Inaudible speech watermarking based on self-compensated echo-hiding and sparse subspace clustering

Shengbei Wang, Weitao Yuan, Jianming Wang, Tianjin Polytechnic University

Masashi Unoki Japan Advanced Institute of Science and Technology, JAIST

Contact Information:
School of Computer Science and Technolody
Tianjin Polytechnic University,
Tianjin, China

Phone: +86 135 0206 8981
Email: wangshengbei@tjpu.edu.cn

Abstract

We proposed an echo-hiding based speech watermarking. Speech signal is analyzed with Sparse subspace clustering (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 echo kernels are designed to have similar delay offsets but opposite amplitudes. As a result, the sound distortion caused by one echo signal can be quickly compensated by the other echo signal, which enables better inaudibility. 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.

keywords:
Echo-hiding, sparse subspace clustering, speech watermarking

Introduction

- **Speech watermarking** is a practical way to protect speech and has been studied for a few decades;
- **An effective watermarking** should satisfy several conflicting requirements, e.g., inaudibility, blindness, robustness, and security;
- **Echo-hiding a challenging task** for speech signals, since the human auditory system is more sensitive to echoes of clean speech than to echoes of general audio;
- **A commom embedding limitation** for echo-hinding is in most cases, the echo kernels can only be applied to the whole signal to realize a cepstrum based watermarking extraction.

Two issues

1. **How to** embed the echo effectively for speech watermarking without degrading the speech quality;
2. **How to** extract the watermarks when the echo kernels are not directly applied to the original whole speech;

Proposed Methods

Feasiblity

Power of speech concentrates on formants. Consequently, the spectrogram about speech has a relatively sparse structure and a speech signal can be separated into a sparse component and a low-rank component.

Sparse subspace clustering for speech separation

High-dimensional data usually can be categorized into several classes and represented by their corresponding low-dimensional subspaces, which can be solved by **Sparse subspace clustering (SSC) [1]**.

1. Given a speech frame, $x(n) \in \mathbb{R}_{n\times 1}$ of n samples ($\sqrt{n}$ is an integer), the $x(n)$ is reshaped into a square matrix $\boldsymbol{X}_F \in \mathbb{R}_{N\times N}$, $N=\sqrt{n}$.
2. Suppose the data points of one column, $\boldsymbol{x}_i \in \mathbb{R}_{N\times 1}$, $1 \leq i \leq N$, of $\boldsymbol{X}_F$ lie in K linear subspaces. 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 a self-expressive dictionary, and the $c_{ii} = 0$ avoids expressing a data point with itself.
3. For Eq. (1), there ideally exists an efficient subspace-sparse representation , $\hat{\boldsymbol{c}}_i$. 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.,
$$\min_{\boldsymbol{c}_i} \|\boldsymbol{c}_i\|_{l_1} \quad \text{s.t.} \quad \boldsymbol{x}_i = \boldsymbol{X}_F \boldsymbol{c}_i, \quad c_{ii} = 0, \tag{2}$$
$$\min_{\boldsymbol{C}} \|\boldsymbol{C}\|_{l_1} \quad \text{s.t.} \quad \boldsymbol{X}_F = \boldsymbol{X}_F \boldsymbol{C}, \quad \mathrm{diag}(\boldsymbol{C}) = 0, \tag{3}$$
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 $\boldsymbol{x}_i$.
4. For speech contains both sparse and low-rank components, $\boldsymbol{X}_F = \boldsymbol{X}_F \boldsymbol{C}$, $\mathrm{diag}(\boldsymbol{C}) = 0$ in Eq. (3) should be generalized as,
$$\boldsymbol{X}_F = \boldsymbol{X}_F \boldsymbol{C} + \boldsymbol{S}, \quad \mathrm{diag}(\boldsymbol{C}) = 0, \tag{4}$$
where $\boldsymbol{S}$ corresponds to the matrix of sparse outlying entries. Accordingly, we have
$$\min_{\boldsymbol{C},\boldsymbol{S}} \quad \|\boldsymbol{C}\|_{l_1} + \lambda_s \|\boldsymbol{S}\|_{l_1} \tag{5}$$
$$\text{s.t.} \quad \boldsymbol{X}_F = \boldsymbol{X}_F \boldsymbol{C} + \boldsymbol{S}, \quad \mathrm{diag}(\boldsymbol{C}) = 0,$$
where $\lambda_s > 0$ balances $\boldsymbol{C}$ and $\boldsymbol{S}$ and l_1-norm promotes sparsity in the columns of $\boldsymbol{C}$ and $\boldsymbol{S}$. The optimal $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{S}}$ express $\boldsymbol{X}_F$ with $\boldsymbol{L}_F \in \mathbb{R}_{N\times N}$ (low-rank, $\boldsymbol{L}_F = \boldsymbol{X}_F \hat{\boldsymbol{C}}$) and $\boldsymbol{S}_F \in \mathbb{R}_{N\times N}$ (sparse, equals $\hat{\boldsymbol{S}}$ and $\boldsymbol{S}_F = \boldsymbol{X}_F - \boldsymbol{L}_F$). The $\boldsymbol{L}_F$ and $\boldsymbol{S}_F$ are reshaped into low-rank signal $l(n) \in \mathbb{R}_{n\times 1}$ and sparse signal $s(n) \in \mathbb{R}_{n\times 1}$.

Watermark embedding algorithm

Self-compensated echo kernels consisting of $h_p(n)$ and $h_q(n)$:
$$h_p(n) = a\delta(n-d_*) + a\delta(n+d_*), \tag{6}$$
$$h_q(n) = -a\delta(n-d_*-\Delta) - a\delta(n+d_*+\Delta), \tag{7}$$

Advantage Opposite amplitudes and small Δ: sound distortion introduced by the first echo is quickly weakened by the second echo;
Performing $h_p(n)$ and $h_q(n)$ on $l(n)$ and $s(n)$ separately, i.e.,
$$\tilde{l}(n) = l(n) + \xi(s(n) \otimes h_p(n)), \tag{8}$$
$$\tilde{s}(n) = s(n) + \xi(s(n) \otimes h_q(n)). \tag{9}$$
$$y(n) = \tilde{l}(n) + \tilde{s}(n) \tag{10}$$
$$= x(n) + \xi(s(n) \otimes (h_p(n) + h_q(n))).$$

Watermark extraction algorithm

- General case: $y(n) = x(n) \otimes h(n) \rightarrow \mathcal{C}_{y(n)} = \mathcal{C}_{x(n)} + \mathcal{C}_{h(n)}$.
- If $y(n) = x(n) + F(x(n)) \otimes h(n)$ ($F(\cdot)$ is the non-linear transformation), then $\mathcal{C}_{y(n)} \neq \mathcal{C}_{x(n)} + \mathcal{C}_{h(n)}$ [2].

In our method, the echoes of $s(n)$ have the same sparsity as $s(n)$. As a result, the echoes will be completely assigned to the sparse component,
$$\check{l}(n) \approx l(n), \tag{11}$$
$$\check{s}(n) \approx s(n) + \xi(s(n) \otimes (h_p(n) + h_q(n))), \tag{12}$$
By re-writing $s(n)$ in form of $s(n) \otimes \delta(n)$,
$$\check{s}(n) \approx s(n) \otimes \underbrace{(\delta(n) + \xi(h_p(n) + h_q(n)))}_{h_s(n)}, \tag{13}$$
$$\mathcal{C}_{\check{s}(n)} \approx \mathcal{C}_{s(n)} + \mathcal{C}_{h_s(n)} \tag{14}$$

The cepstrum of $h_s(n)$ can be expressed as
$$\mathcal{C}_{h_s(n)} = a\xi[\delta(n-d_*) + \delta(n+d_*)] - a\xi[\delta(n-d_*-\Delta) + \delta(n+d_*+\Delta)] + \cdots. \tag{15}$$
The most dominant peaks appear at $n = d_*$ and $n = d_* + \Delta$ can be used for watermark extraction.

Evaluations

- **Dataset**: ATR database (B set) (8.1-sec, 20 kHz, and 16 bits) ;
- **Parameter setting**: λ_s=50, a=0.45, ξ=0.5, d_0=31, d_1=60;
- **Inaudibility**: Log-spectrum distortion (LSD) and Perceptual evaluation of speech quality (PESQ);
- **Robustness**: Bit detection rate (BDR);

Results

Inaudibility affected by offset Δ**:**

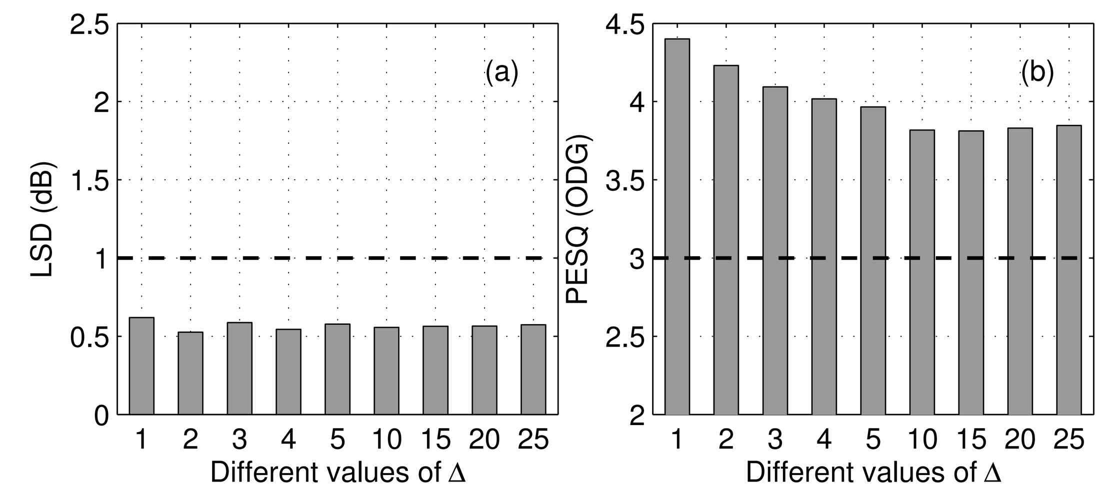

Figure 1: A shorter offset enables two opposite echoes to be better compensated.

***Effectiveness of self-compensated echoes*:**

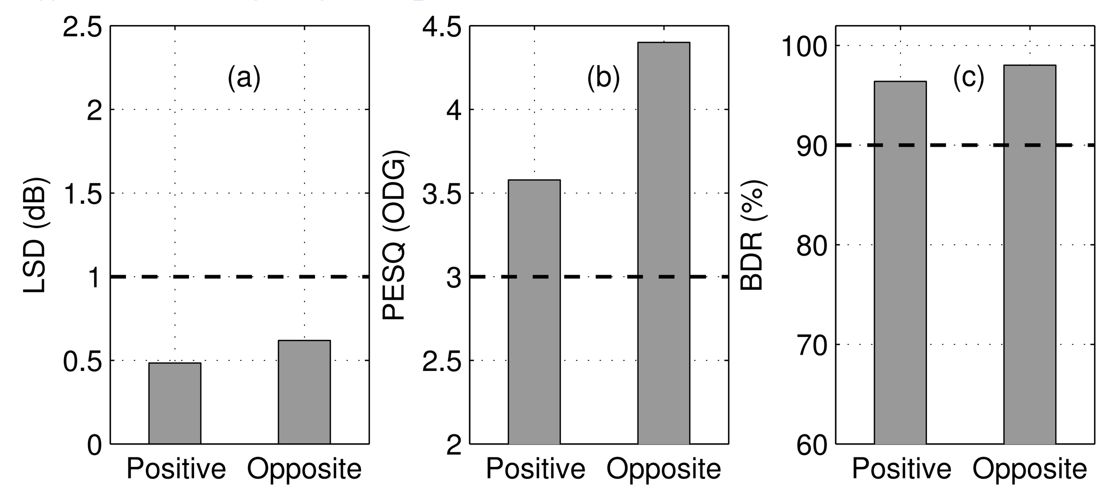

Figure 2: Performance of proposed method using positive and opposite echo kernels.

***Watermark extraction based on sparsity of embedded echoes*:**

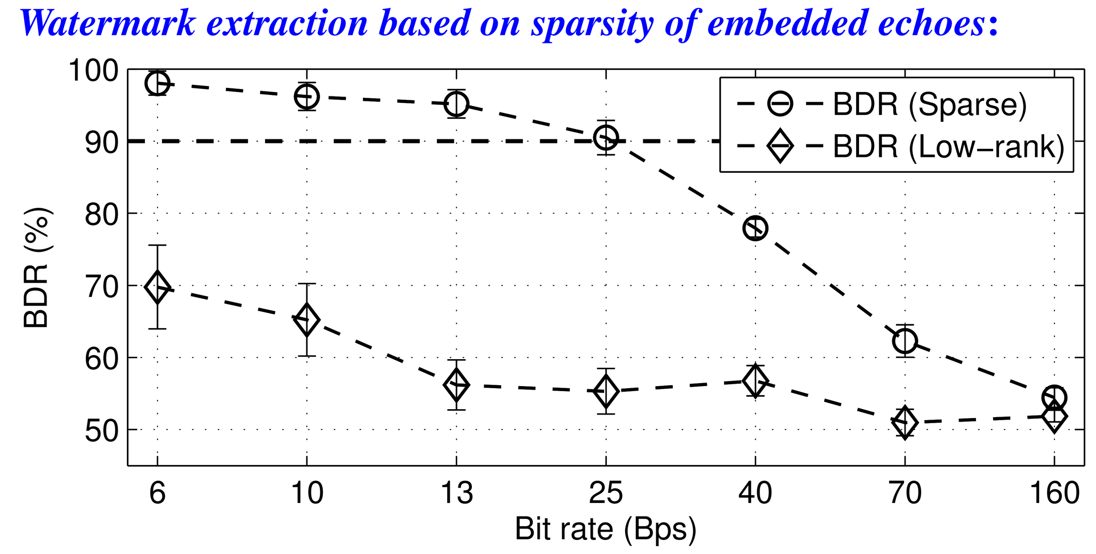

Figure 3: Watermark extraction based on sparsity of embedded echoes.

***Comparative evaluations*:**

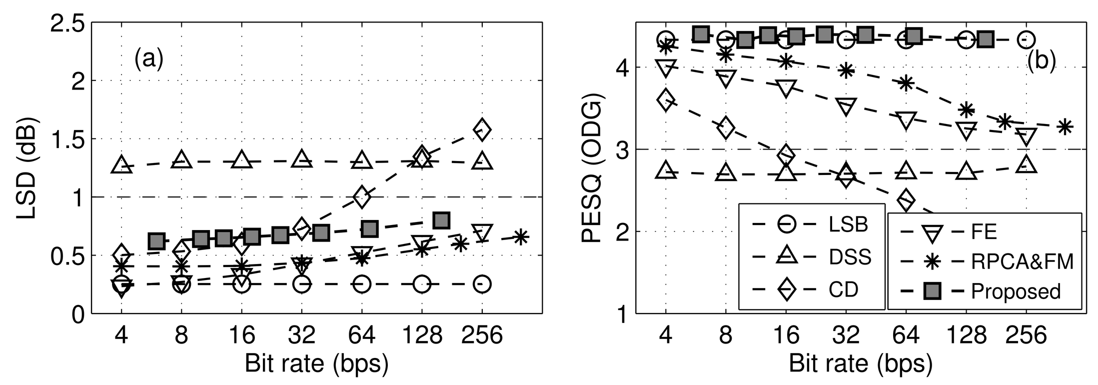

Figure 4: Comparative results on inaudibility.

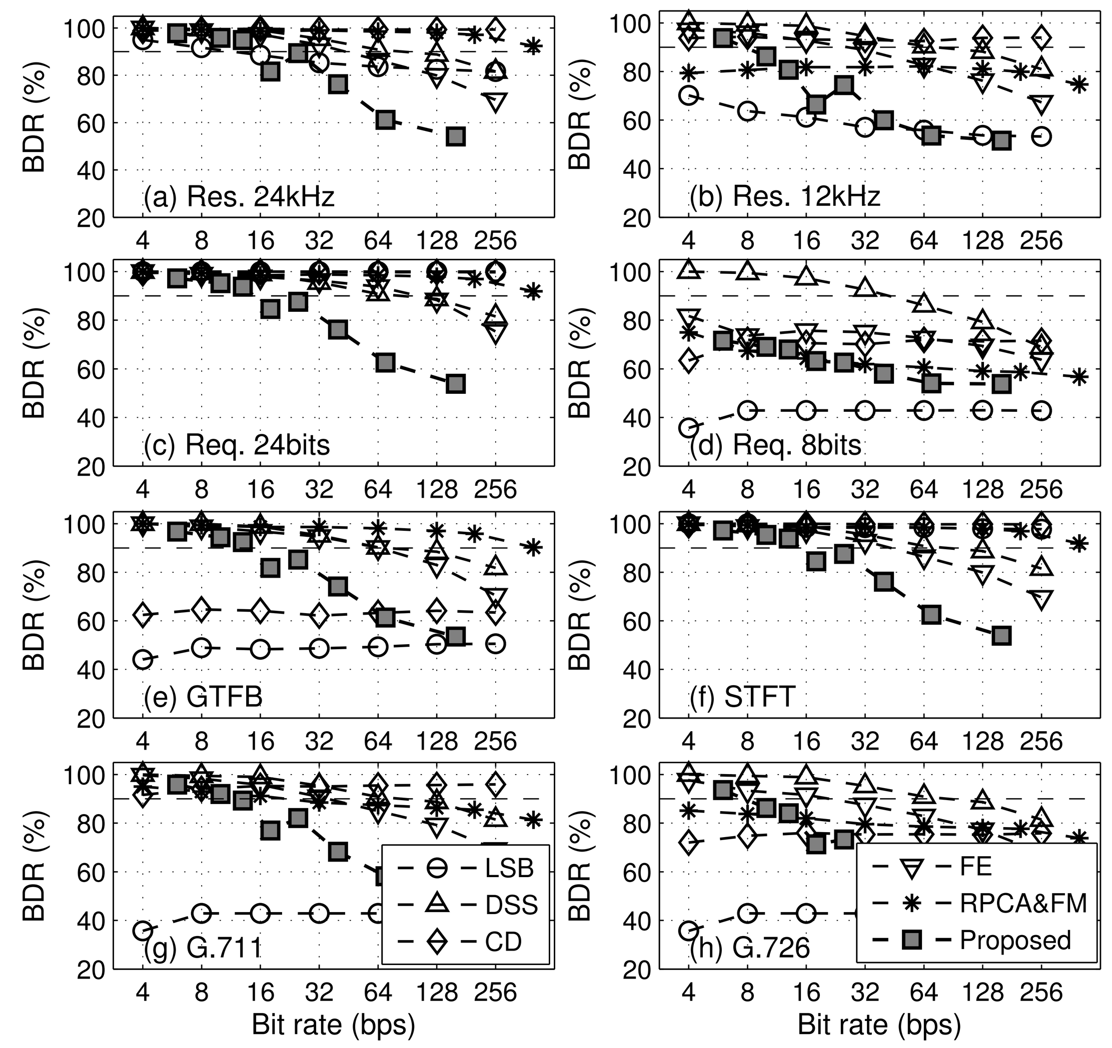

Figure 5: Comparative results on robustness.

Conclusions

We have introduced a watermarking method for speech signals based on echo-hiding and sparse subspace clustering. **Two independent echo kernels with similar delay times but opposite amplitudes are used to reduce the sound distortion.** The evaluation results suggested that **it is possible to extract the watermarks with a general cepstrum analysis by taking advantage of the attributes of subsignals.** This finding shows promise for developing new ways of echo-hiding.

References

[1] 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.

[2] 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.