INAUDIBLE SPEECH WATERMARKING BASED ON SELF-COMPENSATED ECHO-HIDING AND SPARSE SUBSPACE CLUSTERING

Shengbei Wang[⋆] *Weitao Yuan*[⋆] *Jianming Wang*[⋆∗] *Masashi Unoki* [‡†]

[⋆] Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems, Tianjin Polytechnic University, Tianjin, China
[‡]Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, Japan

ABSTRACT

The method reported here realizes an inaudible echo-hiding based speech watermarking by using sparse subspace clustering (SSC). Speech signal is first analyzed with SSC to obtain its sparse and low-rank components. Watermarks are embedded as the echoes of the sparse component for robust extraction. Self-compensated echoes consisting of two independent echo kernels are designed to have similar delay offsets but opposite amplitudes. A one-bit watermark is embedded by separately performing the echo kernels on the sparse and low-rank components. As a result, the sound distortion caused by one echo signal can be quickly compensated by the other echo signal, which enables better inaudibility. Since the embedded echoes have the same sparsity as the sparse component, watermarks can be extracted with a basic cepstrum analysis even if the echo kernels are not directly performed on the original speech. The evaluation results verify the feasibility and effectiveness of this method.

Index Terms— Echo-hiding, sparsity, sparse subspace clustering, speech watermarking

1. INTRODUCTION

Speech watermarking is the process of hiding imperceptible information (watermarks) in the original speech [1, 2]. It is considered to be a practical way to protect speech and has been studied for a few decades [3]. An effective watermarking method should satisfy several conflicting requirements, e.g., inaudibility, blindness, robustness, and security. Since these requirements mutually restrict and rely on each other, the design of a speech watermarking scheme is usually treated as a problem of seeking an artful balance among them [4].

In literature, audio watermarking has been sufficiently studied compared with speech watermarking. In particular, because of the similarities between audio and speech, many successful audio watermarking techniques have been adapted to speech signals and found effective for them, e.g., least significant bit-replacement (LSB) [5], direct spread spectrum (DSS) modulation [6, 7], cochlear delay (CD) [8], and phase modulation [9]. As a representative technique for audio watermarking, echo-hiding [10, 11], has attracted considerable attention for its remarkable inaudibility, robustness, etc. However, the echo-hiding method has been rarely used on speech. Many studies claim that echo-hiding is uniquely for audio [10, 12]. One reason behind this claim is that the human auditory system (HAS) is more sensitive to echoes of clean speech than to echoes of general audio. This makes echo-hiding based watermarking a rather challenging task for speech signals. Moreover, it has been found that most echo-hiding methods apply the echo kernel directly to the original whole signal [10, 11, 12, 13]. One advantage is that the cepstrum of the watermarked signal can thus be simplified into a sum of the cepstrum of the original signal and the cepstrum of the echo kernel, which facilitates watermark extraction [13, 14]. The mathematical derivation for cepstrum analysis would be rather complicated if the echo kernel were not performed in this way. Although [15] proposes to apply the echo kernel to two subsignals of the original signal, it still requires the values of two adjacent samples in two subsignals to be amlost identical. Under this strict constraint, a general cepstrum analysis can be used for watermark extraction [16].

Two main issues that we deal with here are (i) how to embed the echo effectively for speech watermarking without degrading the speech quality and (ii) how to extract the watermarks when the echo kernels are not directly applied to the original whole speech. To address these problems, we investigated a new echo-hiding mechanism, i.e., self-compensated echoes, by taking advantage of the sparsity of the original speech. Unlike most methods that perform the echo kernel directly on the original signal, the proposed kernels are performed on only part of the speech, i.e., its sparse component, which is obtained by sparse subspace clustering (SSC). One benefit leading to better security is that the embedded echoes can hardly be extracted without prior knowledge on how the sparse component was extracted and what parameters were used. The self-compensated echoes consisting of two echo kernels were designed to have opposite amplitudes to maximally reduce the perceptual distortion. Benefiting from SSC, the watermarks can be extracted with a basic cepstrum analysis due to the sparsity of the embedded echoes, even the echo kernels are not performed on the original speech, which can be considered to be a new manner of echo-hiding.

2. PROPOSED METHOD

Human speech varies significantly over time, and its power concentrates on formants. Consequently, the spectrogram about speech has a relatively sparse structure [17, 18] and a speech signal can be separated into at least two main components, i.e., a sparse component and a low-rank component. The sparse component contains important information of speech, so we embed watermarks as echoes of the sparse component to strengthen it for robust watermark extraction.

[∗]Thanks to Natural Science Foundation of Tianjin (No. 17JCQNJC00100 and No. 16JCYBJC41500), the Science&Technology Development Fund of Tianjin Education Commission for Higher Education (No. 2017KJ089 and No. 2018KJ218), National Natural Science Foundation of China (No. 6137104 and No. 61602344), and the Program for Innovative Research Team in University of Tianjin (No. TD13-5032).

[†]This work was also supported by a Grant-in-Aid for Scientic Research (B) (No. 17H01761) and I-O DATA foundation.

2.1. Sparse subspace clustering for speech separation

The problem of separating speech into its sparse component and low-rank component can be solved with Robust Principal Component Analysis (RPCA), as long as the data matrix with points as column vectors has approximately low-rank [18, 19, 20]. However, this assumption cannot always be satisfied in real situations, since the data may lie in a number of lower-dimensional subspaces other than a single subspace, e.g., as in the case of audio data [20]. Considering a future generalization of the proposed method to audio signals, we chose to use SSC [21] to do the separation, which is capable of separating the data samples to multiple lower-dimensional subspaces according to their underlying attributes, using sparse representation techniques [22].

The process of SSC for sparse and low-rank separation is formulated as follows. Given a speech signal, we divide it into non-overlapping frames. Each frame is expressed as $x(n) \in \mathbb{R}_{n\times 1}$ of n samples. Here, we require $\sqrt{n}$ to be an integer and larger than σ (e.g., σ=10). Each frame $x(n)$ can be reshaped into a square matrix $\boldsymbol{X}_F \in \mathbb{R}_{N\times N}$, $N=\sqrt{n}$. Starting from the simplest case of SSC, we suppose that the data points of one column, $\boldsymbol{x}_i \in \mathbb{R}_{N\times 1}$, $1 \le i \le N$, of speech frame $\boldsymbol{X}_F$ lie in K linear subspaces and the dimension of each subspace is smaller than N. According to the self-expressiveness property, $\boldsymbol{x}_i$ in $\boldsymbol{X}_F$ can be written as a linear combination of the other points in $\boldsymbol{X}_F$, i.e.,

$$\boldsymbol{x}_i = \boldsymbol{X}_F \boldsymbol{c_i}, \quad c_{ii} = 0, \tag{1}$$

where $\boldsymbol{c_i} = [c_{i1}, c_{i2}, \cdots, c_{iN}]^T$, $\boldsymbol{X}_F$ is called the self-expressive dictionary, and the constraint $c_{ii} = 0$ avoids expressing a data point as a linear combination of itself. For Eq. (1), there ideally exists an efficient subspace-sparse representation, $\hat{\boldsymbol{c}}_i$, whose nonzero entries correspond to data points from the same subspace, as $\boldsymbol{x}_i$. To find this $\hat{\boldsymbol{c}}_i$, Eq. (1) is restricted by minimizing the objective function $\boldsymbol{c_i}$ under the l_1-norm, i.e.,

$$\min_{\boldsymbol{c_i}} \|\boldsymbol{c_i}\|_{l_1} \quad \text{s.t.} \quad \boldsymbol{x}_i = \boldsymbol{X}_F \boldsymbol{c_i}, \quad c_{ii} = 0, \tag{2}$$

This can be rewritten in matrix form for all data points,

$$\min_{\boldsymbol{C}} \|\boldsymbol{C}\|_{l_1} \quad \text{s.t.} \quad \boldsymbol{X}_F = \boldsymbol{X}_F \boldsymbol{C}, \quad \text{diag}(\boldsymbol{C}) = 0, \tag{3}$$

where the i-th column of $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_N] \in \mathbb{R}_{N\times N}$ corresponds to the sparse representation of $\boldsymbol{x}_i$.

Since the speech signal contains both sparse and low-rank components, its data does not lie in a union of low-dimensional subspaces. Therefore, the basic form $\boldsymbol{X}_F = \boldsymbol{X}_F \boldsymbol{C}, \text{diag}(\boldsymbol{C}) = 0$ in Eq. (3) should be generalized as,

$$\boldsymbol{X}_F = \boldsymbol{X}_F \boldsymbol{C} + \boldsymbol{S}, \quad \text{diag}(\boldsymbol{C}) = 0, \tag{4}$$

where $\boldsymbol{S}$ corresponds to the matrix of sparse outlying entries. Accordingly, we have

$$\begin{aligned} &\min_{\boldsymbol{C},\boldsymbol{S}} \quad \|\boldsymbol{C}\|_{l_1} + \lambda_s \|\boldsymbol{S}\|_{l_1} \\ &\text{s.t.} \quad \boldsymbol{X}_F = \boldsymbol{X}_F \boldsymbol{C} + \boldsymbol{S}, \quad \text{diag}(\boldsymbol{C}) = 0, \end{aligned} \tag{5}$$

where $\lambda_s > 0$ balances the two terms in the objective function and l_1-norm promotes sparsity in the columns of $\boldsymbol{C}$ and $\boldsymbol{S}$. Equation (5) can be solved using convex programming tools [21]. The optimal solution $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{S}}$ of Eq. (5) expresses the $\boldsymbol{X}_F$ in the form of its low-rank component $\boldsymbol{L}_F \in \mathbb{R}_{N\times N}$ and its sparse component $\boldsymbol{S}_F \in \mathbb{R}_{N\times N}$, where $\boldsymbol{L}_F = \boldsymbol{X}_F \hat{\boldsymbol{C}}$ is the component of $\boldsymbol{X}_F$ expressed with the union of multiple low-dimensional subspace signals and $\boldsymbol{S}_F$ (equals $\hat{\boldsymbol{S}}$), $\boldsymbol{S}_F = \boldsymbol{X}_F - \boldsymbol{L}_F$, respectively. Finally, $\boldsymbol{L}_F$ and $\boldsymbol{S}_F$ are reshaped into low-rank signal $l(n) \in \mathbb{R}_{n\times 1}$ and sparse signal $s(n) \in \mathbb{R}_{n\times 1}$, respectively. Note that when adapting the current SSC model to audio signals, the multiple low-dimensional subspace audio signals can be separately calculated by using $\boldsymbol{X}_F \hat{\boldsymbol{c}}_i$, $1 \le i \le N$, where $\hat{\boldsymbol{c}}_i$ is the i-th column of $\hat{\boldsymbol{C}} \in \mathbb{R}_{N\times N}$.

2.2. Watermark embedding algorithm

Watermarks are embedded to emphasize the sparse component with self-compensated echo kernels, which consists of two independent echo kernels $h_p(n)$ and $h_q(n)$, as follows:

$$h_p(n) = a\delta(n-d_*) + a\delta(n+d_*), \tag{6}$$

$$h_q(n) = -a\delta(n-d_*-\Delta) - a\delta(n+d_*+\Delta), \tag{7}$$

where $\delta(\cdot)$ denotes the Dirac delta function, a ($0 < a < 1$) is the amplitude of the echoes, and $d_* = \{d_0, d_1\}$ is the delay of the echo signal determined by the watermark bits.

Inaudibility is achieved by collaboration between $h_p(n)$ and $h_q(n)$. Here, the first item of $h_p(n)$, i.e., $a\delta(n-d_*)$, produces an echo with a positive amplitude a, while the first item of $h_q(n)$, i.e., $-a\delta(n-d_*-\Delta)$, produces an echo with a negative amplitude $-a$, where the offset Δ (a small integer) in $h_q(n)$ determines the delay time of the first echo in $h_q(n)$ relative to the first echo in $h_p(n)$. Because of their opposite amplitudes and the short offset Δ between them, the sound distortion introduced by the first echo is quickly weakened by the second echo, which leads to better sound quality. The second items in $h_p(n)$ and $h_q(n)$ are designed in the same manner. Moreover, both $h_p(n)$ and $h_q(n)$ contain forward and backward kernels themselves. It has been proven that such kernels can increase the peak of cepstrum for more robust watermark extraction [23].

It seems reasonable to perform the echo kernels $h_p(n)$ and $h_q(n)$ on $l(n)$ and $s(n)$ separately. However, as the offset Δ between $h_p(n)$ and $h_q(n)$ will cause the echoes of $l(n)$ and $s(n)$ to be misaligned, the speech quality will be distorted. More importantly, since $l(n)$ and $s(n)$ are different from each other, their echoes cannot be compensated effectively even with the opposite amplitude. Hence, the watermarked signal $y(n)$ is obtained by only emphasizing the sparse component, $s(n)$, i.e.,

$$\begin{aligned} y(n) &= x(n) + \xi s(n) \otimes h_p(n) + \xi s(n) \otimes h_q(n) \\ &= x(n) + \xi(s(n) \otimes (h_p(n) + h_q(n))), \end{aligned} \tag{8}$$

where the operator $\otimes$ denotes convolution and ξ ($0 < \xi < 1$) controls the energy of the echo.

To clearly explain the watermark extraction process in subsection 2.3, in fact, we implement Eq. (8) by performing $h_p(n)$ and $h_q(n)$ on $l(n)$ and $s(n)$ separately, i.e.,

$$\tilde{l}(n) = l(n) + \xi(s(n) \otimes h_p(n)), \tag{9}$$

$$\tilde{s}(n) = s(n) + \xi(s(n) \otimes h_q(n)). \tag{10}$$

The watermarked signal $y(n)$ is the sum of $\tilde{l}(n)$ and $\tilde{s}(n)$, which is derived the same as Eq. (8),

$$\begin{aligned} y(n) &= \tilde{l}(n) + \tilde{s}(n) \\ &= x(n) + \xi(s(n) \otimes (h_p(n) + h_q(n))). \end{aligned} \tag{11}$$

2.3. Watermark extraction algorithm

When the echo kernel, e.g., $h(n)$, is directly performed on the original signal $x(n)$, i.e., $y(n) = x(n) \otimes h(n)$, the cesptrum of $y(n)$ simplifies to $\mathcal{C}_{y(n)} = \mathcal{C}_{x(n)} + \mathcal{C}_{h(n)}$, where $\mathcal{C}_{(\cdot)} = \mathcal{F}^{-1}(\log(\mathcal{F}(\cdot)))$, $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ stand for Fourier transform and inverse Fourier transform, respectively. However, if the echo kernel is performed on a non-linear transformed signal of $x(n)$, e.g., $y(n) = x(n) + F(x(n)) \otimes h(n)$, where $F(\cdot)$ stands for the non-linear transformation, $\mathcal{C}_{y(n)} = \mathcal{F}^{-1}(\log(\mathcal{F}(x(n) + F(x(n)) \otimes h(n))))$ can no longer be simplified into a sum of $\mathcal{C}_{x(n)}$ and $\mathcal{C}_{h(n)}$ [13]. Moreover, it is difficult to ensure the convergence of the Taylor series in this case [14, 16].

Obviously, the proposed method suffers from the above problem (see Eq. (8)) and the $s(n)$ obtained by SSC is typically a non-linear transformed signal of $x(n)$. Nevertheless, since the embedded echoes are generated by the sparse component $s(n)$ in our method, theoretically, they should have the same sparsity as $s(n)$. As a result, the embedded echoes will be completely assigned to the sparse component if we use the same parameters λ_s in the embedding process to analyze the watermarked signal. Accordingly, we have

$$\breve{l}(n) \approx l(n), \tag{12}$$

$$\breve{s}(n) \approx s(n) + \xi(s(n) \otimes (h_p(n) + h_q(n))), \tag{13}$$

where $\breve{l}(n)$ and $\breve{s}(n)$ are the extracted low-rank signal and sparse signal, respectively. According to Eq. (13), watermark extraction is only related to $\breve{s}(n)$. By re-writing $s(n)$ in form of $s(n) \otimes \delta(n)$, Eq. (13) can be formulated as

$$\breve{s}(n) \approx s(n) \otimes \underbrace{(\delta(n) + \xi(h_p(n) + h_q(n)))}_{h_s(n)}, \tag{14}$$

where $h_s(n)$ stands for the kernel for $s(n)$. Watermarks can be extracted with a basic cepstral analysis of Eq. (14), i.e.,

$$\begin{aligned} \mathcal{C}_{\breve{s}(n)} &\approx \mathcal{C}_{s(n)} + \mathcal{C}_{h_s(n)} \\ &= \mathcal{F}^{-1}(\log S(w)) + \mathcal{F}^{-1}(\log H_s(w)). \end{aligned} \tag{15}$$

More specifically, $H_s(w)$ can be rewritten as

$$H_s(w) = 1 + 2a\xi(\cos w d_* - \cos w(d_* + \Delta)) \tag{16}$$

Using a trigonometric function $\cos w d_* - \cos w(d_* + \Delta) = 2(\sin(2wd_* + w\Delta)/2 \times \sin w\Delta/2)$, when we adjust a and ξ to make sure $4a\xi < 1$, Eq. (16) can be expanded in a Taylor series:

$$\begin{aligned} \log H_s(w) = {} & 2a\xi(\cos w d_* - \cos w(d_* + \Delta)) \\ & - \frac{(2a\xi)^2}{2}(\cos w d_* - \cos w(d_* + \Delta))^2 + \cdots . \end{aligned} \tag{17}$$

The cepstrum of $h_s(n)$ can be expressed as

$$\begin{aligned} \mathcal{C}_{h_s(n)} = {} & a\xi[\delta(n - d_*) + \delta(n + d_*)] \\ & - a\xi[\delta(n - d_* - \Delta) + \delta(n + d_* + \Delta)] + \cdots . \end{aligned} \tag{18}$$

The most dominant peaks appear at $n = d_*$ and $n = d_* + \Delta$ can be used for watermark extraction.

It should be noted that SSC provides a way to cluster speech data according to their underlying characteristics, and this facilitates and simplifies echo extraction when echo kernels are not performed on the original whole signal. Since the echoes generated from the subspace signals have the same characteristics as them, they will be well preserved within the subspace signals for extraction. Moreover, the proposed method can potentially be further developed by separately applying suitable echo kernels to the low-rank component, sparse component, and multi-subspace signals.

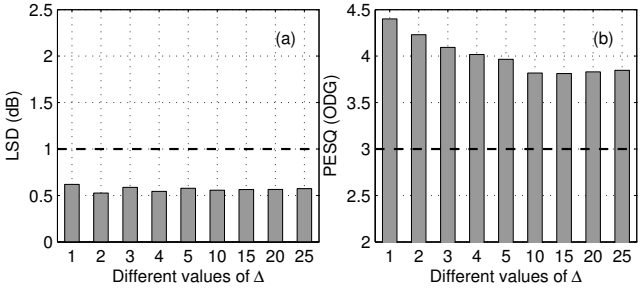

Fig. 1. Speech quality for varying offsets Δ.

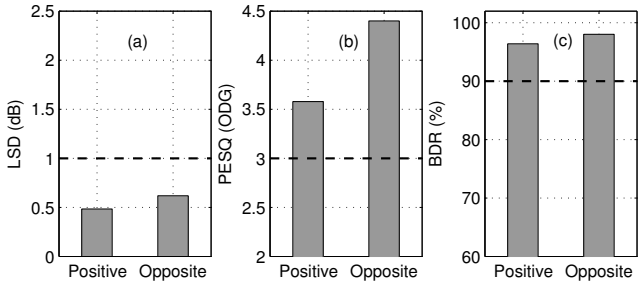

Fig. 2. Performance of proposed method using positive and opposite echo kernels (6 bps).

3. EVALUATIONS

We experimentally evaluated our method in terms of its inaudibility and robustness. Inaudibility was measured in terms of the log-spectrum distortion (LSD) [24] and a perceptual evaluation of speech quality (PESQ) [25]. An LSD of 1.0 dB was chosen as the criterion, and a lower value indicates less distortion. The PESQ evaluated the speech quality with Objective Difference Grades (ODGs), where ODGs were graded from -0.5 (very annoying) to 4.5 (imperceptible), corresponding to Mean Opinion Score (MOS) values of 1.0 to 5.0. The ODG of 3.0 (slightly annoying) was set as the criterion, and a higher value indicates better quality. The bit detection rate (BDR), which is defined as the ratio between the correctly extracted bits and the embedded bits, was selected to evaluate robustness. A higher BDR indicates stronger robustness.

The 12 speech signals in the ATR database (B set) (8.1-sec, 20 kHz, and 16 bits) were used as stimuli [26]. The performance of the proposed method depends heavily on its parameters. According to our preliminary experiments, the λ_s was set as 50 to attain the best results. The a and ξ were set as 0.45 and 0.5, respectively, to balance inaudibility and robustness while assuring the convergence of the Taylor series. The delay d_* in Eqs. (6) and (7) was set as $d_0 = 31$ for bit 0 and $d_1 = 60$ for bit 1. The embedded watermark was a random binary sequence. The embedding capacities were set as 6, 10, 13, 18, 25, 40, 70, and 160 bps. All of the reported results were calculated on the average of 12 speech signals.

1) Inaudibility affected by offset Δ**:** The proposed method takes advantage of successive but opposite echoes for inaudibility. To check how the offset Δ of the successive echoes affects speech quality, we used a gradually increasing Δ in $h_q(n)$ for embedding, i.e., $\Delta = \{1, 2, 3, 4, 5, 10, 15, 20, 25\}$. Figure 1 plots the inaudibility results, where embedding capacity was fixed at 6 bps. The LSD results remained almost unchanged as Δ increased. In contrast, the PESQ results got worse when Δ increased. These results suggest that a

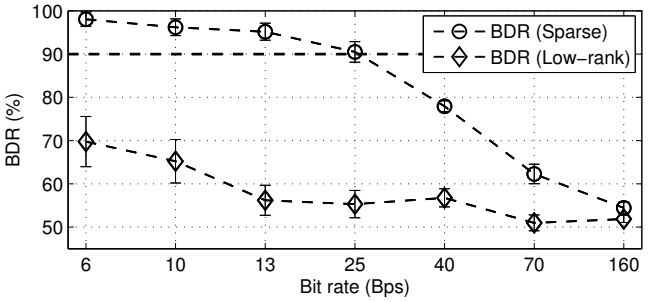

Fig. 3. Watermark extraction based on sparsity of embedded echoes.

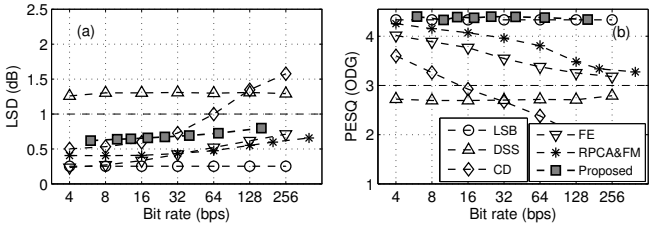

Fig. 4. Comparative results on inaudibility.

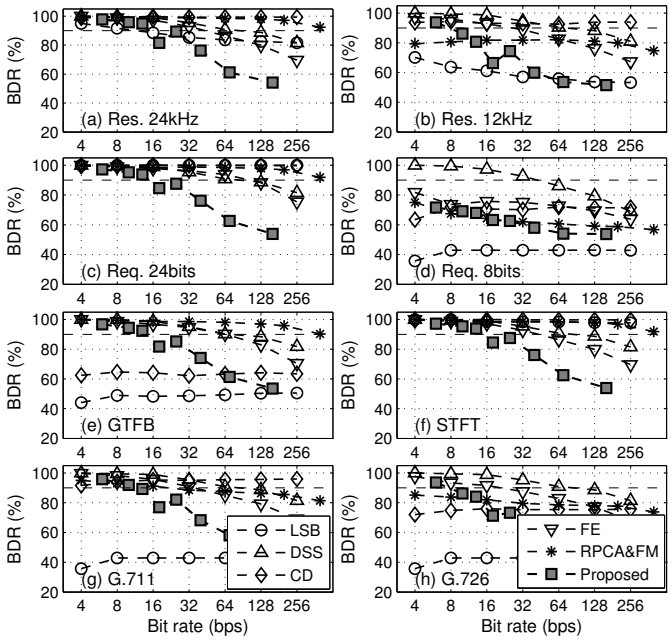

Fig. 5. Comparative results on robustness.

shorter offset enables two opposite echoes to be better compensated and that using a small Δ would yield satisfactory inaudibility.

***2) Effectiveness of self-compensated echoes*:** We compared two types of kernel to check the effectiveness of the self-compensated echo kernels in preserving speech quality: one was the same as $h_p(n)$ and $h_q(n)$ in Eqs. (6) and (7), and the other was $h_p(n)$ and $|h_q(n)|$, i.e., both $h_p(n)$ and $|h_q(n)|$ had positive amplitudes. In both cases, we set a one-sample offset, $\Delta = 1$. The LSD and PESQ results are plotted in Figs. 2(a) and 2(b). The self-compensated echo kernels with opposite amplitudes provided better speech quality compared with positive amplitudes. This result verifies the effectiveness of the proposed echo kernels. In addition, according to Eq. (18), the BDR results of the two cases should be the same. The results in Fig. 2 (c) indeed show that the BDRs of the two cases were similar to each other.

***3) Watermark extraction based on sparsity of embedded echoes*:** The simple watermark extraction of our method relies on the assumption that the embedded echoes generated by the sparse component have the same sparsity. To verify that this assumption is valid, we separately extracted the watermarks from the sparse component and the low-rank component. The BDR results in Fig. 3 shows that the watermarks were correctly extracted from the sparse component. In contrast, the BDR of the low-rank component was quite low. Even though the averaged BDR of the low-rank part reached almost 70% at 6 bps, the deviation was quite high, suggesting that the results were not stable on different speech signals. Overall, these results verify our assumption.

***4) Comparative evaluations*:** Finally, we compared our method with other methods, i.e., LSB [5], DSS [6], CD [8], watermarking based on formant enhancement (FE) [27], and watermarking based on Robust PCA and formant manipulations (RPCA&FM) [28]. The embedding capacities of the LSB, DSS, CD, and FE methods were 4, 8, 16, 32, 64, 128, and 256 bps, while the embedding capacities of RPCA&FM were 4, 8, 16, 32, 64, 128, 200, and 400 bps, in accordance with their original implementations.

The comparative results on inaudibility are plotted in Fig. 4. The LSB method, which is known for its inaudibility, performed the best. CD had satisfactory inaudibility when the embedding capacity was lower than 16 bps. DSS did not satisfy the criteria for either LSD or PESQ. The FE and RPCA&FM methods satisfied the criteria for both LSD and PESQ, and their LSD results were better than those of the proposed method. Our method was better than CD, DSS, FE, and RPCA&FM for PESQ, and its results were close to those of the LSB method for PESQ. Overall, our method had satisfactory inaudibility (LSD ≤ 1.0 dB and PESQ ≥ 3.0 ODG).

Robustness was evaluated against several speech processings and two typical speech codecs. These included re-sampling (24 kHz and 12 kHz), re-quantization (24 bits and 8 bits), speech analysis/synthesis using gammatone filter-bank (GTFB) and short-time Fourier transform (STFT), and speech codecs G.711 and G.726. The BDR results are plotted in Fig. 5. The DSS method performed the best. The LSB method was only robust against a few kinds of processing. The CD method was robust against all processings except for re-quantization with 8 bits, GTFB, and G.726. Our method, the FE method, and the RPCA&FM method were basically robust against all processings except for re-quantization with 8 bits. Overall, our method had satisfactory inaudibility and robustness. However, its robustness at high capacities needs to be improved. This will be a topic of our future work.

4. CONCLUSIONS

We described a watermarking method for speech signals based on echo-hiding and sparse subspace clustering. In this method, watermarks are embedded as echoes of the sparse component for robust extraction. Two independent echo kernels with similar delay times but opposite amplitudes are used to reduce the sound distortion. The evaluation results suggested that echo-hiding with satisfactory inaudibility can be performed on speech. Furthermore, the results showed that it is possible to extract the watermarks with a general cepstrum analysis by taking advantage of the attributes of subsignals when the echo kernels are not applied to the original signal. This finding shows promise for developing new ways of echo-hiding.

5. REFERENCES

[1] Reza Kazemi, Fernando Pérez-González, Mohammad Ali Akhaee, and Fereydoon Behnia, "Data hiding robust to mobile communication vocoders," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2345–2357, 2016.

[2] Hwai-Tsu Hu, Shiow-Jyu Lin, and Ling-Yuan Hsu, "Effective blind speech watermarking via adaptive mean modulation and package synchronization in DWT domain," *EURASIP J. Audio, Speech and Music Processing*, vol. 2017, pp. 10, 2017.

[3] Bin Yan and Yinjing Guo, "Speech authentication by semi-fragile speech watermarking utilizing analysis by synthesis and spectral distortion optimization," *Multimedia Tools Appl.*, vol. 67, no. 2, pp. 383–405, 2013.

[4] Guang Hua, Jiwu Huang, Yun Q. Shi, Jonathan Goh, and Vrizlynn L. L. Thing, "Twenty years of digital audio watermarking - a comprehensive review," *Signal Processing*, vol. 128, pp. 222–242, 2016.

[5] P. Bassia, Ioannis Pitas, and Nikos Nikolaidis, "Robust audio watermarking in the time domain," *IEEE Trans. Multimedia*, vol. 3, no. 2, pp. 232–241, 2001.

[6] Laurence Boney, Ahmed H. Tewfik, and Khaled N. Hamdy, "Digital watermarks for audio signals," in *8th European Signal Processing Conference, EUSIPCO 1996, Trieste, Italy, 10-13 September*, 1996, pp. 1–4.

[7] Yong Xiang, Iynkaran Natgunanathan, Dezhong Peng, Guang Hua, and Bo Liu, "Spread spectrum audio watermarking using multiple orthogonal PN sequences and variable embedding strengths and polarities," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 26, no. 3, pp. 529–539, 2018.

[8] Masashi Unoki and Ryota Miyauchi, "Detection of tampering in speech signals with inaudible watermarking technique," in *Proceedings of Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP*, Piraeus-Athens, Greece, 2012, pp. 118–121.

[9] Nhut Minh Ngo and Masashi Unoki, "Robust and reliable audio watermarking based on phase coding," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2015, Australia, April 19-24*, 2015, pp. 345–349.

[10] Daniel Gruhl, Anthony Lu, and Walter Bender, "Echo hiding," in *Information Hiding, First International Workshop, Cambridge, UK, May 30 - June 1, 1996, Proceedings*, 1996, pp. 293–315.

[11] Byeong-Seob Ko, Ryouichi Nishimura, and Yôiti Suzuki, "Time-spread echo method for digital audio watermarking," *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 212–221, 2005.

[12] Oscal T.-C. Chen and Wen-Chih Wu, "Highly robust, secure, and perceptual-quality echo hiding scheme," *IEEE Trans. Audio, Speech & Language Processing*, vol. 16, no. 3, pp. 629–638, 2008.

[13] Yong Xiang, Dezhong Peng, Iynkaran Natgunanathan, and Wanlei Zhou, "Effective pseudonoise sequence and decoding function for imperceptibility and robustness enhancement in time-spread echo-based audio watermarking," *IEEE Trans. Multimedia*, vol. 13, no. 1, pp. 2–13, 2011.

[14] Guang Hua, Jonathan Goh, and Vrizlynn L. L. Thing, "Time-spread echo-based audio watermarking with optimized imperceptibility and robustness," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 23, no. 2, pp. 227–239, 2015.

[15] Yong Xiang, Iynkaran Natgunanathan, Dezhong Peng, Wanlei Zhou, and Shui Yu, "A dual-channel time-spread echo method for audio watermarking," *IEEE Trans. Information Forensics and Security*, vol. 7, no. 2, pp. 383–392, 2012.

[16] Guang Hua, Jonathan Goh, and Vrizlynn L. L. Thing, "Cepstral analysis for the application of echo-based audio watermark detection," *IEEE Trans. Information Forensics and Security*, vol. 10, no. 9, pp. 1850–1861, 2015.

[17] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Japan, March 25-30*, 2012, pp. 57–60.

[18] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis," *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[19] Ali Zare, Alp Ozdemir, Mark A. Iwen, and Selin Aviyente, "Extension of PCA to higher order data structures: An introduction to tensors, tensor decompositions, and tensor PCA," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1341–1358, 2018.

[20] Chun-Guang Li and René Vidal, "A structured sparse plus structured low-rank framework for subspace clustering and completion," *IEEE Trans. Signal Processing*, vol. 64, no. 24, pp. 6557–6570, 2016.

[21] Ehsan Elhamifar and René Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[22] Chun-Guang Li, Chong You, and René Vidal, "Structured sparse subspace clustering: A joint affinity learning and subspace clustering framework," *IEEE Trans. Image Processing*, vol. 26, no. 6, pp. 2988–3001, 2017.

[23] Hyoung Joong Kim and Yong Hee Choi, "A novel echo-hiding scheme with backward and forward kernels," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 13, no. 8, pp. 885–889, 2003.

[24] Eugen Hoffmann, Dorothea Kolossa, Bert-Uwe Köhler, and Reinhold Orglmeister, "Using information theoretic distance measures for solving the permutation problem of blind source separation of speech signals," *EURASIP J. Audio, Speech and Music Processing*, vol. 2012, pp. 14, 2012.

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

[26] K. Takeda, *Speech database user's manual*, 2010, ATR Technical Report TR-I-0028.

[27] Shengbei Wang and Masashi Unoki, "Speech watermarking method based on formant tuning," *IEICE Transactions*, vol. 98-D, no. 1, pp. 29–37, 2015.

[28] Shengbei Wang, Weitao Yuan, Jianming Wang, and Masashi Unoki, "Speech watermarking based on robust principal component analysis and formant manipulations," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2018, Canada, April 15-20*, 2018, pp. 2082–2086.