Antiforensics of Speech Resampling Using Dual-Path Strategy

Diqun Yan, Yongkang Gong, and Tianyun Liu

Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China

Correspondence should be addressed to Diqun Yan; yandiqun@nbu.edu.cn

Received 8 December 2020; Revised 27 December 2020; Accepted 31 December 2020; Published 15 January 2021

Copyright © 2021 Diqun Yan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Resampling is an operation to convert a digital speech from a given sampling rate to a different one. It can be used to interface two systems with different sampling rates. Unfortunately, resampling may also be intentionally utilized as a postoperation to remove the manipulated artifacts left by pitch shifting, splicing, etc. To detect the resampling, some forensic detectors have been proposed. Little consideration, however, has been given to the security of these resampling detectors themselves. To expose weaknesses of these resampling detectors and hide the resampling artifacts, a dual-path resampling antiforensic framework is proposed in this paper. In the proposed framework, 1D median filtering is utilized to destroy the linear correlation between the adjacent speech samples introduced by resampling on low-frequency component. And for high-frequency component, Gaussian white noise perturbation (GWNP) is adopted to destroy the periodic resampling traces. The experimental results show that the proposed method successfully deceives the existing resampling forensic algorithms while keeping good perceptual quality of the resampled speech.

1. Introduction

With the wide availability of powerful audio-editing tools such as Adobe Audition, Audacity, and GoldWave, one can easily modify a digital speech with little or no obvious perceptual artifacts. For law enforcement agencies, the authenticity and integrity [1, 2] are of importance when a speech is provided as evidence. Therefore, various speech forensic techniques have been proposed to identify different kinds of forgery, such as replaying [3, 4], pitch shifting [5, 6], and double compression [7]. Resampling is an operation widely used to convert a digital speech from a given sampling rate to a different sampling rate. However, it is worth noting that resampling is also a necessary operation when a speech undergoes other manipulations like pitch shifting, splicing, and fake-quality mp3 compositing [8]. Additionally, resampling could be used as a postprocessing operation to hide the artifacts left by the forgery operations [9].

In recent years, many forensic techniques have been proposed to detect the traces left in resampled speech. Since specific periodic correlations between the samples of the resampled speech will be introduced, Expectation Maximization (EM) algorithm is applied to estimate such periodic artifacts [10]. In addition, the statistical moments in frequency domain are chosen as features to classify the original and the resampled speech. Inspired by the successful application of derivative features in image resampling forensics [11] and audio double compression forensics [12], Xu and Xia [13] found that if an original audio signal is resampled, significant peaks can be found in the second-order derivative of the spectrum, and the peak position is related to resampling factor. They utilized K-singular value decomposition to distinguish the resampled speech from the natural one. In our previous work [14], we found that resampling will cause the inconsistency between the bandwidth and sampling rate. It proves that the logarithmic ratio of band energy can effectively detect the resampling trace.

In real scenario, a malicious attacker may launch an antiforensic attack against the existing forensic detectors. The purpose of the antiforensic techniques is to fool or mislead the forensic detectors by creating difficulties in forgery detection. In fact, the study of antiforensics can help researchers be aware of weaknesses in the existing forensic detectors. In this work, we focus on antiforensics of speech resampling. The objective of this work is to remove the resampling trace without degrading the perceptual quality of the resampled speech. To the best of our knowledge, there is no prior work on antiforensic of speech resampling. Specifically, a dual-path
strategy is applied in the proposed algorithm. For the low-frequency component of the speech, median filter is used to destroy the linear relationship between the adjacent sampling in the resampled speech. And for the high-frequency component, Gaussian perturbation is adaptively added with Gaussian white noise. The experimental results show that the proposed antiforensic algorithm can not only remove the resampling trace but also ensure the speech quality after the antiforensic process.

The rest of this paper is organized as follows. Section 2 briefly reviews the basics of speech resampling and the related work on resampling forensics. In Section 3, a dual-path antiforensic framework is given and then the proposed antiforensic resampling algorithm is introduced. The experimental results are presented in Section 4. Finally, Section 5 summarizes the paper and discusses future work.

2. Related Works

In this section, the basics of speech resampling are given and two typical forensic algorithms for speech resampling are investigated.

2.1. Speech Resampling. Resampling, also known as sample rate conversion, is the process of converting a signal from one sampling rate to another. The resampling function $R$ can be defined by

$$Y = R(X, r),$$

where $X$ is the original speech, $Y$ is the resampled speech, and $r$ is the resampling factor. $r$ can be defined as

$$r = \frac{P}{Q},$$

where $P$ and $Q$ are the original and new sampling rates after the conversion, respectively. According to the value of $r$, it can be divided into downsampling $r < 1$ and upsampling $r > 1$.

The implementing of the resampling scheme is shown in Figure 1. The original speech $X$ is first upsampled by a factor of $P$ and followed by interpolation filtering. Then, the interpolated speech is downsampled by a factor of $Q$. Here, the role of the filter is typically a lowpass filter to act as an antialiasing filter; the stopband frequency $f_{\text{stop}}$ of the ideal filter $H(\omega)$ should be

$$f_{\text{stop}} = \min \left( \frac{\pi}{P}, \frac{\pi}{Q} \right),$$

where min denotes the minimum function. More details of resampling can be found in [11].

2.2. Existing Works on Detecting Resampling Trace. Popescu and Farid [11] found that the linear interpolated signal will exhibit a periodic correlation. Such correlation could be well fitted by the EM algorithm and successfully applied to the detection of image resampling. To make the EM algorithm more suitable for speech resampling detection, Chen and Wu [10] introduced signal histogram to describe the speech’s distribution and extracted the statistical moments in frequency domain as the identification features. Let $f_{np}(i)$, $i = 1, 2, 3, \cdots, K$, be the value of the $i$th spectrum peak, and $K$ is the total number of spectrum peaks. The detection algorithm can be described as follows.

1. Normalize $f_{np}(i)$ to $[0, 1]$

$$f_{np}(j) = \frac{f_{np}(i)}{\max(f_{np}(i))}. \quad (4)$$

where $j$ is defined by equation (5) and its purpose is to make the spectrum peaks consistent for the speech with various lengths, so as to adapt the situation that the speech is often not in a fixed length in real scenes.

$$j = \frac{\text{abs}(i - (K/2))}{K/2}. \quad (5)$$

2. In order to highlight the primary peaks, the small values of $f_{np}(j)$ are suppressed by

$$f_{np}(j) = 10^{-6}, \text{if } f_{np}(j) < 5 \times \text{mean}(f_{np}(j)), \quad (6)$$

where mean($\cdot$) denotes the average function.

3. Calculate the 3rd-order central moment of $f_{np}(j)$ and normalize it by

$$S_q = \frac{1}{K} \sum(f_{np}(j))^q, \quad q = 1, 2, 3. \quad (7)$$

4. Once $S_q$ is obtained, it will be compared with a threshold to determine whether the suspected speech is resampled or not.

Xu and Xia [13] found that significant peaks will arise in the second-order derivative of the spectrum in a resampled speech. Meanwhile, the peak position is related to the resampling factor. For a given speech, we can obtain the FFT spectrum on its second-order derivatives. According to the sum of the mean $m$ and standard variance of the spectrum $s$, the flat critical boundary can be estimated. Meanwhile, the maximum value of the spectrum peak $p_{\text{max}}$ can be found by searching the whole normalized frequency domain. Under a hypothesis framework, the decision rule for distinguishing resampled and original speech can be
Figure 2: Proposed dual-path antiforensic scheme.

Figure 3: Accuracy of various detection algorithms after antiforensic attack: (a) EM algorithm; (b) second-order derivative algorithm.
formulated as follows:

\[
H_0 : \text{Test speech is not resampled,} \\
H_1 : \text{Test speech is resampled,}
\]

\[
\text{decision} = \begin{cases} 
H_0 & : \text{if } \frac{p_{\text{max}}}{m+s} < Th, \\
H_1 & : \text{if } \frac{p_{\text{max}}}{m+s} \geq Th,
\end{cases}
\]

where Th is a decision threshold.

3. Proposed Antiforensic Method

Generally, it is a hard task to defeat the resampling forensic detectors by a single-step attack [15]. Hence, in this work, a dual-path framework is designed to remove the resampling trace, which has been shown in Figure 2. In this framework, different strategies are applied to the different frequency-domain components of the resampled speech. For the low-frequency component, 1D median filtering is utilized to destroy the linear correlation between the adjacent samples introduced by the resampling. For the high-frequency component, perturbation is created by adding Gaussian white noise on the samples in order to destroy the periodic resampling traces.

Given the speech \(X\) and the resampling factor \(r\), the speech \(X\) can be modelled as

\[
X = X^L + X^H,
\]

where \(X^L\) and \(X^H\) are the low- and high-frequency components, respectively.

The proposed antiforensic algorithm can be described as follows.

1. First, these two components are separated by 1D median filter

\[
X^L = \text{median}(X, \alpha), \\
X^H = X - X^L,
\]

where \(\alpha\) is the size of the median filter.
Figure 5: Continued.
Here, 1D median filter works with a sliding window that passes through each speech sample one by one. At each sample, numerically sort the list of adjacent neighbor samples and the middle one from the list is just the output of the median filter.

(2) For $X^H$, a zero mean Gaussian white noise $n \sim N(0, \sigma)$ is added and the strength of the noise is controlled by the standard variance $\sigma$

$$X^H_n = X^H + n.$$  \hfill (11)

(3) Then, the noised signal is resampled with the specific resampling factor $r$

$$Y^H = R(X^H_n, r).$$  \hfill (12)

(4) Meanwhile, $X$ is resampled with $r$ to get $\hat{X} = R(X, r)$. Then, the low-frequency component $Y^L$ can be obtained by applying median filtering on $\hat{X}$

$$Y^L = \text{median} (\hat{X}, \alpha).$$  \hfill (13)

(5) Finally, the resampled speech is obtained by

$$Y = Y^L + Y^H.$$  \hfill (14)

4. Experimental Results

4.1. Experimental Setup. To evaluate the antiforensic effectiveness of the proposed algorithm quantitatively, we perform experiments to attack two typical resampling detectors [10, 13]. To this end, TIMIT, which is a dataset, consisted of 6300 speech with the average duration of 3 seconds from 630 North America speakers. The format of each speech is WAV, 16kHz sampling rate, 16-bit quantization, and mono. TIMIT is taken as the original speech dataset. And the resampling dataset is created by resampling function resample in Matlab from 0.8 to 2.0 (step is 0.1). Two different window sizes $\alpha$ (3 and 5) of median filter are chosen, and the strength of the Gaussian noise $\sigma$ is set as 0.2.

4.2. Removing Resampling Artifacts. From the description in Section 2.1, interpolation algorithm is the key component in speech resampling. Meanwhile, it will cause unavoidable linear dependencies between adjacent audio samples. As shown in [16, 17], the strength of such linear dependencies varies periodically with the resampling parameters. Other detection methods [9, 13] also prove the existence of such periodic artifacts.

The detection accuracy of the forensic methods is defined as the percentage of true positive and true negative in all testing cases. Figure 3 shows the detection accuracies in [10, 13]. First, when the resampling factor $r$ is less than 1.0, the accuracy of the EM algorithm is reduced by about 10%, while the second-order derivatives are reduced by about 2%. With the increase of the resampling factor, when the resampling factor is greater than 1.2, the detection performance of two methods is reduced by an average of 30%. Meanwhile, if we decrease the window size of the median filter, the attack performance will be improved further. The above experimental results show that the proposed antiforensic algorithm better erases the periodic resampling traces.

Figure 4 shows the normalized power spectrums of the EM algorithm, and Figure 5 shows the second-order derivative spectrum. We can see that the periodic traces caused by resampling have been effectively removed. It further confirms the effectiveness of the proposed algorithm.
It is known that median filtering will cause regions with nearly constant intensity values [18]. Moreover, median filtering could be taken as an attack against resampling detection due to its nonlinear smoothing ability. Thus, in the proposed method, the periodic dependencies between the neighboring samples are effectively destroyed by the median filtering. On the other hand, the high-frequency component is resampled with adding Gaussian white noise on the samples. By the proposed dual-path attack, the resampling trace can be successfully removed.

4.3. Perceptual Quality. Segmental SNR (SegSNR) [19] and Perceptual Evaluation of Speech Quality (PESQ) [20], which have been extensively used for evaluating the sound quality objectively, are adopted as the speech quality evaluation metrics. SegSNR is a variant of the signal-to-noise ratio. It is a method for determining the signal-to-noise ratio by dividing the speech signal into a short time of 10-30 ms and then taking the average value over the full speech time interval. The definition of SegSNR is

$$\text{SegSNR} = \frac{1}{J} \sum_{j=0}^{J-1} 10 \log \frac{\sum_{n=0}^{k-1} s(n,j)^2}{\sum_{n=0}^{k-1} [s(n,j) - \bar{s}(n,j)]^2},$$

where $J$ is the number of total frames, $j$ is the frame index, and $k$ is the frame size.

Figure 6: Speech quality evaluation: (a) SegSNR; (b) ODG.
PESQ is standardized as ITU-T recommendation P.862 and can provide an end-to-end quality assessment. The Objective Difference Grade (ODG) is a main parameter of PESQ and has a range from 0 to -4. The highest score 0 means imperceptible difference, and the lowest score -4 means very annoying.

Figure 6 shows the SegSNR and ODG scores of the speech processed through the proposed antiforensic algorithm with various filtering window sizes. First, the quality of the resampled speech is improved with the increase of the resampling factor, which meets the actual expectation. Meanwhile, ODG scores for various resampling factors are higher than -2.0, which means that the proposed antiforensic algorithm can maintain good perceptual quality of the resampled speech. Additionally, we can see that the algorithm with the smaller window size provides better quality results. Hence, in practice, a smaller window size should be more suitable for choosing.

5. Conclusion

In this work, a resampling antiforensic method based on dual-path strategy is proposed to attack the existing forensic detectors. The linear correlation introduced by resampling is processed by a median filtering on low-frequency component. And the periodic traces introduced by resampling are removed by adding Gaussian perturbation. A better tradeoff is obtained between speech perceptual quality and forensic undetectability by the proposed resampling antiforensic technique. In the future, we expect that our approach can be extended to antiforensics of other speech operations.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61300055), Ningbo Natural Science Foundation (Grant No. 202003N4089), and K.C. Wong Magna Fund in Ningbo University.

References

[1] Z. L. Zhou, Q. M. Jonathan Wu, Y. M. Yang, and X. M. Sun, "Region-level visual consistency verification for large-scale partial-duplicate image search," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 2, pp. 1–25, 2020.
[2] Z. L. Zhou, Y. Mu, and Q. M. Jonathan Wu, "Coverless image steganography using partial-duplicate image retrieval," *Soft Computing*, vol. 23, no. 13, pp. 4927–4938, 2019.
[3] X. D. Lin, J. X. Liu, and X. G. Kang, "Audio recapture detection with convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1480–1487, 2016.
[4] K. M. Malik, A. Javed, H. Malik, and A. Irtaza, "A light-weight replay detection framework for voice controlled IoT devices," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 982–996, 2020.
[5] H. T. Wu, Y. Wang, and J. W. Huang, "Identification of electronic disguised voices," *IEEE Transactions on Information Security and Forensics*, vol. 9, no. 3, pp. 489–500, 2014.
[6] M. K. Singh, A. K. Singh, and N. Singh, "Multimedia analysis for disguised voice and classification efficiency," *Multimedia Tools and Applications*, vol. 78, pp. 29395–29411, 2019.
[7] D. Luo, R. Yang, and J. W. Huang, "Identification of AMR decompressed audio," *Digital Signal Processing*, vol. 37, pp. 85–91, 2015.
[8] X. M. Yu, R. D. Wang, D. Q. Yan, and P. F. Ma, "Detecting fake-quality MP3 based on Huffman table index," *Journal of Software*, vol. 9, no. 4, pp. 907–912, 2014.
[9] K. Singh, A. Kansal, and G. Singh, "An improved median filtering anti-forensics with better image quality and forensic undetectability," *Multidimensional Systems and Signal Processing*, vol. 30, no. 11, pp. 1951–1974, 2019.
[10] Y. Chen and X. Wu, "An EM-based resampling detection method in audio blind forensics," *Journal of Circuits and Systems, vol. 17*, no. 4, pp. 118–123, 2012.
[11] A. C. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of re-sampling," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 16–26, 2005.
[12] C. Grigoras, "Statistical tools for multimedia forensics: compression effects analysis," in *Proceedings of International Conference on Audio Forensics: Practices and Challenges*, pp. 1–3, Hillerod, Denmark, 2010.
[13] J. Xu and J. Xia, "Digital audio resampling detection based on sparse representation classifier and periodicity of second derivative," *Journal of Digital Information Management*, vol. 13, pp. 101–109, 2015.
[14] Z. F. Wang, D. Q. Yan, and R. D. Wang, "Speech resampling detection based on inconsistency of band energy," *Computers Materials & Continua*, vol. 56, no. 2, pp. 247–259, 2018.
[15] M. Kirchner and R. Bohme, "Hiding traces of resampling in digital images," *IEEE Transactions on Information Security and Forensics*, vol. 3, no. 4, pp. 582–592, 2008.
[16] A. C. Gallagher, "Detection of linear and cubic interpolation in JPEG compressed images," in *The 2nd Canadian Conference on Computer and Robot Vision (CRV’05)*, pp. 65–72, Victoria, BC, Canada, 2005.
[17] B. Mahdian and S. Saic, "Blind authentication using periodic properties of interpolation," *IEEE Transactions on Information Security and Forensics*, vol. 3, no. 3, pp. 529–538, 2008.
[18] D. Kim, H. U. Jang, and S. M. Mun, "Median filtered image restoration and anti-forensics using adversarial networks," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 278–282, 2018.
[19] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings* (Cat. No.01CH37221), pp. 749–752, Salt Lake City, UT, USA, 2001.
[20] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 16, no. 1, pp. 229–238, 2008.