IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 6, 2019, accepted August 29, 2019, date of publication September 5, 2019, date of current version September 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939629

A Common Method for Detecting Multiple Steganographies in Low-Bit-Rate Compressed Speech Based on Bayesian Inference

JIE YANG[1,2], PENG LIU[1], AND SONGBIN LI[1]
[1]Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China
[2]Jiyang College, Zhejiang A & F University, Zhuji 311800, China

Corresponding author: Songbin Li (lisongbin@mail.ioa.ac.cn)

This work was supported in part by the Important Science and Technology Project of Hainan Province under Grant ZDKJ201807, in part by the National Natural Science Foundation of China under Grant U1636113, in part by the Hainan Provincial Natural Science Foundation of China under Grant 618QN309, in part by the Scientific Research Foundation Project of Haikou Laboratory, Institute of Acoustics, Chinese Academy of Sciences, in part by the IACAS Young Elite Researcher Project under Grant QNYC201829 and Grant QNYC201747, and in part by the Scientific Research Startup Fund for Talent of Jiyang College under Grant JY2018RC04.

ABSTRACT Analysis-by-synthesis linear predictive coding (AbS-LPC) is widely used in a variety of low-bit-rate speech codecs. The existing steganalysis methods for AbS-LPC low-bit-rate compressed speech steganography are specifically designed for one certain category of steganography methods, thus lacking generalization capability. In this paper, a common method for detecting multiple steganographies in low-bit-rate compressed speech based on a code element Bayesian network is proposed. In an AbS-LPC low-bit-rate compressed speech stream, spatiotemporal correlations exist between the code elements, and steganography will eventually change the values of these code elements. Thus, the method presented in this paper is developed from the code element perspective. It consists of constructing a code element Bayesian network based on the strong correlations between code elements, learning the network parameters by utilizing a Dirichlet distribution as the prior distribution, and finally implementing steganalysis based on Bayesian inference. Experimental results demonstrate that the proposed method performs better than the existing steganalysis methods for detecting multiple steganographies in the AbS-LPC low-bit-rate compressed speech.

INDEX TERMS AbS-LPC, compressed speech, steganography, steganalysis, Bayesian inference.

I. INTRODUCTION

Information hiding, also called steganography, is an ancient but effective technique of embedding secret information into innocent carriers without perception. Although its history can date back to 440 BC [1], the steganographic carriers have been developing over time [2]. In the past years, the carriers have evolved from images [3]–[5] to almost all media forms (e.g., videos [6], [7], audios [8], [9], texts [10], [11], network protocol [12], [13]). In recent years, with the rapid increase in bandwidth requirements and the trend of increasing network convergence, network streaming media services for communication have undergone unprecedented development. Since Voice over Internet Protocol (VoIP) technology has been widely used for real-time communication, it serves as a suitable carrier for transmitting secret information over the Internet. VoIP steganography is a means of imperceptibly embedding secret information into VoIP-based cover speech. There are many VoIP speech codecs, including G.711, G.723.1, G.726, G.728, G.729, internet Low Bitrate Codec (iLBC), and the Adaptive Multi-Rate (AMR) codec. Most of them, including G.723.1, G.729, AMR and iLBC, are low-bit-rate speech codecs that use analysis-by-synthesis linear predictive coding (AbS-LPC). At present, most methods of speech steganography utilize AbS-LPC low-bit-rate speech codecs to embed secret information for covert communication. Although the ability to conduct covert communication can be convenient in the contexts of both daily life and work, it also provides opportunities for criminal activity. Thus, for certain

The associate editor coordinating the review of this manuscript and approving it for publication was Malik Najmus Saqib.

 This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

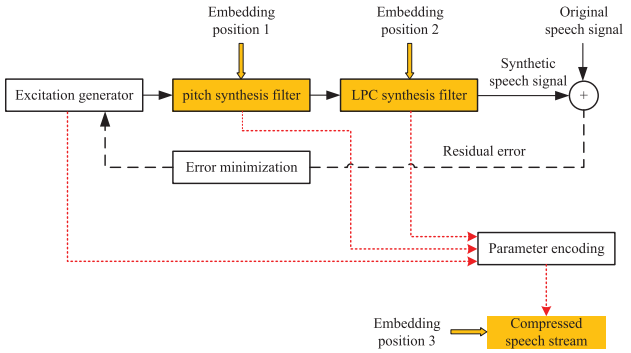

FIGURE 1. Three major categories of steganography methods for AbS-LPC low-bit-rate compressed speech. Embedding position 1 corresponds to steganography using the pitch synthesis filter. Embedding position 2 corresponds to steganography using the LPC synthesis filter. Embedding position 3 corresponds to steganography using the code elements in the compressed speech stream.

sensitive agencies, it is necessary to evaluate compressed speech streams to determine whether they carry hidden secret information.

There are three major categories of steganography methods for AbS-LPC low-bit-rate compressed speech, depending on the embedding position of the secret information, as shown in Fig. 1. Methods in the first category use the pitch synthesis filter to hide information [14]–[19]. Methods in the second category use the LPC synthesis filter for information hiding [20]–[24]. Methods in the third category hide information in the code elements (CEs) of the compressed speech stream [25]–[32].

For the detection of hidden secret information in AbS-LPC low-bit-rate compressed speech, speech steganalysis technology has developed in tandem with speech steganography technology. Current steganalysis methods for AbS-LPC low-bit-rate compressed speech can be divided into three categories corresponding to the different categories of steganography methods. Methods in the first category are designed to detect steganography based on the pitch synthesis filter [33]–[41]. Methods in the second category detect steganography based on the LPC synthesis filter [42]–[44]. Methods in the third category detect CE-based steganography [45], [46].

All of the current steganalysis methods have a feature-based design and consist of the same three steps. First, multiple features are extracted from the AbS-LPC low-bit-rate compressed speech data. Second, multi-feature fusion is performed, and dimensionality reduction is implemented on the high-dimensional fused features. At last, a support vector machine (SVM) is employed for classification. There are two main characteristics of these steganalysis methods. One is that they use a steganalysis framework based on SVM classification. The other is that they are specifically designed for one certain category of steganography methods and lack generality. Thus, for thorough detection, steganalysis methods for multiple categories of steganography need to be performed on a single compressed speech stream, which is a cumbersome and time-consuming process. Therefore, there is a need to develop a common method to achieve the simultaneous detection of multiple steganography methods. Some existing general steganalysis methods are based on uncompressed domain features, such as mel-frequency cepstral coefficient (MFCC) features [47]–[49]. Although these methods can be used for the general steganalysis of AbS-LPC low-bit-rate compressed speech, they are designed for the detection of steganography in uncompressed speech files. Since these methods do not incorporate the AbS-LPC principle, their detection accuracies for steganography in AbS-LPC low-bit-rate compressed speech are unsatisfactory. Therefore, a common method for detecting multiple steganographies in AbS-LPC low-bit-rate compressed speech is needed.

All steganography methods for AbS-LPC low-bit-rate compressed speech will eventually modify the CEs in the compressed speech stream. For instance, steganography methods in the first category will modify the CEs related to pitch, and these in the second category will modify the CEs related to the LPC. According to this, a common method for detecting multiple steganographies in AbS-LPC low-bit-rate compressed speech is proposed from the CE perspective. A code element Bayesian network (CEBN) is built based on the strong spatiotemporal correlations between the CEs. The network parameters are learned by utilizing a Dirichlet distribution as the prior distribution, and steganalysis is then implemented based on Bayesian inference.

In this paper, a new steganalysis method based on Bayesian inference is proposed, which is distinct from the traditional methods based on SVM classification. In the traditional methods, the features of each speech segment are extracted. However, these features cannot satisfactorily describe speech segments of short durations, resulting in low detection accuracies. By contrast, in the proposed steganalysis method based on Bayesian inference, the network parameters are learned for each speech frame, and the speech duration will not affect the conditional probabilities of the network nodes. Thus, high detection accuracies can be ensured even when the speech segment duration is short. Moreover, all of the CEs' values and their strong spatiotemporal correlations are mapped into the CEBN, thereby enabling effective steganalysis for AbS-LPC low-bit-rate compressed speech steganography.

The remainder of this paper is organized as follows. Section II reviews the related work on steganalysis. Section III analyses the spatiotemporal correlations between the CEs and describes the construction of the code element spatiotemporal correlation network (CESCN). The proposed Bayesian-inference-based steganalysis method for AbS-LPC low-bit-rate compressed speech steganography is detailed in Section IV, followed by the presentation and evaluation of the experimental results in Section V. Finally, the paper is concluded in Section VI.

II. RELATED WORK

For the detection of steganography methods in the first category (i.e., those using the pitch synthesis filter), many

SVM-based steganalysis algorithms have been proposed. Ding *et al.* extract three and five categories of histogram features related to the pulse position distribution to train SVM classifiers in [33] and [34], respectively. Miao *et al.* [35] utilize the Markov transition probabilities, joint entropies, and conditional entropies of the pulse positions as the features considered for classification. Tian *et al.* [36] achieve better performance at low embedding rates than the two methods presented in [33], [34] by utilizing the probability distributions, Markov transition probabilities and joint probabilities of the pulse positions as the features considered for classification. Ren *et al.* [37] find that steganography based on the modulation of the pulse positions during the fixed codebook search process causes the probability of same pulse positions (PSPP) in the same track to increase. Based on this phenomenon, they propose a set of steganalysis features based on the PSPP. Since PSPP features describe only the distributions of pulses being in the same position in two tracks, Tian *et al.* [38] employ probability distributions, Markov transition probabilities and joint probabilities to characterize pulse pairs and use the adaptive boosting technique to reduce the feature dimensionality. Based on the above features, they propose three classification schemes to achieve steganalysis for unknown embedding rates in [39]. Li *et al.* [40] propose a steganalysis method for pitch period modification steganography based on a codebook correlation network model, in which the conditional probabilities of strongly correlated nodes are used as features and principal component analysis (PCA) is applied for dimensionality reduction before training the SVM classifier. Ren *et al.* [41] propose a steganalysis method based on the matrix of the second-order differences in pitch delay (MSDPD) and obtain calibrated MSDPD features through recompression to further improve the detection accuracy.

For the detection of steganography methods in the second category (i.e., those using the LPC synthesis filter), Li et al. [42], [43] propose two SVM-based steganalysis methods for QIMS, using independent and joint VQ codebooks, respectively. In [42], they train an SVM classifier for each VQ codebook by exploiting the distribution histogram of the corresponding LSP CE and the state transition probability of that histogram between adjacent frames as features. In [43], they construct a quantization codeword correlation network (QCCN) model based on the intra-frame and inter-frame correlation indices between the LSP CEs. Then, they use the transition probabilities of the QCCN edges as high-dimensional features for classification and reduce their dimensionality via PCA. In addition, Yang and Li [44] present a novel steganalysis method based on a Codeword Bayesian Network (CBN). The CBN is constructed based on the probability distribution and the steganography-sensitive transition relationships of codewords.

For the detection of steganography methods in the third category (i.e., those using the CEs in AbS-LPC low-bit-rate compressed speech streams), Tian *et al.* [45] present a distributed steganalysis scheme based on four types of features: histograms, differential histograms, Markov transition matrices and differential Markov transition matrices. Each feature type is utilized to train a different SVM classifier. Thus, four classifiers are trained for each CE, and the best feature type for each CE is used for classification. As an alternative to the SVM-based steganalysis method in [45], Huang *et al.* [46] propose a steganalysis method based on second statistical detection and regression analysis. This method can detect the hidden information and estimate the embedding rate in addition to determining whether a speech signal contains any hidden information in the first place. However, the detection accuracy of this method is lower than that of the SVM-based method.

Current steganalysis methods are designed based on the selection of a certain coding method or CE to identify features that will perform well in the classification process for identifying the corresponding type of steganography. However, they cannot detect multiple steganographies at the same time. In this paper, we propose a common method for AbS-LPC low-bit-rate compressed speech steganography from the perspective of CE. The proposed method considers the information contained in all CEs along with the AbS-LPC speech coding principles.

III. ANALYSIS OF THE SPATIOTEMPORAL CORRELATIONS OF THE CES IN ABS-LPC LOW-BIT-RATE COMPRESSED SPEECH STREAMS

Current AbS-LPC low-bit-rate compressed speech codecs essentially adopt the coding scheme shown in Fig. 1, although they have slight differences in their implementation details. G.723.1 uses multi-pulse maximum likelihood quantization coding in the high-rate 6.3 kbit/s mode and algebraic code-excited linear prediction (ACELP) coding in the low-rate 5.3 kbit/s mode, and the frame size is 30 ms. G.729 uses conjugate-structure ACELP coding, and the frame size is 10 ms. AMR uses ACELP coding with 8 rates, and the frame size is 20 ms. In this paper, we will use G.723.1 in the high-rate 6.3 kbit/s mode as the speech codec to illustrate the proposed method. Each frame in a compressed G.723.1 speech stream consists of 24 CEs: three LPC VQ index CEs, VQ_1, VQ_2 and VQ_3; four adaptive codebook lag CEs, ACL_0, ACL_1, ACL_2 and ACL_3; four combined adaptive and fixed gain CEs, $GAIN_0$, $GAIN_1$, $GAIN_2$ and $GAIN_3$; five pulse position index CEs, POS_0, POS_1, POS_2, POS_3 and MPOS; four pulse sign index CEs, $PSIG_0$, $PSIG_1$, $PSIG_2$ and $PSIG_3$; and four grid index CEs, $GRID_0$, $GRID_1$, $GRID_2$ and $GRID_3$.

A. ANALYSIS OF THE SPATIOTEMPORAL CORRELATIONS OF THE CES

A speech signal can be divided into unvoiced and voiced speech segments according to the phonemes it contains. Voiced speech carries most of the energy in a speech signal and has an obvious periodicity in the time domain. Unvoiced speech is similar to white noise and has no obvious

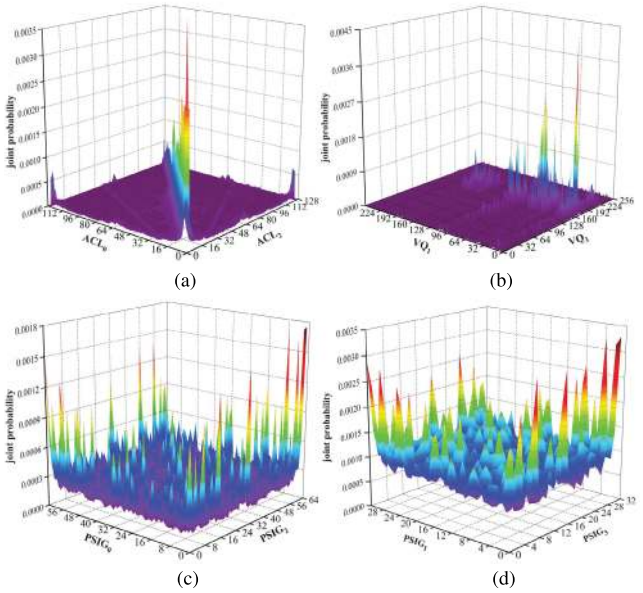

FIGURE 2. The joint intra-frame probability distributions of selected CE pairs. (a) ACL_0 and ACL_2. (b) VQ_1 and VQ_3. (c) $PSIG_0$ and $PSIG_2$. (d) $PSIG_1$ and $PSIG_3$.

periodicity. A speech signal is a non-stationary signal, but it can approximately be seen as stable over a period of 30 ms, i.e., exhibits short-term stability. The frame sizes of the commonly used AbS-LPC low-bit-rate compressed speech codecs all lie in this range. Therefore, there are intra-frame correlations between the CEs in a compressed speech stream. For instance, ACL_0 and ACL_2, which represent the pitch periods, have similar values in the same frame. We randomly collect 3000 speech segments with a speech length of 10s from the Internet to analyse the spatiotemporal correlations of the CEs. The joint intra-frame probability distributions for several CE pairs are shown in Fig. 2.

A strong correlation between two CEs will result in an uneven distribution in terms of the joint probability. The joint probabilities of ACL_0 and ACL_2 on and near the diagonal are much greater than those at other coordinates in Fig. 2(a), indicating a strong intra-frame correlation between ACL_0 and ACL_2. Similarly, the joint probabilities of VQ_1 and VQ_3 at some coordinates are much greater than those at others in Fig. 2(b), indicating a strong intra-frame correlation between these two CEs. By contrast, in Fig. 2(c), the joint probability distribution of $PSIG_0$ and $PSIG_2$ is relatively smooth, which indicates that the correlation between them is weak. Similarly, as seen from Fig. 2(d), the correlation between $PSIG_1$ and $PSIG_3$ is also weak, as reflected by the relatively smooth joint probability distribution of these CEs. To further objectively evaluate the joint probability distributions of the CEs, we divide the joint probabilities of each CE pair above into three regions, labelled as ''I'', ''II'' and ''III'', respectively. Let the average value of the joint probabilities of two CEs be denoted by u. ''I'' represents the proportion of the joint probabilities that are less than $0.5u$, ''II'' represents the proportion between $0.5u$ and $1.5u$, and ''III'' represents the proportion of joint probabilities larger than $1.5u$. Table 1 shows the proportions of the joint intra-frame CE probabilities in these three regions corresponding to Fig. 2.

TABLE 1. The proportions of several joint intra-frame CE probabilities separated into three regions.

Region	ACL_0-ACL_2	VQ_1-VQ_3	$PSIG_0$-$PSIG_2$	$PSIG_1$-$PSIG_3$
I	8.56%	11.85%	7.86%	5.76%
II	31.85%	12.07%	90.68%	94.14%
III	59.59%	76.08%	1.46%	0.10%

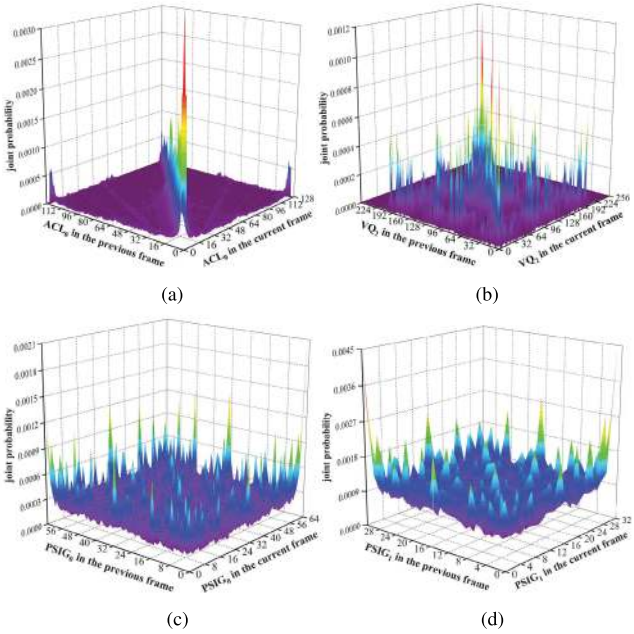

FIGURE 3. The joint inter-frame probability distributions of selected CEs. (a) ACL_0. (b) VQ_2. (c) $PSIG_0$. (d) $PSIG_1$.

As seen from the data in Table 1, more than half of the joint probabilities of ACL_0 and ACL_2 and of VQ_1 and VQ_3 are larger than $1.5u$, whereas those of $PSIG_0$ and $PSIG_2$ and of $PSIG_1$ and $PSIG_3$ are mainly distributed around u. These findings further indicate that the correlations between ACL_0 and ACL_2 and between VQ_1 and VQ_3 are stronger than those between $PSIG_0$ and $PSIG_2$ and between $PSIG_1$ and $PSIG_3$.

In addition, because of the local periodicity of speech signals, different frames may exactly correspond to periodically repeating speech signals. Therefore, there are also inter-frame correlations between the CEs. To illustrate these correlations, we analyse the joint probability distributions of several CEs between adjacent frames, as shown in Fig. 3.

As seen from Fig. 3(a), the joint probability distribution of ACL_0 between adjacent frames is quite uneven, as is that of VQ_2 in Fig. 3(b), whereas those of $PSIG_0$ in Fig. 3(c) and $PSIG_1$ in Fig. 3(d) are smoother. These findings indicate that ACL_0 and VQ_2 have strong inter-frame correlations, whereas $PSIG_0$ and $PSIG_1$ have weak inter-frame correlation. Table 2 shows the proportions of the joint inter-frame

TABLE 2. The proportions of several joint inter-frame CE probabilities separated into three regions.

Region	ACL_0	VQ_2	$PSIG_0$	$PSIG_1$
I	10.73%	14.85%	7.71%	6.15%
II	33.84%	20.02%	90.85%	93.75%
III	55.43%	65.12%	1.44%	0.10%

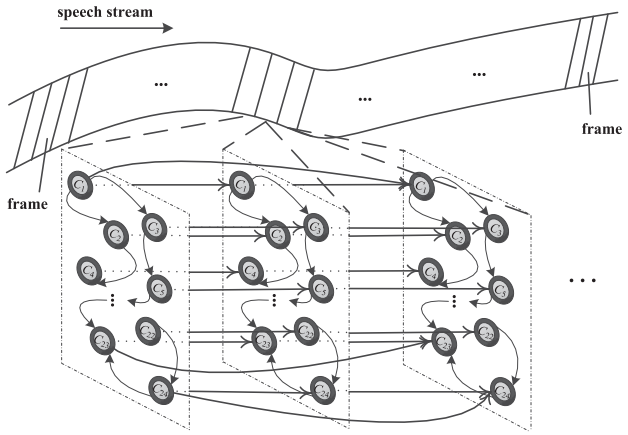

FIGURE 4. The code element spatiotemporal correlation network (CESCN). It contains 24 CE vertices, 276 intra-frame correlation edges and 576 inter-frame correlation edges.

CE probabilities in the three regions described previously corresponding to Fig. 3.

As seen from the data in Table 2, more than half of the inter-frame joint probabilities for ACL_0 and VQ_2 are larger than $1.5u$, whereas those for $PSIG_0$ and $PSIG_1$ are mainly distributed around u, thus further indicating that the inter-frame correlations of ACL_0 and VQ_2 are stronger than those of $PSIG_0$ and $PSIG_1$.

B. MODELLING THE SPATIOTEMPORAL CORRELATIONS OF THE CES

According to the analysis above, spatiotemporal correlations exist between the CEs in AbS-LPC low-bit-rate compressed speech streams. For convenience of description, the 24 CEs are denoted here by C_i ($i = 1, 2, \ldots, 24$). A correlation network that describes the spatiotemporal correlations between the CEs can be built from these 24 CEs. We call this network the code element spatiotemporal correlation network (CESCN) in this paper. It is constructed as shown in Fig. 4.

The CESCN is a directed graph composed of vertices and edges, where the vertices represent the 24 CEs and the edges represent the intra-frame and inter-frame correlations between the CEs. Mathematically, the CESCN is described as $D =< V, E >$, where V and E are defined as follows:

$$\begin{cases} V = \{v_i[m] | i \in \{1, 2, \ldots, 24\}, m \in \{0, 1, 2, \ldots\}\} \\ E = \{\langle v_i[p], v_j[q]\rangle | v_i[p] \in V, v_j[q] \in V\} \end{cases} \quad (1)$$

where V is the vertex set of D, and each element $v_i[m]$ represents C_i in the mth frame. E is the edge set of D, and each element $\langle v_i[p], v_j[q]\rangle$ represents a directed edge from vertex $v_i[p]$ to vertex $v_j[q]$. When $i \neq j$ and $q = p$, $\langle v_i[p], v_j[q]\rangle$ represents an intra-frame edge. When $q > p$, $\langle v_i[p], v_j[q]\rangle$ represents an inter-frame edge. Since the strengths of the inter-frame correlations are affected by time, i.e., a larger time interval results in weaker inter-frame correlations, only the correlations between adjacent frames are considered to simplify the analysis. In other words, we analyse only correlations for which $q-p \in \{0, 1\}$. Therefore, the CESCN contains 852 correlation edges: 276 intra-frame correlation edges and 576 inter-frame correlation edges. Since the CESCN contains too many correlation edges to be conducive to analysis, we wish to simplify it further. According to the analysis presented in Section III-A, the strengths of the correlations between each pair of CEs (i.e ., corresponding to each edge) are different. Therefore, edges representing weak correlations can be removed to simplify the CESCN.

TABLE 3. Bit allocation of the CEs in a G.723.1 stream in the high-rate 6.3 kbit/s mode.

CE	VQ_1	VQ_2	VQ_3	ACL_0	ACL_1	ACL_2
bits	8	8	8	7	2	7
CE	ACL_3	$GAIN_0$	$GAIN_1$	$GAIN_2$	$GAIN_3$	POS_0
bits	2	12	12	12	12	16
CE	POS_1	POS_2	POS_3	MPOS	$PSIG_0$	$PSIG_1$
bits	14	16	14	13	6	5
CE	$PSIG_2$	$PSIG_3$	$GRID_0$	$GRID_1$	$GRID_2$	$GRID_3$
bits	6	5	1	1	1	1

To quantify the correlation strength, we define the correlation index for C_i and C_j as

$$I_c(C_i, C_j) = \sum_{c_i=0}^{r_i} \sum_{c_j=0}^{r_j} \left| p(C_i = c_i)p(C_j = c_j) - p(C_i = c_i,\ C_j = c_j) \right| \quad (2)$$

where r_i and r_j are the maximum values of C_i and C_j, respectively. $p(C_i = c_i)$ and $p(C_j = c_j)$ represent the marginal probabilities of $C_i = c_i$ and $C_j = c_j$, respectively. And $p(C_i = c_i, C_j = c_j)$ represents the joint probability of $C_i = c_i$ and $C_j = c_i$. If C_i and C_j are independent of each other, these probabilities satisfy $p(C_i = c_i)p(C_j = c_i) = p(C_i = c_i, C_j = c_i)$ for any c_i and c_j, that is, $I_c = 0$. In other words, a stronger correlation between C_i and C_j results in a larger value of I_c. Table 3 shows the bit allocation of the 24 CEs.

As seen from the data in Table 3, the bit allocations of the different CEs vary significantly. Both POS_0 and POS_2 are allocated 16 bits each, with a range of [0, 65535]. The joint probability distribution of POS_0 and POS_2 is thus represented by a 65536×65536 matrix, the accurate calculation of which would require billions of speech frames. Therefore, we divide the values of the CEs into multiple value intervals to calculate the joint distribution. Each interval corresponds to the calculation of one probability in the joint distribution. First, the histogram distribution of a CE is calculated from

TABLE 4. Strong correlation edges and the corresponding correlation indices of CESCN.

Intra-frame	ACL_0-ACL_2	VQ_1-VQ_2	VQ_1-VQ_3	VQ_2-VQ_3
I_c	0.87	0.68	0.58	0.70
Inter-frame	ACL_0-ACL_0	ACL_0-ACL_2	ACL_2-ACL_0	ACL_2-ACL_2
I_c	0.82	0.72	0.73	0.82
Inter-frame	VQ_1-VQ_1	VQ_2-VQ_2	VQ_3-VQ_3	
I_c	0.73	0.84	0.93	

the 3000 speech segments chosen for analysis, as mentioned above. Second, the values of the CE are sorted in descending order according to this histogram distribution. Finally, the sorted values are uniformly partitioned into T intervals. Note that a CE whose maximum value is less than T is not divided. We use $T = 256$ to analyse the spatiotemporal correlations between the CEs and calculate the correlation indices for the 852 correlation edges. $Ic = 0.5$ is used as the criterion to determine whether an edge represents a strong or a weak correlation. If I_c is larger than 0.5, the edge represents a strong correlation; otherwise, it represents a weak correlation. The strong correlation edges and their indices as calculated for the 3000 speech segments chosen for analysis are shown in Table 4.

As seen from the data in Table 4, the simplified CESCN contains 4 strong intra-frame correlation edges and 7 strong inter-frame correlation edges. There are no strong spatiotemporal correlations between CEs in different categories, whereas there are strong spatiotemporal correlations among VQ_1, VQ_2 and VQ_3 as well as between ACL_0 and ACL_2. This is because the LPC VQ index CEs (VQ_1, VQ_2 and VQ_3) are the results of short-term analysis of the speech signals, whereas the adaptive codebook lag CEs (ACL_0 and ACL_2) are the results of long-term analysis. ACL_1 and ACL_3 are the differential adaptive codebook lags between the current sub-frame and the previous sub-frame, which are determined from ACL_0 and ACL_2. Therefore, the spatiotemporal correlation between ACL_1 and ACL_3 is weak. Since the other CEs are used to express the residual signal after the short-term and long-term predictions, their spatiotemporal correlations are weak as well.

IV. CODE ELEMENT BAYESIAN NETWORK AND STEGANALYSIS

A Bayesian network is a probability graph whose topological structure is that of a directed acyclic graph. The network consists of network nodes, directed edges and conditional probability tables (CPTs). The CESCN described above is similar to a Bayesian network. Thus, we will further construct a code element Bayesian network (CEBN) for steganalysis based on the strong spatiotemporal correlation edges in the CESCN.

A. CEBN CONSTRUCTION

Since one speech frame is used as the embedding unit' in AbS-LPC low-bit-rate compressed speech steganography, we construct the CEBN based on individual speech frames. The CEBN construction process is as follows:

1) Take the type of the current speech frame, i.e ., cover speech (denoted by 0) or stego speech (denoted by 1), as the root node C.

2) Take the values of the 24 CEs as the child nodes of C. Since the spatiotemporal correlations among $GRID_0$, $GRID_1$, $GRID_2$, and $GRID_3$ are weak and the corresponding number of bits for these CEs is very small (1 bit for each), these 4 nodes are merged into one node with a bit allocation of 4, denoted by GRID. Similarly, ACL_1 and ACL_3 are merged into one node with a bit allocation of 4, denoted by ACL. After merging, there are 20 child nodes and, thus, 20 edges from C to these 20 child nodes. The values of each node correspond to the intervals of the corresponding CEs as identified through the spatiotemporal correlation analysis of the CEs that is described in Section III-A.

3) Use the 4 strong intra-frame correlation edges ACL_0-ACL_2, VQ_1-VQ_2, VQ_1-VQ_3 and VQ_2-VQ_3 to form 4 edges: from ACL_0 to ACL_2, from VQ_1 to VQ_2, from VQ_1 to VQ_3 and from VQ_2 to VQ_3, respectively. Since the correlation index for the edge from VQ_2 to VQ_3 is larger than that for the edge from VQ_1 to VQ_3, which means that VQ_3 is more strongly affected by VQ_2 than by VQ_1, the edge from VQ_1 to VQ_3 is removed.

4) Use the 7 strong inter-frame correlation edges ACL_0-ACL_0, ACL_0-ACL_2, ACL_2-ACL_0, ACL_2-ACL_2, VQ_1-VQ_1, VQ_2-VQ_2 and VQ_3-VQ_3 to form 7 edges: from ACL_0 to ACL_0', from ACL_0 to ACL_2', from ACL_2 to ACL_0', from ACL_2 to ACL_2', from VQ_1 to VQ_1', from VQ_2 to VQ_2' and from VQ_3 to VQ_3', respectively. ACL_0', ACL_2', VQ_1', VQ_2' and VQ_3' are the child nodes of ACL_0, ACL_2, VQ_1, VQ_2 and VQ_3, respectively, and represent the values in the subsequent adjacent frame.

Since the correlation index for the edge from ACL_0 to ACL_0' is larger than that for the edge from ACL_2 to ACL_0', the edge from ACL_2 to ACL_0 is removed. Similarly, the edge from ACL_0 to ACL_2' is removed. In addition, ACL_0', ACL_2', VQ_1', VQ_2' and VQ_3' are influenced by the speech frame type. Therefore, 5 directed edges from C to ACL_0', ACL_2', VQ_1', VQ_2' and VQ_3' are added to the network.

Following the above process, the CEBN is constructed as shown in Fig. 5.

B. CEBN PARAMETER LEARNING

For ease of description, the nodes in Fig. 5 are denoted by the random variables $X_1, X_2, \ldots, X_{26}$, respectively, where X_1 corresponds to the root node and the others correspond to the child nodes, and their possible values are denoted by $x_1, x_2, \ldots, x_{26}$, respectively. The joint probability distribution of the CEBN is defined as follows:

$$P(X_1, X_2, \ldots, X_{26}) = \prod_{i=1}^{26} P(X_i|Pa(X_i)) \tag{3}$$

where $Pa(X_i)$ represents the parent nodes of X_i and $P(X_i|Pa(X_i))$ represents the conditional probability of X_i.

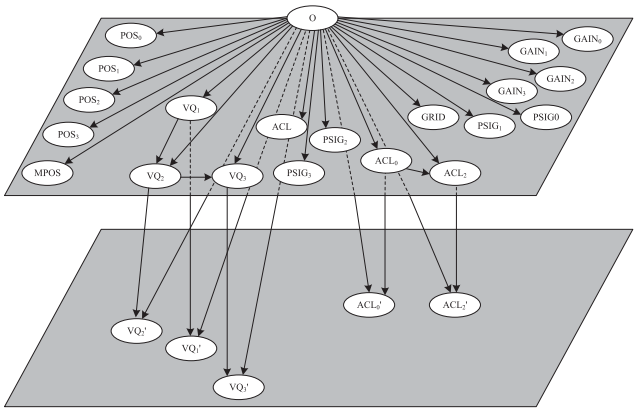

FIGURE 5. **The code element Bayesian network (CEBN). It contains 26 nodes representing information on all CEs, including not only the values of the 24 CEs but also their spatiotemporal correlations.**

Let K_i be the total number of possible values of X_i, and let θ_{ij} be the set of probabilities corresponding to these possible values. Let $\theta_{ij} = (\theta_{ij1}, \theta_{ij2}, \ldots, \theta_{ijk}, \ldots, \theta_{ijK_i})$, where θ_{ijk} is defined as follows:

$$\theta_{ijk} = P(X_i = x_{ik} | Pa(X_i) = Pa(X_i)_j) \tag{4}$$

where x_{ik} is the kth possible value of X_i and $Pa(X_i)_j$ is the jth possible value of X_i's parent nodes. The CEBN parameter learning process is essentially the process of learning θ_{ijk}. In this paper, the parameters are learned through Bayesian analysis. That is, the posterior distribution $\pi(\theta|\chi)$ is considered to be determined by both the prior distribution $\pi(\theta)$ and the sample information χ. In general, the conjugate distribution is used as the prior distribution. Therefore, the prior and posterior distributions both have the same form. We select a Dirichlet distribution as the prior distribution. Then, $\pi(\theta_{ij})$ is given by

$$\begin{aligned}\pi(\theta_{ij}) &= Dir(\alpha_{ij1}, \alpha_{ij2}, \cdots, \alpha_{ijK_i}) \\ &= \frac{\Gamma(\sum_{k=1}^{K_i} \alpha_{ijk}))}{\prod_{k=1}^{K_i} \Gamma(\alpha_{ijk})} \prod_{k=1}^{K_i} \theta_{ijk}{}^{\alpha_{ijk}}\end{aligned} \tag{5}$$

where $\Gamma(\cdot)$ is the gamma function and α_{ijk} is a hyperparameter. Let β_{ijk} be a number that satisfies $X_i = x_{ik}$ and $Pa(X_i) = Pa(X_i)_j$ in the sample χ. Since $\pi(\theta|\chi)$ follows the conjugate distribution, $\pi(\theta_{ij}|\chi)$ is given by

$$\begin{aligned}\pi(\theta_{ij}|\chi) &= Dir(\alpha_{ij1} + \beta_{ij1}, \alpha_{ij2} + \beta_{ij2}, \ldots, \alpha_{ijK_i} + \beta_{ijK_i}) \\ &= \frac{\Gamma(\sum_{k=1}^{K_i} (\alpha_{ijk} + \beta_{ijk}))}{\prod_{k=1}^{K_i} (\alpha_{ijk} + \beta_{ijk})} \prod_{k=1}^{K_i} \theta_{ijk}{}^{(\alpha_{ijk}+\beta_{ijk})}\end{aligned} \tag{6}$$

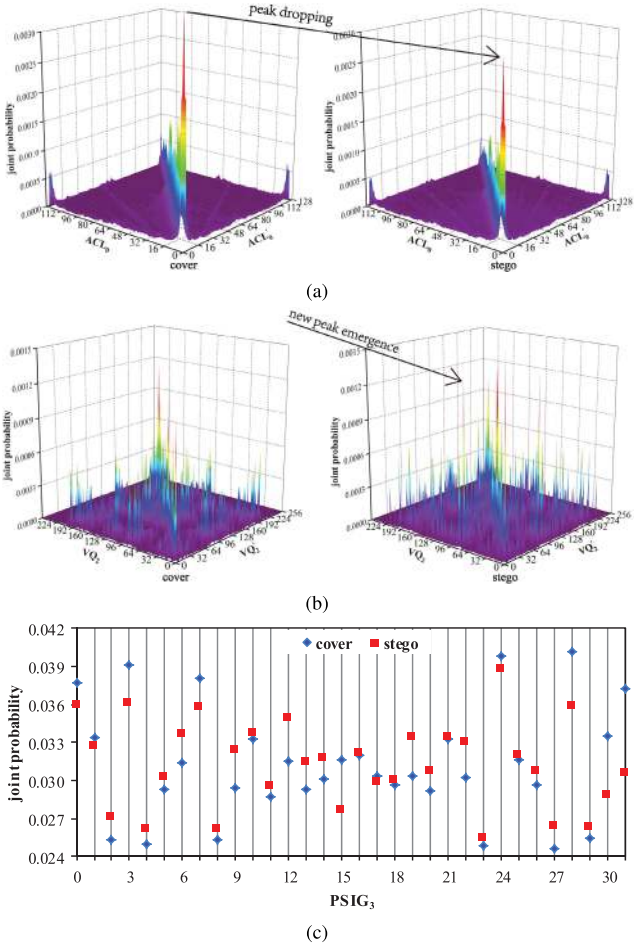

FIGURE 6. **Comparison of the joint probabilities for several nodes under the cover and stego conditions. (a) Joint probability distributions for ACL_0' under the cover and stego conditions for the steganography method of [19]. (b) Joint probability distributions for VQ_2' under the cover and stego conditions for the steganography method of [20]. (c) Comparison of the joint probabilities for $PSIG_3$ under the cover and stego conditions for the steganography method of [29].**

The maximum posterior estimate $\hat{\theta}_{ijk}$ of θ_{ijk} is the network parameter given by

$$\hat{\theta}_{ijk} = \frac{\alpha_{ijk} + \beta_{ijk}}{\sum_{k=1}^{K_i} (\alpha_{ijk} + \beta_{ijk})} \tag{7}$$

The network parameters for each random variable constitute the CPT of the corresponding node. Since the CPTs are too large to be presented as tables, we instead show some of them as plots. We learned the cover conditional probabilities and the stego conditional probabilities for several nodes based on the same 3000 speech segments selected for analysis as mentioned above. Three steganography methods [19], [20], [29] were used to learn three categories of stego conditional probabilities corresponding to the three categories of AbS-LPC low-bit-rate compressed speech steganography methods discussed previously. The results for several nodes are shown in Fig. 6.

In Fig. 6, we can see that the conditional probabilities for a given node are different under the cover and stego conditions and that some of these differences are large. These results illustrate that the CEBN can reflect the changes in the conditional probabilities before and after steganography. Thus, the CEBN can be regarded as an effective classifier for steganalysis.

C. STEGANALYSIS BASED ON THE CEBN

After constructing the CEBN and learning the CPTs, we classify speech frames based on Bayesian inference. Specifically, we exploit the values of the child nodes to infer the probability of the root node. Given a speech segment with N frames, the cover and stego probabilities for each frame can be computed. Let p_i^{cover} and p_i^{stego} be the cover and stego probabilities, respectively, of the ith frame. In theory, $p_i^{cover} > p_i^{stego}$ when the speech frame is a cover frame, and $p_i^{cover} < p_i^{stego}$ when the speech frame is a stego frame. Nevertheless, it is difficult to correctly classify every frame of a speech segment. Therefore, we do not use the frame-level results to judge whether there is secret information hidden in the speech segment. Instead, we introduce a steganography index J, expressed as $J = N^{stego}/N$, to reflect the steganography strength of a speech segment, where N^{stego} is the number of frames in the segment that are judged to be stego frames. We randomly selected 50 original PCM speech segments to be used as cover speech segments and for the generation of 5 types of stego speech segments, with 5 embedding rates of 100%, 80%, 60%, 40%, and 20%. Three steganography methods [19], [20], [29], as mentioned above, were used, and the corresponding steganography indices were computed. The results are shown in Fig. 7.

As shown in Fig. 7, a higher embedding rate results in a larger J. In other words, a larger J indicates a greater probability that secret information is hidden in a speech segment. The steganography index J reflects the difference between stego and cover segments. In this paper, we introduce a steganography index threshold J_{thr} based on which to classify cover and stego speech segments. For a given speech segment, if $J > J_{thr}$, then it is considered a stego speech segment; otherwise, it is considered a cover speech segment. Suppose that there are M speech segments in the training dataset. Let the steganography index sets of these M speech segments before and after steganography be denoted by $J_C = \{J_{c1}, J_{c2}, \ldots, J_{cj} \ldots, J_{cM}\}$ and $J_S = \{J_{s1}, J_{s2}, \ldots, J_{sj} \ldots, J_{sM}\}$, respectively. Then, J_{thr} satisfies the following equation:

$$J_{thr} = \arg \max_{J_{thr} \in J_S \cup J_C} \{CNT(J_S : J_{sj} > J_{thr}) + CNT(J_C : J_{cj} \leq J_{thr})\} \quad (8)$$

where $CNT(J_S : J_{sj} > J_{thr})$ and $CNT(J_C : J_{cj} \leq J_{thr})$ are the numbers of correctly classified stego and cover speech segments, respectively. After obtaining J_{thr}, we can judge whether secret information is hidden in a speech segment of unknown type.

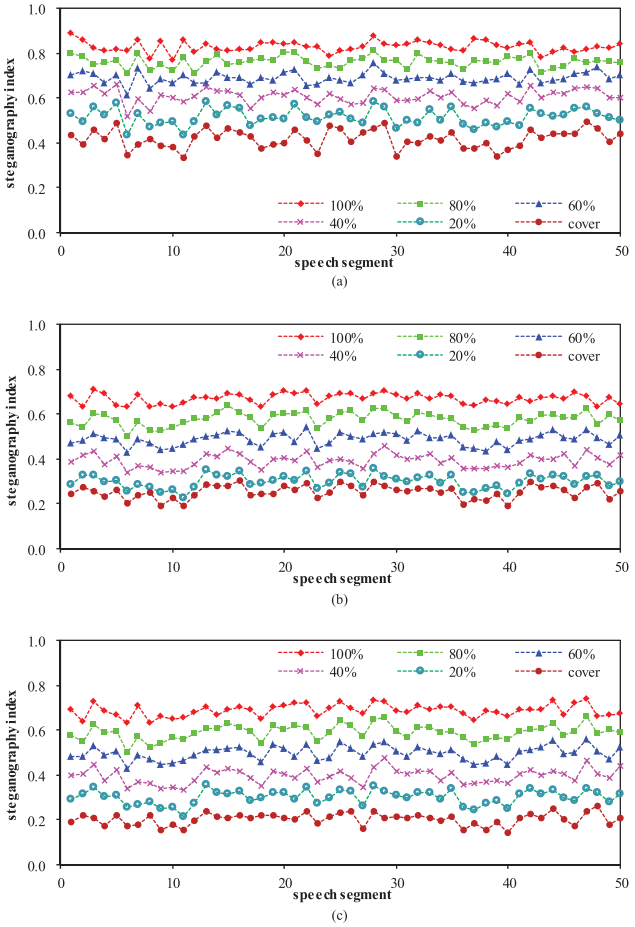

FIGURE 7. Comparison of the steganography indices of cover speech segments and corresponding stego speech segments with different embedding rates. (a) Steganography indices for the steganography method of [19]. (b) Steganography indices for the steganography method of [20]. (c) Steganography indices for the steganography method of [29].

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. EXPERIMENTAL SET-UP

We collected 7000 speech segments from the Internet, including samples from seven human voice categories, to form the speech database. Each category contains 1000 speech segments. The seven categories are Chinese man, Chinese woman, English man, English woman, French, German and Japanese. Each human voice category contains samples from more than five individuals. The duration of each speech segment is 10s, and each segment is formatted as a mono PCM file with an 8,000 Hz sampling rate and 16-bit quantization. The G.723.1 (6.3 kbit/s) codec is used as the low-bit-rate speech codec. The speech segments in each category are divided into a training dataset and a testing dataset at a 3:2 ratio. In other words, 4200 speech segments are used to learn the network parameters and calculate the steganography index threshold J_{thr}, and 2800 speech segments are used to assess the detection accuracy.

To evaluate the performance of the proposed CEBN-based steganalysis method, the methods of [19], [20], [22], [24], [29] are selected as representative steganography methods.

TABLE 5. The general detection accuracies of CEBN and MFCC for the five steganography methods at different embedding rates.

Steganography Method	Steganalysis Method	Embedding Rate									
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Stego-19	CEBN	60.40%	69.99%	78.36%	84.01%	88.25%	91.66%	93.98%	95.63%	96.74%	97.48%
	MFCC	51.43%	51.58%	52.26%	52.33%	51.90%	52.40%	53.48%	56.34%	59.43%	61.47%
Stego-20	CEBN	62.21%	73.53%	82.54%	88.41%	92.53%	95.66%	97.70%	98.93%	99.54%	99.78%
	MFCC	58.36%	59.17%	59.96%	60.16%	61.55%	62.89%	63.17%	64.25%	66.36%	68.76%
Stego-22	CEBN	60.35%	70.44%	76.45%	81.49%	85.47%	88.39%	92.17%	95.23%	97.12%	98.35%
	MFCC	55.37%	56.21%	57.03%	57.65%	58.25%	59.31%	61.3%	62.34%	64.03%	65.35%
Stego-24	CEBN	55.36%	62.36%	68.29%	74.41%	79.37%	83.96%	88.63%	92.88%	95.08%	95.46%
	MFCC	53.14%	54.28%	55.23%	56.41%	57.21%	58.03%	58.66%	59.13%	60.21%	61.32%
Stego-29	CEBN	62.41%	73.66%	82.63%	88.50%	92.53%	95.69%	97.76%	98.93%	99.55%	99.79%
	MFCC	61.32%	61.66%	60.74%	63.96%	65.47%	66.84%	68.08%	68.91%	70.68%	71.38%

As far as we known, there is no general method designed for the detection of steganographies in AbS-LPC low-bit-rate compressed speech. The MFCC-based steganalysis method [49] can detect any type of steganographies based on the decoded audio/speech data in theory. In this sense, we think that this method is also a general method. In order to evaluate our method thoroughly, we also compared our method with it. There is no steganalysis method that is specifically designed for the steganography method of [29], whereas the MSDPD-based method of [41] is specifically designed for the steganography method of [19]. In addition, the QCCN-based method of [43] and the CBN-based method of [44] are specifically designed for the steganography methods of [20], [22] and [24]. Therefore, the MSDPD-, QCCN- and CBN-based methods are selected as specialized steganalysis methods for comparison with the proposed method. For ease of description, the five steganography methods of [19], [20], [22], [24], [29] are denoted by Stego-19, Stego-20, Stego-22, Stego-24 and Stego-29, respectively, and the CEBN-, MFCC-, MSDPD-, QCCN- and CBN-based steganalysis methods are simply denoted by CEBN, MFCC, MSDPD, QCCN and CBN, respectively.

B. PERFORMANCE ANALYSIS OF THE GENERAL STEGANALYSIS MODELS AT DIFFERENT EMBEDDING RATES

One way to improve steganographic security is to reduce the embedding rate at which secret information is embedded into speech samples. Table 5 shows the general detection accuracies of CEBN and MFCC at 10 different embedding rates when the speech length is 10s and the network complexity is $T = 256$.

As seen from the data in Table 5, the detection accuracies of CEBN and MFCC increase with a increasing embedding rate. A higher embedding rate results in larger modifications to the CEs and causes the correlation characteristics of the CEs to change more significantly. Therefore, the Bayesian network can more easily distinguish these changes, and the detection accuracy of CEBN is higher. Similarly, a higher embedding rate results in a larger difference in the MFCC features used for classification between the cover and stego samples. Therefore, the accuracy of MFCC increases with an increasing embedding rate. When the embedding rate is 60% or above, the detection accuracies of CEBN for Stego-19 are higher than 90%, whereas when the embedding rate is 20% or below, the detection accuracies are lower than 70%. By contrast, the detection accuracies of MFCC for Stego-19 are lower than 62% at all embedding rates. Moreover, CEBN achieves good performance for the detection of Stego-20, Stego-22, Stego-24, and Stego-29. The detection accuracies of CEBN for Stego-20 and Stego-29 are higher than 80% at embedding rates of 30% or above, while those of MFCC are lower than 61%. In summary, CEBN performs better than MFCC does at all embedding rates. When the embedding rate is 80% or above, the general detection accuracies of CEBN for all five steganography methods are higher than 92% and are significantly higher than those of MFCC. These results demonstrate that the proposed method achieves effective general steganalysis performance for all three steganography categories.

C. PERFORMANCE ANALYSIS OF THE GENERAL STEGANALYSIS MODELS FOR DIFFERENT SPEECH LENGTHS

For a more comprehensive performance evaluation, we analyse the impact of the speech length on the general steganalysis performance. We tested the detection accuracies of CEBN and MFCC for each of the three steganography methods at ten different speech lengths with an embedding rate of 100% and a network complexity of $T = 256$. The experimental results are shown in Table 6.

As seen from the data in Table 6, the detection accuracies of CEBN for all five steganography methods are higher than 87% for each speech length, and the detection accuracies decrease with decreasing speech length. When the speech length is 1s, the detection accuracy of CEBN is 87.14% for Stego-24 and is higher than 91% for Stego-19, Stego-20, Stego-22 and Stego-29, while that of MFCC is lower than 54% for each of the five steganography methods. These results demonstrate that the proposed method performs well at short speech lengths. More specifically, the impact of the speech length on the general steganalysis performance of the proposed Bayesian-inference-based method is less than that for the SVM-classification-based method.

TABLE 6. The general detection accuracies of CEBN and MFCC for the three steganography methods at different speech lengths.

Steganography Method	Steganalysis Method	Speech Length									
		1s	2s	3s	4s	5s	6s	7s	8s	9s	10s
Stego-19	CEBN	91.83%	94.24%	96.08%	96.01%	96.56%	96.85%	97.18%	97.30%	97.28%	97.48%
	MFCC	50.19%	50.61%	51.06%	52.83%	54.91%	56.32%	58.55%	59.48%	60.66%	61.47%
Stego-20	CEBN	97.89%	99.10%	99.56%	99.63%	99.61%	99.71%	99.75%	99.76%	99.78%	99.78%
	MFCC	51.96%	53.31%	55.47%	60.67%	62.14%	64.37%	66.55%	67.32%	68.12%	68.76%
Stego-22	CEBN	91.27%	93.1%	94.56%	95.63%	96.19%	96.78%	97.35%	97.76%	98.24%	98.35%
	MFCC	50.88%	52.5%	54.31%	56.46%	58.32%	60.03%	61.59%	62.98%	64.4%	65.35%
Stego-24	CEBN	87.14%	88.77%	90.18%	91.36%	92.55%	93.37%	93.95%	94.75%	95.22%	95.46%
	MFCC	50.47%	51.88%	53.33%	54.88%	56.22%	57.26%	58.37%	59.55%	60.35%	61.32%
Stego-29	CEBN	97.93%	99.10%	99.56%	99.63%	92.62%	99.73%	99.75%	99.78%	99.79%	99.79%
	MFCC	53.17%	56.36%	60.24%	63.08%	65.29%	67.54%	68.91%	69.25%	70.85%	71.38%

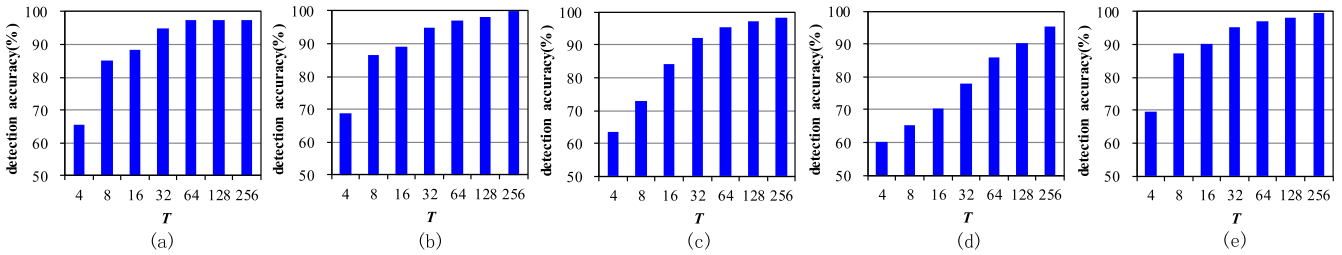

FIGURE 8. General detection accuracies of CEBN with different network complexities for the five steganography methods. (a) Stego-19. (b) Stego-20. (c) Stego-22. (d) Stego-24. (e) Stego-29.

D. PERFORMANCE ANALYSIS OF THE PROPOSED GENERAL STEGANALYSIS MODEL WITH DIFFERENT NETWORK COMPLEXITIES

In this section, we analyse the impact of the network complexity on the general steganalysis results. As seen from the CEBN construction process in Section IV-A, T determines the size of the CPTs in the CEBN. A larger T results in a more complex network. We tested the general steganalysis results of CEBN for all three steganography methods using seven T values of 4, 8, 16, 32, 64, 128 and 256. The experimental results are shown in Fig. 8.

As seen from the results in Fig. 8, the accuracy of the CEBN results increases with increasing T. This is because the larger the CPT size is, the better the CEBN can reflect the changes in the conditional probabilities before and after steganography. The detection accuracies of CEBN for all five steganography methods are higher than 86% when T is 64 or above. Moreover, the detection accuracies remain higher than 70% with a T of 16. This finding proves that the proposed method can achieve a satisfactory detection accuracy even with a low network complexity.

E. PERFORMANCE ANALYSIS OF THE SPECIALIZED STEGANALYSIS MODELS

Usually, only one steganography method is used to hide secret information. That is, all stego speech samples are generated with the same steganography method. For each of the general steganalysis methods (MFCC and CEBN), we trained three specialized steganalysis models, one for each steganography method, and we compare their performances with those of the corresponding general models and of the three inherently specialized methods (QCCN, CBN and MSDPD) in this section.

Fig. 9 shows the receiver operating characteristic (ROC) curves of the general and specialized steganalysis models for an embedding rate of 30%, a speech length of 10s and a network complexity of $T = 256$, where G-MFCC, S-MFCC, G-CEBN and S-CEBN denote the steganalysis results of the general MFCC, specialized MFCC, general CEBN and specialized CEBN models, respectively.

As seen from the results in Fig. 9, the specialized steganalysis methods perform better than the corresponding general steganalysis methods do, i.e., S-MFCC performs better than G-MFCC, and S-CEBN performs better than G-CEBN. In Fig. 9(a), we can see that for Stego-19, the specialized steganalysis method MSDPD performs better than G-CEBN and worse than S-CEBN, whereas QCCN and CBN are ineffective for Stego-19. Fig. 9(b)(c)(d) shows the results for Stego-20, Stego-22, and Stego-24, respectively. Similarly, the specialized steganalysis methods QCCN and CBN perform better than G-CEBN and worse than S-CEBN, whereas MSDPD is ineffective for Stego-20, Stego-22, and Stego-24. As seen in Fig. 9(e), QCCN and CBN also perform well for Stego-29. This is because Stego-29 modifies the LPC VQ index CEs. Both G-CEBN and S-CEBN perform better than QCCN and CBN do for Stego-29, and MSDPD is ineffective for this steganography method. Overall, for the detection of each of the five steganography methods, G-CEBN, S-CEBN and the corresponding specialized steganalysis method perform better than both G-MFCC and S-MFCC do. In summary, although G-CEBN performs slightly worse than QCCN, CBN and MSDPD do for the steganography

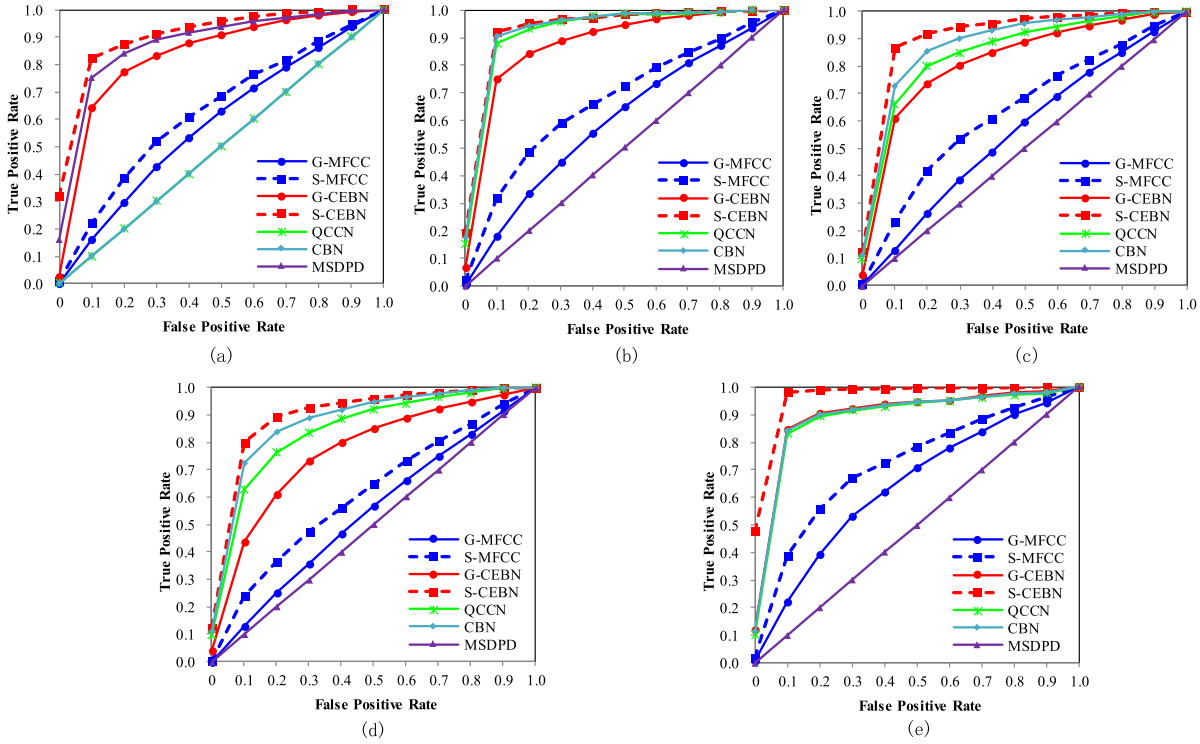

FIGURE 9. The ROC curves of the general steganalysis models (G-MFCC and G-CEBN) and the specialized steganalysis models (S-MFCC, S-CEBN, QCCN, CBN and MSDPD) for the five steganography methods with an embedding rate of 30%, a speech length of 10s and a network complexity of $T = 256$. (a) Stego-19. (b) Stego-20. (c) Stego-22. (d) Stego-24. (e) Stego-29.

methods for which they are specifically designed, G-CEBN works well for all five steganography methods, while QCCN, CBN and MSDPD are effective only for their corresponding steganography methods.

VI. CONCLUSION

In this paper, we propose a common method for detecting multiple steganographies in AbS-LPC low-bit-rate compressed speech. The correlations between the CEs are analysed from the spatiotemporal perspective, and a CEBN is constructed based on the strong correlations. The experimental results demonstrate that the proposed method performs better than the existing steganalysis methods for detecting multiple steganographies in AbS-LPC low-bit-rate compressed speech. The proposed method can achieve a satisfactory detection accuracy when the network complexity is low. In the future, we will further investigate steganalysis based on Bayesian networks for steganographies in compressed video stream.

ACKNOWLEDGMENT

The two institutions contributed equally to this work.

REFERENCES

[1] N. Provos and P. Honeyman, "Hide and seek: An introduction to steganography," *IEEE Security Privacy*, vol. 99, no. 3, pp. 32–44, May/Jun. 2003.

[2] E. Zieli ska, W. Mazurczyk, and K. Szczypiorski, "Trends in steganography," *Commun. ACM*, vol. 57, no. 3, pp. 86–95, Mar. 2014.

[3] T. Pevný, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," in *Proc. Int. Workshop Inf. Hiding*. Berlin, Germany: Springer, 2010, pp. 161–177.

[4] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2012, pp. 234–239.

[5] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP J. Inf. Secur.*, vol. 2014, Jan. 2014, Art. no. 1.

[6] M. M. Sadek, A. S. Khalifa, and M. G. Mostafa, "Video steganography: A comprehensive review," *Multimedia Tools Appl.*, vol. 74, no. 17, pp. 7063–7094, Sep. 2015.

[7] J. Yang and S. Li, "An efficient information hiding method based on motion vector space encoding for HEVC," *Multimedia Tools AND Appl.*, vol. 77, no. 10, pp. 11979–12001, May 2018.

[8] F. Djebbar, B. Ayad, K. A. Meraim, and H. Hamam, "Comparative study of digital audio steganography techniques," *Eurasip J. Audio Speech Music Process.*, vol. 2012, Oct. 2012, Art. no. 25.

[9] G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. L. Thing, "Twenty years of digital audio watermarking—A comprehensive review," *Signal Process.*, vol. 128, pp. 222–242, Nov. 2016.

[10] E. Satir and H. Isik, "A Huffman compression based text steganography method," *Multimedia Tools Appl.*, vol. 70, no. 3, pp. 2085–2110, Jun. 2014.

[11] C. Chang and S. Clark, "Practical linguistic steganography using contextual synonym substitution and a novel vertex coding method," *Comput. Linguistics*, vol. 40, no. 2, pp. 403–448, Jun. 2014.

[12] S. Zander, G. Armitage, and P. Branch, "Covert channels and countermeasures in computer network Protocols," *IEEE Commun. Mag.*, vol. 45, no. 12, pp. 136–142, Dec. 2007.

[13] J. Lubacz, W. Mazurczyk, and K. Szczypiorski, "Principles and overview of network steganography," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 225–229, May 2014.

[14] G. Bernd and V. Peter, "High rate data hiding in ACELP speech codecs," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Mar./Apr. 2008, pp. 4005–4008.

[15] A. Nishimura, "Data hiding in pitch delay data of the adaptive multi-rate narrow-band speech codec," in *Proc. Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Sep. 2009, pp. 483–486.

[16] H. Miao, L. Huang, Z. Chen, W. Yang, and A. Al-Hawbani, "A new scheme for covert communication via 3G encoded speech," *Comput. Elect. Eng.*, vol. 38, no. 6, pp. 1490–1501, Nov. 2012.
[17] S. Yan, G. Tang, and Y. Chen, "Incorporating data hiding into G.729 speech codec," *Multimedia Tools Appl.*, vol. 75, no. 18, pp. 11493–11512, Sep. 2016.
[18] Y. Ren, H. Wu, and L. Wang, "An AMR adaptive steganography algorithm based on minimizing distortion," *Multimedia Tools Appl.*, vol. 77, no. 6, pp. 12095–12110, May 2017.
[19] Y. Huang, C. Liu, S. Tang, and S. Bai, "Steganography integration into a low-bit rate speech codec," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1865–1875, Dec. 2012.
[20] B. Xiao, Y. Huang, and S. Tang, "An approach to information hiding in low bit-rate speech stream," in *Proc. Global Telecommun. Conf.*, Nov./Dec. 2008, pp. 1–5.
[21] Y. Huang, H. Tao, B. Xiao, and C. Chang, "Steganography in low bit-rate speech streams based on quantization index modulation controlled by keys," *Sci. China Technol. Sci.*, vol. 60, no. 10, pp. 1585–1596, Oct. 2017.
[22] H. Tian, J. Liu, and S. Li, "Improving security of quantization-index-modulation steganography in low bit-rate speech streams," *Multimedia Syst.*, vol. 20, no. 2, pp. 143–154, Mar. 2014.
[23] P. Liu, S. Li, and H. Wang, "Steganography integrated into linear predictive coding for low bit-rate speech codec," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 2837–2859, Jan. 2017.
[24] P. Liu, S. Li, and H. Wang, "Steganography in vector quantization process of linear predictive coding for low-bit-rate speech codec," *Multimedia Syst.*, vol. 23, no. 4, pp. 485–497, Jul. 2017.
[25] L. Liu, M. Li, Q. Li, and Y. Liang, "Perceptually transparent information hiding in G.729 bitstream," in *Proc. Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Aug. 2008, pp. 406–409.
[26] Z. Wu, H. Cao, and D. Li, "An approach of steganography in G.729 bitstream based on matrix coding and interleaving," *Chin. J. Electron.*, vol. 24, no. 1, pp. 157–165, Jan. 2015.
[27] S. Yan, G. Tang, Y. Sun, Z. Gao, and L. Shen, "A triple-layer steganography scheme for low bit-rate speech streams," *Multimedia Tools Appl.*, vol. 74, no. 24, pp. 11763–11782, Dec. 2015.
[28] T. Xu and Z. Yang, "Simple and effective speech steganography in G.723.1 low-rate codes," in *Proc. Int. Conf. Wireless Commun. Signal Process.*, Nov. 2009, pp. 1–4.
[29] Y. F. Huang, S. Tang, and J. Yuan, "Steganography in inactive frames of VoIP streams encoded by source codec," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 296–306, Jun. 2011.
[30] J. Liu, K. Zhou, and H. Tian, "Least-significant-digit steganography in low bitrate speech," in *Proc. Int. Conf. Commun.*, Jun. 2012, pp. 1133–1137.
[31] R. S. Lin, "An imperceptible information hiding in encoded bits of speech signal," in *Proc. Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Sep. 2016, pp. 37–40.
[32] X. Peng, Y. Huang, and F. Li, "A steganography scheme in a low-bit rate speech codec based on 3D-sudoku matrix," in *Proc. IEEE Int. Conf. Commun. Softw. Netw.*, Jun. 2016, pp. 13–18.
[33] Q. Ding and X. Ping, "Steganalysis of compressed speech based on histogram features," in *Proc. 6th Int. Conf. Wireless Commun. Netw. Mobile Comput. (WiCOM)*, Sep. 2010, pp. 1–4.
[34] Q. Ding and P. Xijian, "Steganalysis of analysis-by-synthesis compressed speech," in *Proc. Int. Conf. Multimedia Inf. Netw. Secur.*, Nov. 2010, pp. 681–685.
[35] H. Miao, L. Huang, Y. Shen, X. Lu, and Z. Chen, "Steganalysis of compressed speech based on Markov and entropy," in *Proc. 12th Int. Workshop Digit. Watermarking*. Berlin, Germany: Springer, 2014, pp. 63–76.
[36] H. Tian, Y. Wu, Y. Huang, J. Liu, Y. Chen, T. Wang, and Y. Cai, "Steganalysis of low bit-rate speech based on statistic characteristics of pulse positions," in *Proc. Int. Conf. Availability, Rel. Secur.*, Aug. 2015, pp. 455–460.
[37] Y. Ren, T. Cai, M. Tang, and L. Wang, "AMR steganalysis based on the probability of same pulse position," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 9, pp. 1801–1811, Sep. 2015.
[38] H. Tian, Y. Wu, C.-C. Chang, Y. Huang, Y. Chen, T. Wang, Y. Cai, and J. Liu, "Steganalysis of adaptive multi-rate speech using statistical characteristics of pulse pairs," *Signal Process.*, vol. 134, pp. 9–22, May 2017.
[39] H. Tian, J. Sun, Y. Huang, T. Wang, Y. Chen, and Y. Cai, "Detecting steganography of adaptive multirate speech with unknown embedding rate," *Mobile Inf. Syst.*, vol. 2017, May 2017, Art. no. 5418978.
[40] S. Li, Y. Jia, J. Fu, and Q. Dai, "Detection of pitch modulation information hiding based on codebook correlation network," *Chin. J. Comput.*, vol. 37, no. 10, pp. 2107–2117, Oct. 2014.
[41] Y. Ren, J. Yang, J. Wang, and L. Wang, "AMR steganalysis based on second-order difference of pitch delay," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 6, pp. 1345–1357, Jun. 2017.
[42] S.-B. Li, H.-Z. Tao, and Y.-F. Huang, "Detection of quantization index modulation steganography in G.723.1 bit stream based on quantization index sequence analysis," *J. Zhejiang Univ. SCI. C*, vol. 13, no. 8, pp. 624–634, Aug. 2012.
[43] S. Li, Y. Jia, and C.-C. J. Kuo, "Steganalysis of QIM steganography in low-bit-rate speech signals," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 5, pp. 1011–1022, May 2017.
[44] J. Yang and S. Li, "Steganalysis of joint codeword quantization index modulation steganography based on codeword Bayesian network," *Neurocomputing*, vol. 313, no. 3, pp. 316–323, Nov. 2018.
[45] H. Tian, Y. Wu, Y. Cai, Y. Huang, J. Liu, T. Wang, Y. Chen, and J. Lu, "Distributed steganalysis of compressed speech," *Soft Comput.*, vol. 21, no. 3, pp. 795–804, Feb. 2017.
[46] Y. Huang, S. Tang, C. Bao, and Y. J. Yip, "Steganalysis of compressed speech to detect covert voice over Internet protocol channels," *IET Inf. Secur.*, vol. 5, no. 1, pp. 26–32, Mar. 2011.
[47] C. Kraetzer and J. Dittmann, "Mel-cepstrum-based steganalysis for VoIP steganography," *Proc. SPIE*, vol. 6505, Mar. 2007, Art. no. 650505.
[48] C. Kraetzer and J. Dittmann, "Pros and cons of mel-cepstrum based audio steganalysis using SVM classification," in *Proc. Int. Conf. Inf. Hiding*. Berlin, Germany: Springer, 2007, pp. 359–377.
[49] Q. Liu, A. H. Sung, and M. Qiao, "Temporal derivative-based spectrum and Mel-Cepstrum audio steganalysis," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 3, pp. 359–368, Sep. 2009.
[50] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 3, pp. 920–935, Sep. 2011.
[51] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

JIE YANG received the B.S. degree in electronic science and technology from Beijing Normal University, Beijing, China, in 2013, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, in 2018. He has been an Assistant Professor with the Jiyang College, Zhejiang A & F University, Zhuji, China, since 2018.

His current research interests include multimedia signal processing and data hiding.

PENG LIU received the B.S. degree in communication engineering from Hainan University, in 2011, and the Ph.D. degree from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, in 2016, where he has been an Associate Professor, since 2018.

His current research interest includes information forensics.

SONGBIN LI received the Ph.D. degree from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, in 2010. He was a Postdoctoral Fellow and a Visiting Professor with Tsinghua University and the University of Southern California, respectively. He has been a Professor with the Institute of Acoustics, Chinese Academy of Sciences, since 2018. He has been the Principle Investigator on several projects of the National Natural Science Foundation of China.

His current research interests include machine learning, multimedia signal processing, and information forensics.