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

Shengbei Wang, Weitao Yuan, Jianming Wang, Masashi Unoki
Tianjin Polytechnic University
Japan Advanced Institute of Science and Technology, JAIST

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 in 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 caper 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 common embedding limitation for echo-hiding 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

Feasibility

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 separation

High-dimensional data can usually 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, \( x(n) \in \mathbb{R}^N \) of \( N \) samples (\( f \) in an integer), the \( x(n) \) is reshaped into a square matrix \( X \in \mathbb{R}^{N \times N} \).
2. Suppose the data points of one column, \( x_i \in \mathbb{R}^{N \times 1} \), \( 1 \leq i \leq N \), of \( X \) lie in \( K \) linear subspaces. According to the self-expressiveness property, \( x_i \) can be written as a linear combination of the other points in \( X \), i.e.,

\[
x_i = X x_c, \quad c_i = 0, \quad \text{or} \quad c_i = 1.
\]

where \( c_i = (c_{i1}, c_{i2}, \ldots, c_{iN})^T \) and \( X \) is a self-expressive dictionary, and the \( c_i = 0 \) avoiding expressing a data point with itself.

3. For Eq. (1), there ideally exists an efficient subspace-sparse representation \( c_i \). To find this \( c_i \), Eq. (1) is restricted by minimizing the objective function \( c_i \) under the \( l_1 \)-norm, i.e.,

\[
\min_{c_i} ||c_i||_{1}, \quad \text{s.t.} \quad x_i = X x_c, \quad c_i = 0.
\]

The cepstrum of \( b(n) \) can be expressed as

\[
\hat{C}_y(n) = a[n] - d(n) + \hat{d}(n),
\]

\[
\hat{d}(n) = -a[n-d(n)-\Delta] + \hat{d}(n-d(n)+\Delta) + \cdots.
\]

The most dominant peaks appear at \( n - d(n) \) and \( n - d(n) + \Delta \) can be used for watermark extraction.

Evaluations

• Dataset: ATR database (B set) (8.3-sec, 2kHz) and (16 bits);
• Parameter setting: \( \alpha = 30, \mu = 0.45, \zeta = 0.5, \delta_1 = 11, \delta_2 = 60, \); 6 dB.
• Inaudibility: Log-spectrum distortion (LSD) and Perceptual evaluation of speech quality (PESQ);
• Robustness: Bit detection rate (BDR);

Results

Inaudibility affected by offset \( \Delta \)

Effectiveness of self-compensated echoes

Watermark extraction based on sparsity of embedded echoes

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 can be 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