

RESEARCH

Open Access



# Existence detection and embedding rate estimation of blended speech in covert speech communications

Lijuan Li\* and Yong Gao

\*Correspondence:  
LLJ\_SCU@163.com  
College of Electronics  
and Information Engineering,  
Sichuan University,  
Chengdu 610064, Sichuan,  
China

## Abstract

Covert speech communications may be used by terrorists to commit crimes through Internet. Steganalysis aims to detect secret information in covert communications to prevent crimes. Herein, based on the average zero crossing rate of the odd–even difference (AZCR-OED), a steganalysis algorithm for blended speech is proposed; it can detect the existence and estimate the embedding rate of blended speech. First, the odd–even difference (OED) of the speech signal is calculated and divided into frames. The average zero crossing rate (ZCR) is calculated for each OED frame, and the minimum average ZCR and AZCR-OED of the entire speech signal are extracted as features. Then, a support vector machine classifier is used to determine whether the speech signal is blended. Finally, a voice activity detection algorithm is applied to determine the hidden location of the secret speech and estimate the embedding rate. The results demonstrate that without attack, the detection accuracy can reach 80 % or more when the embedding rate is greater than 10 %, and the estimated embedding rate is similar to the real value. And when some attacks occur, it can also reach relatively high detection accuracy. The algorithm has high performance in terms of accuracy, effectiveness and robustness.

**Keywords:** Blended speech, Covert speech communication, Embedding rate estimation, Existence detection, Odd–even difference (OED)

## Background

Covert audio communication refers to the technique of embedding secret information into carrier audio such that important information can be transmitted safely and reliably via public communication. This technology can also be used by terrorists to commit criminal activities, which pose a serious danger to society (Qiao et al. 2013). Traffickers and terrorists have begun to hide secret information in audio files, such as MP3 and WAV files, and spread them by file sharing or e-mail through high-quality Internet (Gelfand 2007).

The speech, as an important branch of the audio, is one of the most important methods for human communication and is ubiquitous and accessible. Currently, existing covert speech communication algorithms (Singh 2016; Tayel et al. 2016; Hartoko et al. 2015; Krishnan and Abdullah 2016; Nutzinger and Juergen 2011; Matsuoka 2006; Byeong-Seob

et al. 2005; Tatsuya and Kotaro 2015; Chen 2001) include the following: least significant bits (LSB) hiding (Tayel et al. 2016; Hartoko et al. 2015; Krishnan and Abdullah 2016), phase coding hiding (Nutzinger and Juergen 2011), direct sequence spread spectrum (DSSS) encoding hiding (Matsuoka 2006), echo hiding (Byeong-Seob et al. 2005; Tatsuya and Kotaro 2015), and blending-based speech hiding (Chen 2001), along with others. Among these algorithms, the blending-based speech hiding algorithm is different from others, where the secret speech can be hidden directly in the carrier speech and does not need to be binary encoded. This algorithm has a good robustness (Rangding et al. 2004) and high hidden capacity.

In contrast with covert speech communication, the aim of speech steganalysis (Natarajan and Nayak 2010; Ghasemzadeh et al. 2016; Bhattacharyya and Sanyal 2012; Chunhui and Yimin 2010; Wei et al. 2008; Hamza et al. 2003) is to detect the existence of secret information in covert speech communication and extract that information. Steganalysis algorithms can be divided into two classes based on their scope of application. One class is specific steganalysis algorithms (Chunhui and Yimin 2010; Wei et al. 2008), and the other is universal steganalysis algorithms (Hamza et al. 2003; Özer et al. 2006; Avcibas 2006). However, no experiment has determined whether existing steganalysis algorithms are applicable to blended speech.

For the above reasons, this paper proposes a steganalysis algorithm for blended speech transmitted via a high-quality Internet. The algorithm is based on the average zero crossing rate (Muhammad 2015; Ali et al. 2011; Ghosal and Suchibrota 2014) of the odd–even difference (AZCR-OED) of the speech, and it combines a support vector machine (SVM) classifier (Mathias and Mohamed 2015; Alex and Bernhard 2004) and a voice activity detection (VAD) algorithm (Shota et al. 2016; Qi et al. 2002; Xinyan and Guojun 2013; Jongseo et al. 1999). The algorithm can detect the existence and estimate the embedding rate of blended speech. The experimental results demonstrate the high accuracy, effectiveness and robustness of this algorithm for a variety of embedding rates.

This paper is organized as follows: “**Blending-based speech hiding algorithm**” section provides a brief introduction to the blending-based speech hiding algorithm. The features, such as the AZCR-OED of the speech, are analyzed in “**Feature analysis**” section. A steganalysis algorithm for blended speech, which is used for existence detection and embedding rate estimation, is proposed in “**Steganalysis algorithm for blending speech**” section. Experimental results and analyses are given in “**Experimental results and analysis**” section. Finally, “**Conclusion**” section presents the conclusions of the work.

### **Blending-based speech hiding algorithm**

$P$  denotes the carrier speech of length  $N$ , and  $S$  denotes the secret speech of length  $M$ . The blending-based speech hiding algorithm is described as follows (Nutzinger and Juergen 2011):

$$\begin{cases} F(2k-1) = P(2k-1) \\ F(2k) = (1-\alpha)P(2k-1) + \alpha S(k) \end{cases} \cdot (1 \leq k \leq M) \quad (1)$$

Considering that the hidden location of secret speech is uncertain, the Eq. (1) can be rewritten as follows:

$$\text{First: } F(k) = P(k), \quad 1 \leq k \leq N \quad (2)$$

$$\text{Then: } \begin{cases} F(2k - 1 + \textit{start}) = P(2k - 1 + \textit{start}) \\ F(2k + \textit{start}) = (1 - \alpha)P(2k - 1 + \textit{start}) + \alpha S(k). \end{cases} \quad (3)$$

where  $1 \leq k \leq M$ , and  $0 \leq \alpha \leq 1$ ,  $\alpha$  is the hidden degree factor; *start* is the location of the start of the secret speech, where  $\textit{start} \in \{0, 1, 2, \dots, N - 2M\}$ ; and  $F$  denotes the blended speech of length  $N$ .  $F$  is more similar to the carrier speech  $P$  as  $\alpha$  decreases. Furthermore, Eq. (3) implies that  $M$  and  $N$  must satisfy  $N \geq 2M$ .

In this paper, we define a stego speech segment as a speech segment in which secret speech is hidden in blended speech. From Eq. (3), we know that the stego segment's length is  $2M$ . The embedding rate is defined as the ratio of the length of the stego speech segment to the entire length of the blended speech and is denoted by  $\eta$ . Thus,

$$\eta = \frac{2M}{N} \times 100\% \quad (4)$$

where  $0 \leq \eta \leq 100\%$ ,  $\eta = 0$  means there is no stego speech segment in the speech signal, and  $\eta = 100\%$  means the carrier speech is completely used to hide the secret speech. Generally, for a fixed length of secret speech, a lower embedding rate requires a longer carrier speech signal, which leads to lower communication efficiency.

Given Eq. (3), the extraction algorithm for secret speech can be defined as follows:

$$\begin{aligned} S(k) &= \frac{F(2k + \textit{start}) - (1 - \alpha)P(2k - 1 + \textit{start})}{\alpha} \\ &= \frac{F(2k + \textit{start}) - (1 - \alpha)F(2k - 1 + \textit{start})}{\alpha} \end{aligned} \quad (5)$$

where  $1 \leq k \leq M$ . From Eq. (5), we can observe that if the receiver wants to extract the secret speech from the blended speech, he/she must know at least the following information:

1. the location of the start of the secret speech;
2. whether the odd–even points are aligned (which is defined below) with the sender;
3. the hidden degree factor  $\alpha$  of the secret speech.

The aims of this paper are as follows: extract the features that can distinguish pure speech from blended speech to detect blended speech, judge whether the odd–even points are aligned with the sender and correct the inverted case, estimate the hidden starting location of the secret speech and the length of the stego speech segment, and calculate the embedding rate.

### Feature analysis

In this section, we first briefly state several definitions that will be used later. Then, the difference in the odd–even difference (OED) between the blended speech and pure carrier speech is analyzed and compared. Finally, we present the features that can distinguish pure and blended speech.

**Definitions**

**Definition 1** For the speech signal  $X(X = \{x(1), x(2), x(3), \dots, x(N)\})$ , the OED is defined to be the difference between the values of odd and even points. Denoting the OED by  $D$ , we have  $D(k) = x(2k) - x(2k - 1)$ , where  $1 \leq k \leq \lfloor N/2 \rfloor$  and  $\lfloor \cdot \rfloor$  denotes the rounded-down value.

**Definition 2** For a sent speech signal  $X(X = \{x(1), x(2), x(3), \dots, x(N)\})$  and received speech signal  $R(R = \{r(1), r(2), r(3), \dots, r(M)\})$ , when  $r(n) = x(n)$  (where  $1 \leq n \leq N$ ), we consider the odd–even points of the received speech to be aligned with the sent speech. When  $r(2k - 1) = x(2k)$  and  $r(2k) = x(2k + 1)$ , where  $1 \leq k \leq \lfloor N/2 \rfloor$ , we define the odd–even points of the received speech to be inverted with respect to the sent speech.

**Definition 3** For the speech signal  $X(X = \{x(1), x(2), x(3), \dots, x(N)\})$ , the average zero crossing rate (ZCR) (Muhammad 2015; Ali et al. 2011; Ghosal and Suchibrota 2014) is defined as follows:

$$Z = \frac{1}{2N} \sum_{k=2}^N |\text{sgn}[x(k)] - \text{sgn}[x(k - 1)]|,$$

where  $\text{sgn}[x(k)] = \begin{cases} 1, & x(k) \geq 0 \\ -1, & x(k) < 0 \end{cases}$

**OED of the speech signal**

$X$  and  $R$  denote the pure sent speech signal and the received speech signal, respectively. Four cases of the OED of received speech are discussed below.

It is assumed that the hidden starting location of the secret speech is  $start = 0$  and that the embedding rate is  $\eta = 100\%$ . For other values, the derivation is similar.

Case 1: the received speech is the blended speech, and the odd–even points are aligned with the sent speech. The OED denoted by  $Drc$  is as follows:

$$\begin{aligned} Drc(k) &= R(2k) - R(2k - 1) \\ &= F(2k) - F(2k - 1) \\ &= (1 - \alpha)P(2k - 1) + \alpha S(k) - P(2k - 1) \\ &= \alpha[S(k) - P(2k - 1)]. \end{aligned} \tag{6}$$

Case 2: the received speech is the pure speech, and the odd–even points are aligned with the sent speech. The OED denoted by  $Dro$  is as follows:

$$Dro(k) = R(2k) - R(2k - 1) = X(2k) - X(2k - 1). \tag{7}$$

Case 3: the received speech is the blended speech, and the odd–even points are inverted with respect to the sent speech. The OED denoted by  $Dwc$  is as follows:

$$\begin{aligned} Dwc(k) &= R(2k) - R(2k - 1) \\ &= F(2k + 1) - F(2k) \\ &= P(2k + 1) - (1 - \alpha)P(2k - 1) + \alpha S(k). \end{aligned} \tag{8}$$

Case 4: the received speech is the pure speech, and the odd–even points are inverted with respect to the sent speech. The OED denoted by  $Dwo$  is as follows:

$$Dwo(k) = R(2k) - R(2k - 1) = X(2k + 1) - X(2k). \tag{9}$$

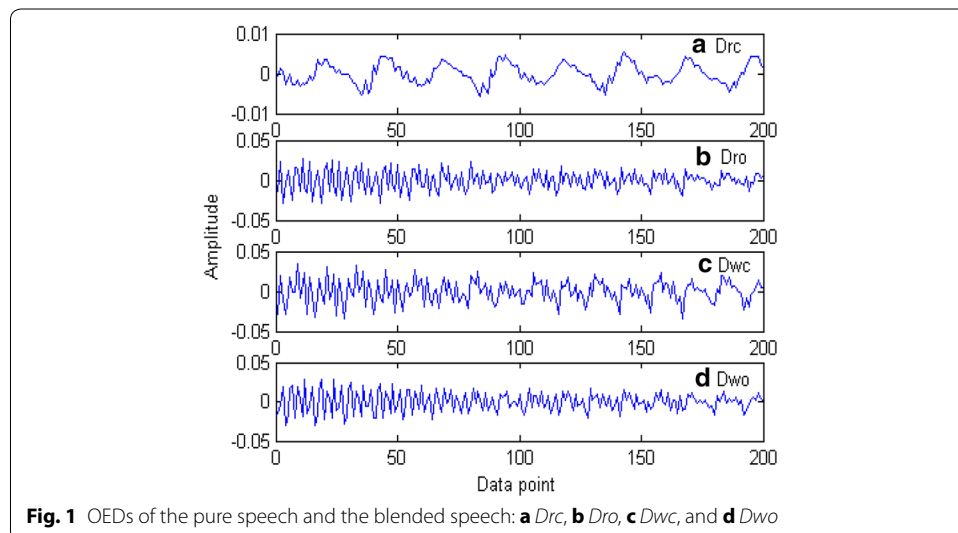
We selected two pure speech samples from the Voice of America (VOA) Special English corpus ([http://www.51voa.com/VOA\\_Special\\_English/](http://www.51voa.com/VOA_Special_English/)) randomly and transformed them into the WAV format with an 8-kHz sampling rate and 16-bit speech encoding. Then, one of the samples was chosen to be the secret speech and hidden in the other using the blending-based speech hiding algorithm, with an embedding rate of 100 % and a hidden degree factor  $\alpha = 0.05$ . Finally, both the pure and blended speech were transmitted through QQ, a popular instant messaging software service in China that is used for chatting, and the receiver analyzed the OED of the received speech. The result is shown in Fig. 1 (to see the result clearly, only data points 1–200 are shown in the figure).

We can intuitively see from Fig. 1 that the change in the OED of the blended speech when the odd–even points are aligned with the sent speech, which is denoted by  $Drc$ , is less rapid than in the other three cases. In this paper, we use the average ZCR to describe this difference.

#### AZCR-OED of the speech signal

According to definition 3 and Eq. (6), the corresponding AZCR-OED of  $Drc$  can be determined as follows:

$$\begin{aligned} Z(Drc) &= \frac{1}{2N} \sum_{k=2}^N |\text{sgn}[Drc(k)] - \text{sgn}[Drc(k - 1)]| \\ &= \frac{1}{2N} \sum_{k=2}^N |\text{sgn}\{\alpha[S(k) - P(2k - 1)]\} - \text{sgn}\{\alpha[S(k - 1) - P(2k - 3)]\}| \\ &= \frac{1}{2N} \sum_{k=2}^N |\text{sgn}[S(k) - P(2k - 1)] - \text{sgn}[S(k - 1) - P(2k - 3)]|. \end{aligned} \tag{10}$$



**Fig. 1** OEDs of the pure speech and the blended speech: **a**  $Drc$ , **b**  $Dro$ , **c**  $Dwc$ , and **d**  $Dwo$

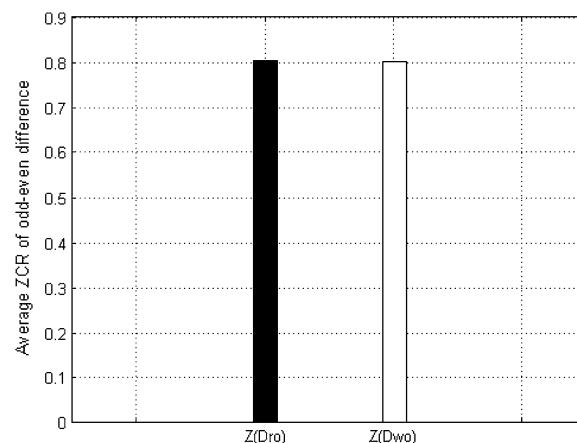
In Eq. (10), we can see that  $Z(Drc)$  is determined only by the value of the secret speech and the carrier speech and is unrelated to the hidden degree factor of the secret speech.

To verify whether the AZCR-OED can serve as feature to distinguish between blended and pure speech, we first obtained 8000 pure speech samples from the VOA Special English corpus ([http://www.51voa.com/VOA\\_Special\\_English/](http://www.51voa.com/VOA_Special_English/)). Then, we transformed all of them into the WAV format with an 8-kHz sampling rate and 16-bit speech encoding and built a speech sample library (which is hereafter referred to as the “VOASE” library). Finally, we divided the VOASE library into two groups, the secret speech group and the carrier speech group. We performed a statistical analysis of the AZCR-OED of the blended speech and pure speech through the following experiments.

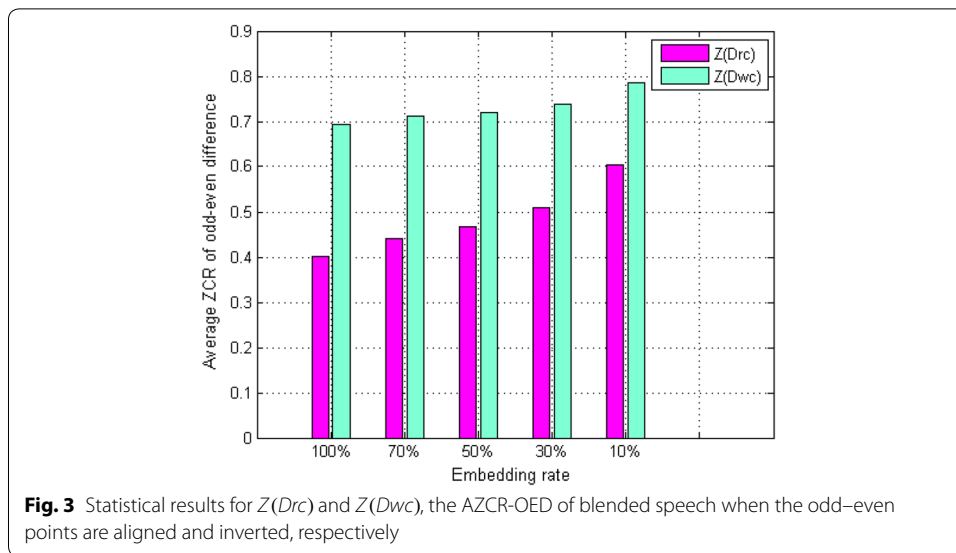
Experiment 1: we calculated the OEDs  $Dro$  and  $Dwo$  for the pure speech in the VOASE library under both conditions, where the odd–even points are aligned and inverted, along with the corresponding average ZCR values,  $Z(Dro)$  and  $Z(Dwo)$ . Figure 2 shows the statistical results.

Experiment 2: we made five copies of the carrier speech group. Then, secret speech was embedded into the carrier speech signal using the blending-based speech hiding algorithm with five different embedding rates. Because the embedding rate is typically high in practical applications, we selected 10, 30, 50, 70, and 100 % as the embedding rates in the experiment. When the odd–even points are aligned, the AZCR-OED of blended speech is unrelated to the hidden degree factor. Thus, we used a hidden degree factor of 0.1 in the experiment. Consequently, we obtained five blended speech groups with different embedding rates. We calculated the OED  $Drc$  of each blended speech signal and the corresponding average ZCR  $Z(Drc)$ . Then, we inverted the odd–even points of each blended speech, and calculated the OED  $Dwc$  of each inverted blended speech signal as well as the corresponding average ZCR  $Z(Dwc)$ . Figure 3 shows the statistical results.

Experiment 3: we made fifteen copies of the carrier speech group and then embedded secret speech into the carrier speech using the blending-based speech hiding algorithm with fifteen combinations of three hidden degree factors and five embedding rates. In



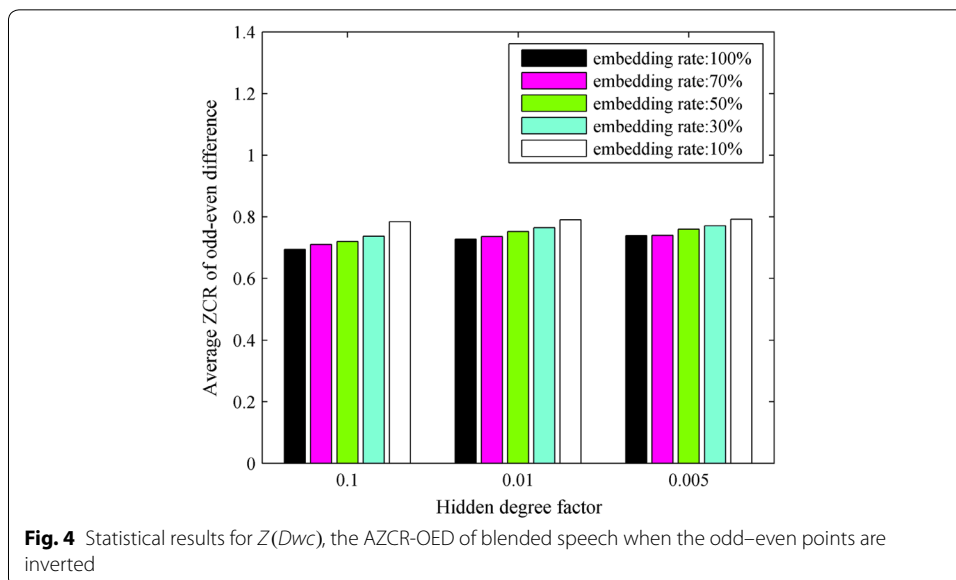
**Fig. 2** Statistical results for  $Z(Dro)$  and  $Z(Dwo)$ , the AZCR-OED of pure speech when the odd–even points are aligned and inverted, respectively



the experiment, the hidden degree factors were 0.1, 0.01, and 0.005, and the embedding rates were 10, 30, 50, 70, and 100 %, respectively. Thus, we obtained fifteen blended speech groups. We inverted the odd-even points of the speech in the blended speech groups and calculated the OED  $Dwc$  of each inverted blended speech signal and the corresponding average ZCR  $Z(Dwc)$ . Figure 4 shows the statistical results.

From Fig. 2, we can observe that whether the odd-even points are aligned has almost no effect on the AZCR-OED of the pure speech, i.e.,  $Z(Dro) \approx Z(Dwo)$ .

From Fig. 3, we can observe that  $Z(Drc)$  and  $Z(Dwc)$ , which are the AZCR-OED of the blended speech when the odd-even points are aligned and inverted respectively, increase as the embedding rate decreases. Comparing  $Z(Drc)$  and  $Z(Dwc)$ , we can see that the former is significantly less than the latter when the embedding rate is the same.



From Fig. 4, we can observe that the hidden degree factor of the secret speech has little effect on  $Z(Dwc)$  when the embedding rate is kept fixed.

Comparing Figs. 2 and 3, we can see that  $Z(Drc)$  is significantly less than  $Z(Dro)$  and  $Z(Dwo)$  when the embedding rates are different.

From the experimental results, we can conclude that it is feasible and effective to use the AZCR-OED of the speech as a feature to distinguish between blended and pure speech.

### Steganalysis algorithm for blending speech

In this section, we use the differences of the AZCR-OED between blended speech and pure speech to build a steganalysis algorithm for blended speech. The algorithm can achieve the following under a variety of embedding rates.

1. It can detect the existence of secret speech.
2. For blended speech, it can determine the hidden location of the secret speech and estimate the embedding rate.

### Existence detection of blended speech

Because the AZCR-OED of blended and that of pure speech have obvious differences, when the odd–even points of the blended speech are aligned, the AZCR-OED is lower than when they are inverted. In this paper, according to the features of the blended speech, we first correct the case in which the odd–even points are inverted and then extract the features and use a SVM classifier to detect the existence of blended speech. Table 1 summarizes the procedure for extracting the features from a speech signal.

Comparing Fig. 2 with Fig. 3, we see that the differences in the AZCR-OED between blended and pure speech are reduced when the embedding rate is low. It is thus easy to make a misjudgment using the SVM classifier. However, when the AZCR-OED of the stego speech segments is low, calculating the average ZCR per frame and extracting the minimum one as the feature can enhance the differences between blended speech and pure speech, and misjudgment can be reduced. For this reason, the second feature  $v_2$  is chosen.

In this paper, we use the above algorithm to extract the feature vector of the blended and pure speech and then use the freely available software package LIBSVM for training to establish a classifier, finally achieving existence detection of the blended speech.

**Table 1 Algorithm for feature extraction**

---

**Input:** A speech signal  $X$  of length  $N$

**Output:** A feature vector  $F$  that contains two features

*Step 1:* For a given speech signal  $X$ , invert its odd–even points to obtain the inverted version  $X_w$

*Step 2:* Calculate the OED of  $X$ , which is denoted by  $D_r$ , and the OED of  $X_w$ , which is denoted by  $D_w$

*Step 3:* Calculate the average ZCRs of  $D_r$  and  $D_w$ , respectively, which are denoted by  $Z(D_r)$  and  $Z(D_w)$ . Obtain the first feature  $v_1$ , where  $v_1 = \min\{Z(D_r), Z(D_w)\}$ . Here, the reason we take the smaller average ZCR is to correct the case in which the odd–even points of the blended speech are inverted

*Step 4:* If  $Z(D_r) \leq Z(D_w)$ , set  $D = D_r$ ; otherwise,  $D = D_w$ . Divide  $D$  into  $N$  frames and calculate the average ZCR per frame, choosing the smallest value as the second feature, which is denoted by  $v_2$

*Step 5:* Construct the feature vector  $F = \langle v_1, v_2, Type \rangle$  from these two features. The type attribute specifies whether the signal is blended (1) or pure (–1) speech object

---



### Estimation of the embedding rate

For the blended speech, the AZCR-OED of the stego speech segments is less than that of the pure speech segments. Table 2 presents the algorithm to detect the hidden location of the secret speech and estimate the embedding rate.

We chose two speech samples randomly from the VOASE library. Then, we chose one to be the secret speech and hid it in the other using the blending-based speech hiding algorithm with an embedding rate of 50 % and a hidden degree factor  $\alpha = 0.01$ . Then, the algorithm presented above was used to detect the hidden location of the secret speech. A frame length of 256 was used. The result is shown in Fig. 5.

From Fig. 5a, we can observe that the AZCR-OED values of most of the stego speech frames are less than the value for the entire blended speech signal, and the AZCR-OED values of most of the pure speech frames are greater than the value for the entire blended speech signal. Comparing panels (a) and (b) of Fig. 5, we see that the estimated hidden location of the secret speech is similar to its actual hidden location. These experimental results demonstrate the effectiveness of the algorithm.

**Table 2 Algorithm for estimating the embedding rate**

---

**Input:** A blended speech signal  $X$  of length  $N$

**Output:** Hidden location of the secret speech and the estimated embedding rate

*Step 1:* For a given speech signal  $X$ , invert its odd-even points to obtain the inverted version  $X_w$

*Step 2:* Calculate the OED of  $X$ , which is denoted by  $D_r$ , and the OED of  $X_w$ , which is denoted by  $D_w$

*Step 3:* Calculate the average ZCRs of  $D_r$  and  $D_w$ , respectively, which are denoted by  $Z(D_r)$  and  $Z(D_w)$ , and set  $Q_{mean} = \min(Z(D_r), Z(D_w))$

*Step 4:* If  $Z(D_r) \leq Z(D_w)$ , set  $D = D_r$ ; otherwise,  $D = D_w$

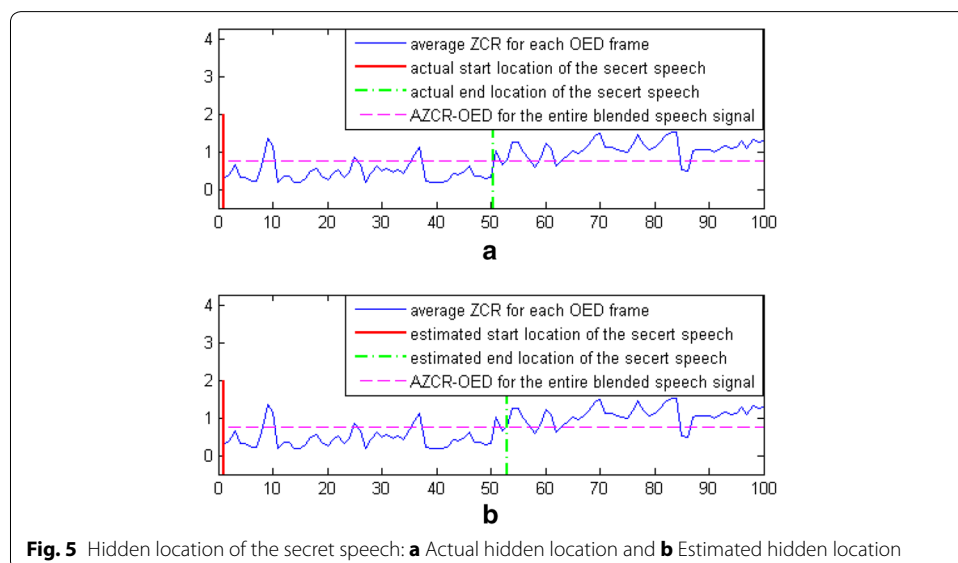
*Step 5:* Divide  $D$  into  $N$  frames and calculate the average ZCR per frame, which is denoted by  $Q(i)$ , where  $i$  denotes the  $i$ th frame

*Step 6:* Let  $Flag(i)$  denote the symbol of the  $i$ th frame:  $Flag(i) \in \{0, 1\}$ . If  $Q(i) < Q_{mean}$ , set  $Flag(i) = 0$ ; otherwise,  $Flag(i) = 1$

*Step 7:* The hang-over scheme (Avciabas 2006; Muhammad 2015), which is a type of VAD algorithm, is used for  $Flag(i)$  to mark the secret speech segments, thereby ensuring that the hidden location of the secret speech is determined

*Step 8:* Calculate the length of the secret speech segments and the embedding rate

---



## Experimental results and analysis

In this section, we evaluate the performance of the algorithms for feature extraction and estimation of the embedding rate.

One thousand speech files were randomly selected from the VOASE library and divided into two groups. One group was the secret speech group, which contained 500 secret speech files. Ten copies of the other 500 speech files were made to be used as carrier speech files, which were embedded with secret speech using the blending-based speech hiding algorithm with 10 different embedding rates of 10, 20, 30... 100 %. Because the AZCR-OED of blended speech is unrelated to the hidden degree factor of the secret speech, the hidden degree factor was chosen to be 0.01 in the experiments. Therefore, 500 blended speech signals were obtained for each embedding rate.

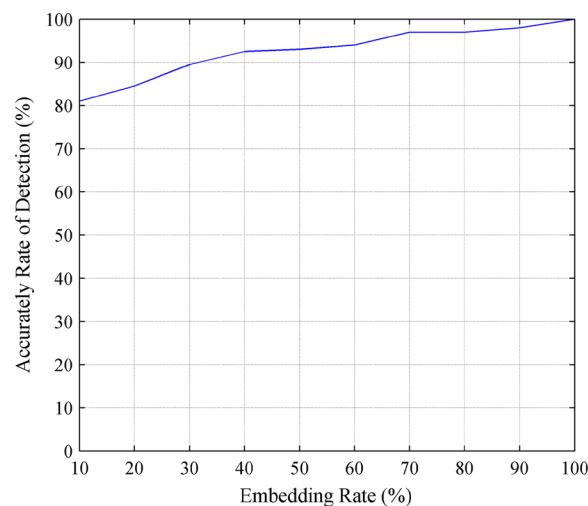
### Basic experiments

#### *Existence detection of the blended speech*

We randomly chose 250 blended speech files and 250 pure carrier speech files for each embedding rate and marked their type as 1 or  $-1$ , respectively. We then extracted the feature parameters for training the SVM classifier and used the remaining 500 speech samples, including 250 blended and 250 pure speech samples, for testing. Figure 6 shows the experimental results.

From Fig. 6, we can observe that the detection accuracy can be greater than 80 % when the embedding rate is greater than 10 %, and the accuracy of detection increases as the embedding rate increases.

From the results, we conclude that the reason that the detection accuracy decreases as the embedding rate decreases is as follows. The amount of secret speech decreases when the embedding rate decreases, thus, the differences in the AZCR-OED between blended and pure speech decrease. It is therefore easy for the SVM classifier to make a misjudgment, thereby decreasing the detection accuracy. However, for a fixed length of secret speech, a lower embedding rate requires a longer carrier speech signal, which leads to



**Fig. 6** Detection accuracy for ten different embedding rates

**Table 3 Embedding rate estimation**

Actual embedding rate (%)	Estimated embedding rate	
	Mean (%)	Variance
10	12.54	$2.80 \times 10^{-2}$
20	21.48	$2.50 \times 10^{-2}$
30	30.13	$1.98 \times 10^{-2}$
40	40.53	$1.79 \times 10^{-2}$
50	51.31	$1.79 \times 10^{-2}$
60	60.23	$9.2 \times 10^{-3}$
70	69.63	$8.1 \times 10^{-3}$
80	81.53	$5.3 \times 10^{-3}$
90	90.43	$6.7 \times 10^{-3}$
100	98.43	$1.64 \times 10^{-2}$

low communication efficiency. Thus, a low embedding rate is not adopted in practical applications of covert speech communication.

#### **Embedding rate estimation**

Under each actual embedding rate, we used 500 blended speech files to estimate the embedding rate using the algorithm described in “[Estimation of the embedding rate](#)” section and then calculated the mean and variance of the estimated embedding rate. The experimental results are presented in Table 3.

From Table 3, we can observe that the estimated embedding rate is similar to the real value, and the variance is small. The AZCR-OED of the blended speech increases as the embedding rate decreases, which increases the threshold that distinguishes the stego and pure speech frames. Therefore, many pure speech frames were misjudged to be stego speech frames, which caused the embedding rate to be overestimated.

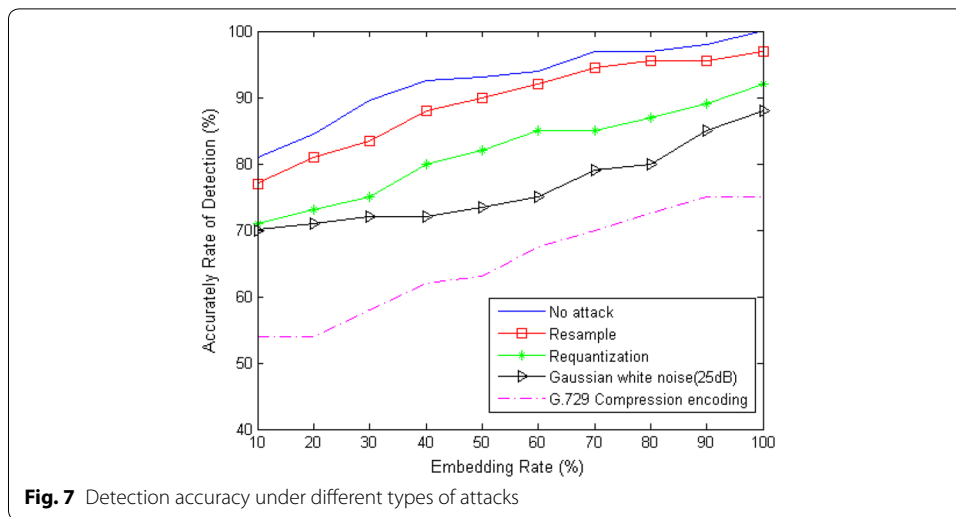
#### **Experiments for robustness**

To testify the robustness of this algorithm, we designed a group of attack experiments on the test speech for each embedding rate, including: (1) Resample: the speech is sampled up to 16-kHz then sampled down to 8-kHz. (2) Requantization: 16-bit encoding speech is converted to 8-bit encoding. (3) Gaussian white noise: the white noise is added with the SNR being 25 dB. (4) G.729 compression encoding.

We extracted the feature parameters from the attacked training speeches, and sent them to trained SVM classifier respectively. Then we can detect the existence of secret speech. Figure 7 shows the experimental results (for a better comparison, we redraw the curve in Fig. 7).

From Fig. 7, we can observe that:

1. for the different embedding rate, the detection accuracy without attack is higher than that with attacks;
2. the attacks of resample and requantization have a lower effect on detection accuracy;
3. the attack of Gaussian white noise has a higher effect on detection accuracy;



**Fig. 7** Detection accuracy under different types of attacks

4. after the speech is encoded with the G.729 compression encoding, the performance of the detection is reduced substantially;
5. under different types of attacks, the detection accuracy increases as the embedding rate increases. Therefore, we can divide a long speech into short speech segments so that we can improve embedding rate in some speech segments of blended speech. Thus, the detection accuracy can be increased.

Similarly, under each actual embedding rate, we processed 500 blended speeches using the aforementioned attacks, and then estimated the embedding rate using the algorithm described in “[Estimation of the embedding rate](#)” section. The experimental results are presented in Table 4.

From Table 4, we can find that no matter what kind of attack it is, there is certain influence on estimated embedding rate. And the attacks lead to a bigger estimated error than the case of no attack. But the estimated embedding rate is around the actual embedding rate.

**Table 4** Embedding rate estimation under different types of attacks

Actual embedding rate (%)	No-attack (%)	The types of attacks			
		Resample (%)	Requantization (%)	Gaussian white noise (25 dB) (%)	G.729 compression encoding (%)
10	12.54	6.87	7.00	12.31	14.07
20	21.48	15.53	16.71	22.32	14.97
30	30.13	24.68	34.06	33.31	22.58
40	40.53	46.01	44.19	43.29	48.36
50	51.31	56.68	54.79	52.61	58.48
60	60.23	65.89	65.18	57.41	67.31
70	69.63	74.52	73.63	67.15	78.12
80	81.53	82.91	83.00	77.78	86.79
90	90.43	87.88	88.35	85.87	82.98
100	98.43	96.64	96.08	94.96	92.21

In conclusion, for the steganalysis algorithm which is proposed in this paper, we can safely conclude that it has a good robustness for the attacks such as resample, requantization. But it is sensitive to Gaussian white noise and G.729 compression encoding.

From the Eq. (3), it can be seen that the algorithm hides the secret speech by modifying the values of even points of carrier speech in the time domain. When Gaussian white noise is imposed on blended speech, the OEDs of the blended speech will be changed significantly due to the randomness of the noise. Especially when the positive or negative prescriptions of the OEDs which are close to zero are changed, it will have large impact on ZCRs of each frame and the entire speech. The probability of misjudgment and the estimated error of embedding rate will then be increased.

G.729 compression encoding divides the speech signal into frames in length of 10 ms. Each encoded signal frame is represented with 80 bits. For the WAV format speech with an 8-kHz sampling rate and 16-bit quantization, the data rate is 128kbps. With a compression rate of 16:1, the data rate will be lowered to 8kbps after the G.729 compression encoding. Consequently, G.729 compression encoding will cause a large loss of numerical information in the time domain of the blended speech. It inevitably leads to that the AZCR-OED is not compliance with the original features any more. Therefore this case also can increase the probability of the misjudgment, as well as the estimated error of embedding rate.

Although the robustness of this steganalysis algorithm is not ideal when speech is attacked by Gaussian white noise or G.729 compression encoding. But in the practical application, the probability is low for the blended speech to be attacked by strong white noise or compression encoding. There are two main reasons. First, the Internet channel has a high-quality. Second, the hidden capacity of the blending-based speech hiding algorithm is high, which results that the blended speech has a low compressibility. In order to extract the secret speech correctly, for both sides of the covert communication, it is less likely to transmit the compression blended speech.

## Conclusion

In this paper, we first briefly introduced the background and significance of this paper along with the blending-based speech hiding algorithm. Then, considering a high-quality Internet, we analyzed the differences in the OED values of blended and pure speech and quantified these differences using the average ZCR. The experimental results verified the correctness of the theoretical analysis. Finally, we proposed a steganalysis algorithm for blended speech. The algorithm successfully achieved existence detection of blended speech and embedding rate estimation under many embedding rates, and the experimental results demonstrate the high accuracy, effectiveness and robustness of this algorithm. Determining how to estimate the hidden degree factor  $\alpha$  of secret speech and how to extract the secret speech signal are our future research goals.

### Authors' contributions

LLJ conceived and carried out the study including design, experiments, and analysis, and drafted the manuscript. GY supervised the whole work and participated in critically assessing the experiments, and reviewed the manuscript. Both authors read and approved the final manuscript.

### Authors' information

Lijuan Li was born in Sichuan, China in 1991. She received the B.S. and M.S. degrees from the College of Electronics and Information Engineering, Sichuan University, Chengdu, China, in 2013 and 2016, respectively. Her research area mainly includes covert audio communication, audio steganalysis, speech signal processing, and speech

enhancement. Yong Gao was born in Xi'an, China in 1969. He received the M.S. and Ph.D. degrees from the school of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu, in 1997 and 2000, respectively. He is a professor in College of Electronics and Information Engineering, Sichuan University. His research area mainly includes speech signal processing, anti-interference and anti-interception technology in communication, modulation recognition, emergency communication, array signal processing, blind analysis of signal.

#### Acknowledgements

The authors would like to acknowledge the authors of the references for the work that they accomplished.

#### Competing interests

The authors declare that they have no competing interests.

Received: 20 January 2016 Accepted: 24 June 2016

Published online: 11 July 2016

#### References

- Alex JS, Bernhard S (2004) A tutorial on support vector regression. *Stat Comput* 14(3):199–222
- Ali OAN, Salina AS, Aini H, Aishah F (2011) A new voice controlled noise cancellation approach. *Int J Adv Sci Eng Inf Technol* 1(4):386–390
- Avcibas I (2006) Audio steganalysis with content-independent distortion measures. *IEEE Signal Process Lett* 13(2):92–95
- Bhattacharyya S, Sanyal G (2012) Feature based audio steganalysis (FAS). *Int J Comput Netw Inf Security* 4(11):62
- Byeong-Seob K, Nishimura R, Suzuki Y (2005) Time-spread echo method for digital audio watermarking. *J IEEE Trans Multimed* 7(2):212–221
- Chen J (2001) Amalgamation-based digital audio information hiding method. *J Fuzhou Univ (Natural Science)* 29(4):104–106
- Chunhui X, Yimin C (2010) Spread-spectrum steganalysis and PN sequence estimation. In: *Proceedings of IEEE 3rd CISP, Yantai, October, 2010*, pp 4143–4147
- Gelfand A (2007) Audio forensics experts reveal (some) secret. Available: [http://www.wired.com/science/discoveries/news/2007/10/audio\\_forensics](http://www.wired.com/science/discoveries/news/2007/10/audio_forensics), October, 2007
- Ghasemzadeh H, Mehdi TK, Meisam KA (2016) Audio steganalysis based on reversed psychoacoustic model of human hearing. *Digital Signal Process* 51:133–141
- Ghosal A, Suchibrota D (2014) Automatic male-female voice discrimination. In: *Proceedings of ICICT, Ghaziabad, February, 2014*, pp 731–735
- Hamza Ö, Ismail A, Bülent S, Nasir M (2003) Steganalysis of audio based on audio quality metrics. In: *Proceedings of SPIE 5020, security and watermarking of multimedia contents V, CA, June, 2003*, pp 55–66
- Hartoko CFS, Tjondronegoro S, Hidayat B (2015) Audio steganography using modified enhanced least significant bit in 802.11n. *J Measure Electron Commun Syst* 1(1):7–13
- Jongseo S, Nam SK, Wonyong S (1999) A statistical model-based voice activity detection. *IEEE Signal Process Lett* 6(1):1–3
- Krishnan S, Abdullah MS (2016) Enhanced security audio steganography by using higher least significant bit. *J Adv Res Comput Appl ISSN (online)* 2(1):39–54
- Mathias MA, Mohamed C (2015) Support vector machine. *Encycl Biom* 1504–1511
- Matsuoka H (2006) Spread spectrum audio steganography using sub-band phase shifting. In: *Proceedings of IIH-MSP'06, Pasadena, December, 2006*, pp 3–6
- Muhammad ZI (2015) Detecting double talk in acoustic echo cancellation using zero-crossing rate: U.S. Patent 9083783[P]. 14 July 2015
- Natarajan M, Nayak L (2010) Steganalysis algorithms for detecting the hidden information in image, audio and video cover media. *Int J Netw Security Appl* 2(1):43–55
- Nutzinger M, Juergen W (2011) A novel phase coding technique for steganography in auditive media. In: *Proceedings of IEEE 6th ARES, Vienna, August, 2011*, pp 91–98
- Özer H, Sankur B, Memon N, Avcıbaşı İ (2006) Detection of audio hidden channels using statistical footprints of hidden messages. *Digital Signal Process* 16(4):389–401
- Qi L, Jinsong Z, Tsai A, Qiru Z (2002) Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Speech Audio Process* 10(3):146–157
- Qiao M, Sung AH, Liu Q (2013) MP3 audio steganalysis. *J. Inf Sci* 231:123–134
- Rangding W, Qian L, Dawen X (2004) The study of information hiding technology based on audio blending. In: *Proceedings of 7th ICSP'04, Beijing, August–September, 2004*, pp 2405–2408
- Shota M, Masashi U, Xugang L, Masato A (2016) Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments. *J Signal Process Syst* 82(2):163–173
- Singh P (2016) A comparative study of audio steganography techniques. *J Int Res J Eng Technol* 3(4):581–585
- Tatsuya M, Kotaro S (2015) Audible secret keying for time-spread-echo based audio watermarking. In: *Proceedings of IEEE international conference on intelligent information hiding and multimedia signal processing (IIH-MSP), Adelaide, September, 2015*, pp 164–167
- Tayel M, Gamal A, Shawky H (2016) A proposed implementation method of an audio steganography technique. In: *Proceedings of IEEE 18th international conference on advanced communication technology (ICACT), PyeongChang, January–February, 2016*, pp 180–184

- Wei Z, Haojun A, Ruimin H, Shang G (2008) An algorithm of echo steganalysis based on Bayes classifier. In: Proceedings of ICIA, Changsha, June, 2008, pp 1667–1670
- Xinyan Z, Guojun Z (2013) A new multi-parameter dual-threshold state discrimination algorithm for voice activity detection. In: Proceedings of 9th ICNC, Shenyang, July, 2013, pp 1239–1243

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Immediate publication on acceptance
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

---

Submit your next manuscript at ▶ [springeropen.com](http://springeropen.com)

---