies," Proc. IEEE, vol.86, no.6, pp.1064–1087, 1998.
[3] P. Bassia and I. Pitas, "Robust audio watermarking in the time domain," Proc. EUSIPCO, pp.25–28, 1998.
[4] H.J. Kim and Y.H. Choi, "A novel echo-hiding scheme with backward and forward kernels," IEEE Trans. Circuits Syst. Video Technol., vol.13, no.8, pp.885–889, 2003.
[5] Y. Erfani and S. Siahpoush, "Robust audio watermarking using improved TS echo hiding," Digital Signal Process., vol.19, no.5, pp.809–814, 2009.
[6] S. Wu, J. Huang, D. Huang, and Y.Q. Shi, "Efficiently self-synchronized audio watermarking for assured audio data transmission," IEEE Trans. Broadcast., vol.51, no.1, pp.69–76, 2005.
[7] B. Lei, I.Y. Soon, F. Zhou, Z. Li, and H. Lei, "A robust audio watermarking scheme based on lifting wavelet transform and singular value decomposition," Signal Process., vol.92, no.9, pp.1985–2001, 2012.
[8] X. Wang and H. Zhao, "A novel synchronization invariant audio watermarking scheme based on DWT and DCT," IEEE Trans. Signal Process., vol.54, no.12, pp.4835–4840, 2006.
[9] H.S. Malvar and A.F. Florêncio, "Improved spread spectrum: A new modulation technique for robust watermarking," IEEE Trans. Signal Process., vol.51, no.4, pp.898–905, 2003.
[10] L. Boney, H.H. Tewfik, and K.H. Hamdy, "Digital watermarks for audio signals," Proc. ICMCS, pp.473–480, 1996.
[11] F. Hartung, J.K. Su, and B. Girod, "Spread spectrum watermarking: Malicious attacks and counter attacks," Proc. SPIE, vol.3657, pp.147–158, 1999.
[12] D. Kirovski and H. Malvar, "Robust spread spectrum audio watermarking," Proc. ICASSP, vol.3, pp.1345–1348, 2001.
[13] M.D. Swanson, B. Zhu, A.H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," Signal Process., vol.66, no.3, pp.337–355, 1998.
[14] F. Battisti, M. Carli, and C. Rinaldi, "Perceptual audio watermarking driven by Human Auditory System," Proc. ISSCS, pp.1–4, 2013.
[15] M. Unoki and D. Hamada, "Method of digital-audio watermarking based on cochlear delay characteristics," Int. J. Inn. Com. Inf., and Cont., vol.6, no.3(B), pp.1325–1346, 2010.
[16] S. Wang and M. Unoki, "Watermarking method for speech signals based on modifications to LSFs," Proc. IIHMSP, pp.283–286, 2013.
[17] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," J. Acoust. Soc. Am., vol.57. no.537(A), pp.35–35, 1975.
[18] H. Brouckxon, W. Verhelst, and B.D. Schuymer, "Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments," Proc. Interspeech, pp.557–560, 2008.
[19] Y. Ueda, S. Hario, and T. Sakata, "Formant based speech enhancement for listening speech sound in noisy place," Proc. ICSV, pp.515–522, 2008.
[20] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Comparison of formant enhancement methods for HMM based speech synthesis," Proc. ISCA Speech Synthesis Workshop, 2010.
[21] HTS, "HMM-based speech synthesis system," http://hts.sp.nitech.ac.jp
[22] F.K. Soong and B.H. Juang, "Line spectrum pair (LSP) and speech data compression," Proc. ICASSP, vol.9, pp.37–40, 1984.
[23] K. Takeda et al, "Speech database user's manual," ATR Technical Report, TR-I-0028, 2010.
[24] A. Gray, Jr. and J. Markel, "Distance measures for speech processing," IEEE Trans. Acoust. Speech Signal Process., vol.24, no.5, pp.380–391, 1976.
[25] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Trans. Audio Speech Language Process., vol.16, no.1, pp.229–238, 2008.
[26] M. Unoki and R. Miyauchi, "Reversible watermarking for digital audio based on cochlear delay characteristics," Proc. IIHMSP, pp.314–317, 2011.
[27] M. Unoki, K. Imabeppu, D. Hamada, A. Haniu, and R. Miyauchi, "Embedding limitations with digital-audio watermarking method based on cochlear delay characteristics," J. Inf. Hid. and Mul. Sig. Proc., vol.2, no.1, 2011.
[28] Information hiding and its criteria for evaluation, http://www.ieice.org/iss/emm/ihc/en/index.php



Shengbei Wang received her B.E and M.E. degrees in Electronic Information Engineering from TianJin Polytechnic University, TianJin, China, in 2009 and 2012, respectively. From October in 2012, she is pursuing the Ph.D. degree with the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan. Her present research interests are digital signal watermarking of speech signals and watermarking-based tampering detection for speech signals. She worked as a Graduate Research Program (GRP) student from Oct., 2012 to Oct., 2013. Now she is a Scholarship Student of the China Scholarship Council.

Masashi Unoki received his M.S. and Ph.D. (Information Science) from the Japan Advanced Institute of Science and Technology (JAIST) in 1996 and 1999. His main research interests are in auditory motivated signal processing and the modeling of auditory systems. He was a Japan Society for the Promotion of Science (JSPS) research fellow from 1998 to 2001. He was associated with the ATR Human Information Processing Laboratories as a visiting researcher from 1999–2000, and he was a visiting research associate at the Centre for the Neural Basis of Hearing (CNBH) in the Department of Physiology at the University of Cambridge from 2000 to 2001. He has been on the faculty of the School of Information Science at JAIST since 2001 and is now an associate professor. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, and the Acoustical Society of America (ASA). He is also a member of the Acoustical Society of Japan (ASJ), and the International Speech Communication Association (ISCA). Dr. Unoki received the Sato Prize from the ASJ in 1999, 2010, and 2013 for an Outstanding Paper and the Yamashita Taro "Young Researcher" Prize from the Yamashita Taro Research Foundation in 2005.