ISSN : 0974 - 7435 Volume 7 Issue 10

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 7(10), 2013 [379-385]

Robust information hiding and extraction algorithms in speech

Bao Yongqiang[1]*, Xi Ji[2], Xu Haiyan[2]

[1]School of Communication Engineering, Nanjing Institute of Technology, 211167, Nanjing, (CHINA)

[2]School of IOT Engineering, Hohai University, 213022, Changzhou, (CHINA)

ABSTRACT

Speech with hidden data will be disturbed and damaged by a variety of interference, such as noises, codec and filters, etc. To improve the robustness, the speech information hiding and extraction algorithm based on PSO-NN (Particle Swarm Optimizer Neural Network) is proposed. To improve the performance of anti-channel interference, the algorithm adds redundant data into the hidden data and then trains at the decoding end. At the same time, to improve the training efficiency and decoding accuracy, the algorithm firstly uses wavelet decomposition to get high-frequency coefficients of the signal, and then calculates the characteristic of high-frequency coefficients. At last, the algorithm selects 32 optimal features to train the neural network based on the FDR (Fish Discriminant Ratio). Simulation results show that the proposed algorithm improves the robustness of speech information hiding approach against filtering attack, noise attack, sampling attack and compression attack. Though the improvement on tensile attacks is ineffective, it was also better than others neural network algorithm.

KEYWORDS

Speech hiding;
PSO-NN;
Fish discriminant ratio;
Wavelet decomposition.

INTRODUCTION

Speech with hidden data would be disturbed and damaged by a variety of interference, such as noises, codec and filters, etc. How to extract the hidden data under these interferences was one of the major difficulties in speech hiding algorithms. Under the interferences, speech with hidden data would lose some information more or less[1]. To improve this situation, speech hiding algorithm based on neural network was a good choice. Compared with the traditional algorithms, it had three advantages: 1) Higher flexibility and robustness; 2) Neural network could identify more status of hidden data with one coefficient to increase the amount of embedded data; 3) Parallel matrix operation resulted in high data extraction speed.

Currently, in the aspects of information hiding of neural networks, some scholars had a targeted research. In image field, Yu[2] embedded image data in time domain and used the cross neighborhood window to calculate neural networks and realize the signal blind detection. In addition, Hung-Hsu Tsai combined human visual characteristics with neural network to achieve image hiding algorithm[3]. In the aspects of audio information hiding, Jinyan Hu[4] studied the blind detection of audio information hiding based on neural networks. Cheng[5] applied neural network to fragile information technology. Hung-Hus Tsai combined human auditory

characteristics with neural network to improve anti-aggressive, and then maintained the naturalness of speech with embedded hidden data[6,7].

To improve the performance of anti-channel interference, an information hiding and extraction algorithm based on PSO-NN was proposed. The algorithm added redundant data into the hidden data and then trained at the decoding side. At the same time, to improve the training efficiency and decoding accuracy, the algorithm used wavelet decomposition to get high-frequency coefficients of signal, and then calculated the characteristic of high-frequency coefficients. At last, the algorithm selected 32 optimal features based on FDR to train the neural network. Comparative experiments of decoding in different neural networks showed that the proposed algorithm had strong anti-interference performance and high decoding rate.

MODEL OF SPEECH INFORMATION HIDING AND EXTRACTION BASED ON NN

Other than traditional coefficient quantization methods, blind information hiding algorithm based on neural work theory used learning samples to train neural networks, and then extracted data from trained networks to realize blind detection algorithm. Therefore, to use neural networks to realize blind detection, while real data was embedded, some training samples was also embedded to ensure that the receiver could build a neural network to extract useful information. PSO[8] was an evolutionary computation technique based on swarm intelligence proposed in recent years. Compared with other evolutionary algorithms, PSO had simple, easy implement and stronger global optimization capacity.

In the data embedded model based on neural networks, sample data generally had the same properties as embedded ones. e.g., image and text data were processed in the same way. The purpose to reduce dimensions in the figure was to make two-dimensional image into one-dimension one, thus they could be embedded into audio data. Regularization was to adjust the dynamic range of data to the same audio data. From practicality, data hiding generally took section as a unit, which meant that each section would embed one bit or a set of data.

The data extraction process was generally the reverse process of embedding. The entire extraction process could be divided into 5 steps: subsection, matrix generation, neural network training, data extraction and judgment, embedded data recovery. Throughout the process, neural network training and data extraction and judgment were much more important, which determined the algorithm efficiency directly. And the most critical contents involved were neutral network selection, parameter settings and selection criteria and implementation strategies of the threshold.

Embedding algorithm of binary image

The image used in this algorithm was binary image. To simulate binary image and train neural network, pixels were constructed and two-dimensional image were reduced dimension and normalized. And then the binary image was embedded into audio information.

Firstly, supposed that there were M pixels in the image, the two-dimensional image was reduced dimension and normalized to become a one-dimensional sequence with length of M. Then the pixels were constructed normalized. The constructed pixels and image sequence merged into one dimensional image were called data merging. Finally, the audio signal was devised into $N+M$ segments and each segment was embedded into one element of the merged sequence until all the data were embedded. The diagram of the image data embedding were shown as Figure 1.

(1) Pixels construction

Supposed the constructed pixel was $\{p(i)\}$, and the value of $p(i)$ was 0 or 1. Then construct PN sequence with length of N, which would be used for synchronization.

(2) Dimension reduction of binary image

Supposed the binary image $\{wm\}$ was a two-dimension matrix of $m \times m$. It would become a one-dimensional sequence by dimension reduction. Supposed that the one-dimension sequence after dimension reduction was $\{wm'\}$, the specific method was:

$$wm'(m(i-1)+j) = wm(i, j) \quad 1 \le \mathrm{i,j} \le \mathrm{m} \qquad (1)$$

(3) Normalization

Here, the non-binary image was mainly processed, such as Q-level gray-scale image. The Q-level gray-

scale image had Q grey levels, so the values were between 0 and Q-1. If the pixel values were directly embedded into the audio, it would cause two problems: 1) The range of audio was [-1,1], so the pixel values of image were higher than magnitude value of audio; 2) The embedding algorithm was additive, and the pixel values were positive. If continued adding, the whole electrical level of audio would rise and influence the audio embedded with the image. So it needed to normalize the pixels embedded.

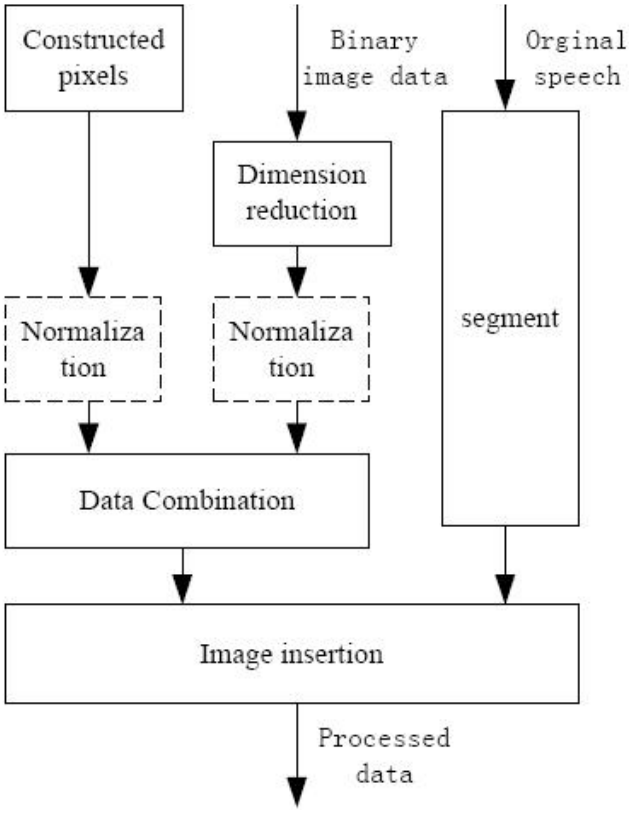

Figure 1 : Diagram of image data embedding

Constructed pixel $\{p\}$ and the image after dimension reduction $\{q\}$ were merged into a one dimension sequence $\{r\}$. Supposed the sequence after normalization was $\{r'\}$, and the specific method was:

$$r'(i) = [r(i) - (Q/2)]/(Q/2) \quad 1 \le i \le N + m^2 \qquad \textbf{(2)}$$

(4)Embedding of binary image

Considered the time of calculation, the image information was embedded into audio signal in the time domain. Divided the signal S into $N + m^2$ segment, and supposed the length of each section was L and the ith section was s_i.

Encoder was used to add the watermark information $p'(i)$ to the speech signal s_i. The processed signal was s_i' and its mathematical expression was:

$$s_i' = s_i + p_j' f(s_i, k, \lambda) = s_i + p_j' k_i, i = 1, ..., (L \times T)$$
$$j = 1, ..., L$$

Here, $p_j' \in \pm 1 \ \ j = 1, ..., L$ represented the embedded hidden data. Compared with unipolar encoding, dual-polarization encoding had higher efficiency at the encoder end. L represented the length of watermark. Watermark data could be inserted repeatedly if necessary, and T represented the period of insertion. K was the generated key and was not relevant to the signal S. $f(s_i, k, \lambda)$ was a nonlinear function of generated key sequence k_i, and λ was used to control the parameters of embedded watermark energy. Here, $p_j' k_i$ represented the embedded watermark signal.

Extraction algorithm of binary image

The audio embedded image would be divided into N+M sections, assuming the length of each section was L. Take the first N sections as the training sequence, and then calculate each section to get 74 features for network training. Select 32 most effective features based on dimension reduction algorithm as network training features. Then begin to train the network, take features of the train sequence as an input and constructed pixels as desired output. After the training, the network could be used to extract image. Then the 74 features of the rest M audio sections were taken as input vector of neural network. After a series processing of threshold judgment and dimension rising, the image data could be restored from the output of neural network. The principle of image data extraction was showed in Figure 2.

For the process of binary image extraction, the audio information embedded binary image was divided into training sequence and extraction sequence firstly. Secondly, the training sequence was used to train the neural network. And finally the binary image could be restored from the audio information embedded binary image after a series processing of threshold judgment, restoring and dimension rising.

The problem of speech information hiding was usu-

FULL PAPER

ally the problem of hypothesis testing. Two hypotheses H_0, H_1 and the sub-hypothesis of H_1 could be defined as:

H_0 : Hidden data did not exist in the voice
H_1 : Hidden data existed in the voice
H_{1a} : Hidden data was +1
H_{1b} : Hidden data was -1

Detection and extraction of hidden data could be achieved at the same time via neural network.

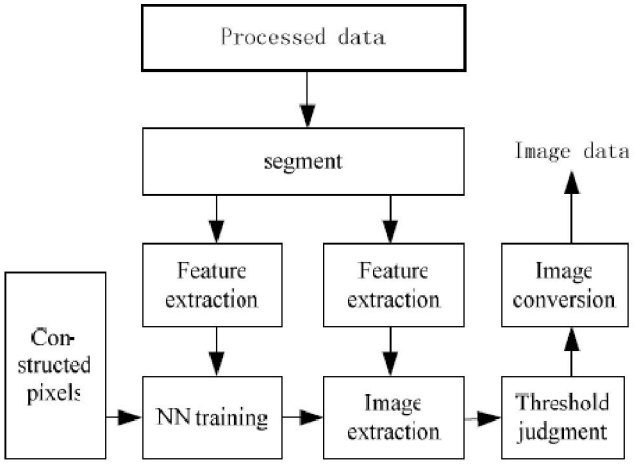

Figure 2 : Diagram of image data extraction

(1) Feature extraction

After wavelet transform, 74 statistic features of the segmented speech were extracted as training features[9]. The reason of wavelet decomposion used to extract high-frequency coefficients were: Firstly, on the judgment of whether watermark existed or not, wavelet decomposion had much higher efficiency than using voice sequence directly. After wavelet processing, the characteristic distribution of high frequency wavelet coefficients was easy for data classification. Secondly, considering the perceptual characteristics of the human ear, most of the watermark information was high-frequency information, so the detail coefficients could be extracted as useful sequence.

(2) Feature dimension reduction

Limited by the size of training samples, the dimension of feature space could not be too high and needed to reduce the dimension. Characteristic analysis method based on FDR obtained good results on the research of HIV virus features[10] and fear emotion features[11]. FDR was defined as:

$$FDA = \sum_{i}^{M}\sum_{j\neq i}^{M}\frac{(\mu_i-\mu_j)^2}{\sigma_i^2+\sigma_j^2} \quad (3)$$

Where in, M was the number of categories, i, j were the category number, μ was the categories center (that is, the feature vectors of categories), and σ^2 was the variance of corresponding sample. Sort the calculated FDR values and then select the top 32 features.

(3) Neural network training

Generate a constructed pixel {p}with the embedded image algorithm, input the train matrix into neural network and then get the desired output {p}. Stop the training process when the error was less than δ.

(4) Binary image extraction

When the neural networks completed the training process, the matrix could be used to extract image. Input the extract matrix into the trained neural network to get the output.

(5) Threshold judgment

Implement the threshold judgment of to generate a new sequence{y'}.

$$\begin{cases} y'(i)=0, & y(i)<-0.3 \\ y'(i)=+1, & y(i)\geq 0.3 \\ \text{No watermark}, & |y(i)|<0.3 \end{cases} \quad (4)$$

(6) Image conversion

Revert to the binary image, the matrix representation was $Y_{m\times m}$.

$$Y(i,j)=y'(m(i-1)+j) \quad (5)$$

EXPERIMENT AND SIMULATION

The data, which included 65 male, 60 female voices of the 10 paragraphs Chinese Corpus, used in this experiment came from the language lab with ITU standard. The audio segment used in the paper was a piece of female voice with the content of "My laughter made the house lively, and my joy fulfilled every room, even the winter was swept out of the door. I danced, jumped and embraced everyone." The sampling frequency was 11.025 KHz, the sampling depth was 16 bits and the duration was 14 seconds.

To test the robustness of the algorithm, the attack strategy selected was from STEP2001, which included filtering attack, noise attack, sample attack, dynamic range compression attack and tensile attack.

PSO neural network had 32 nodes at input layer, 1 node at output layer and 16 neurons at hidden layer. Daubechies-4 wavelet was used for wavelet decomposition.

To reflect the performance of the proposed algorithm, binary gray images with complex seal patterns was selected. Figure 3 is the test image and the size was 64x64.

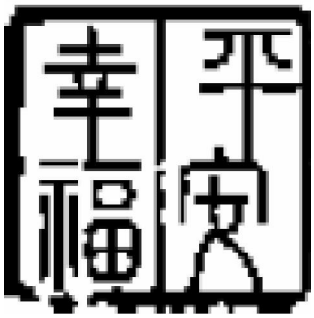

Figure 3 : Embedded speech data

Taking Parity-6bit as an example, the performance of PSO neural network algorithm was validated by comparing with BP neural network and Elman neural network. Results showed that: It only took 1.95 seconds with 5 times for training of PSO neural network to meet the error threshold, while BP neural networks took 3.22 seconds with 11 times. Although Elman networks took only 0.42 second with zero training error, its number of hidden layers was 64. So compared with BP networks, PSO had faster speed in network convergence; and compared with Elman networks, although the performance of convergence wasn't better, the network structure was much more simple. Without special emphasis on convergence time, PSO network had the best overall performance.

The evaluation criterion was error rate.

$$P_e = S_{err} / S_{all} \quad (6)$$

Here, was the number of erroneous data points, was the total number of the data points.

Comparison of algorithm performance under different attacks

To test the robustness of the algorithm, the comparing methods included traditional autocorrelation decision algorithm based on matched filter and three neural network algorithms.

Five kinds of attacks were set as follows:

(1) Filter attack (A1)

Filter attack was to extract hidden information by low-pass filtering the hidden signal through FIR filter. The experiment used a cutoff frequency of 0.1 * fs / 2.

(2) Noise attack (A2)

Noise attack simulation was achieved by embedding white noise of a certain SNR into hidden signal. Here the SNR was 40dB.

(3) Sampling attack (A3)

Sampling attacks was to down-sample the hidden signals twice and up-sample with the original frequency, then extract the secret information.

(4) Dynamic range compression attack (A4)

Compressed the 16 bit sampling depth of hidden signals to 8bit.

(5) Tensile attack (A5)

Stretched steganographic signals by -10% through extraction.

The results of 5 methods were shown as Figure 4. From the attack modes, the error rate of sampling and compression attacks was lower, while the interference of tensile attacks on the hidden data extraction was the largest. From the figure, the effect of autocorrelation method was worse and its error rate was higher than algorithms based on neural networks under different attacks. For the filter attack, the algorithm based on PSO-NN had the highest robustness. By comparing with the error rate of autocorrelation method, the neural network had the best performance under noise attacks and worst performance under tensile attacks.

Statistical results

The experiment compared the hiding effect of 100 pieces of speech with attacks, the statistical results was shown in TABLE 1. From TABLE 1, three speech hiding and extraction algorithms based on neural networks had better robustness than autocorrelation method when facing attacks especially under 40dB white noise attacks, the error rate of autocorrelation method reached 43.8%, while the error rate of algorithms based on neural networks were less than 8%. Among the 3 algo-

FULL PAPER

rithms based on neural networks, PSO-NN had the best performance with the error rate of 5%~10%, which was lower than other 2 algorithms. However, the table also showed that under tensile attacks, all the 4 algorithms had poor performance and needed further improvement.

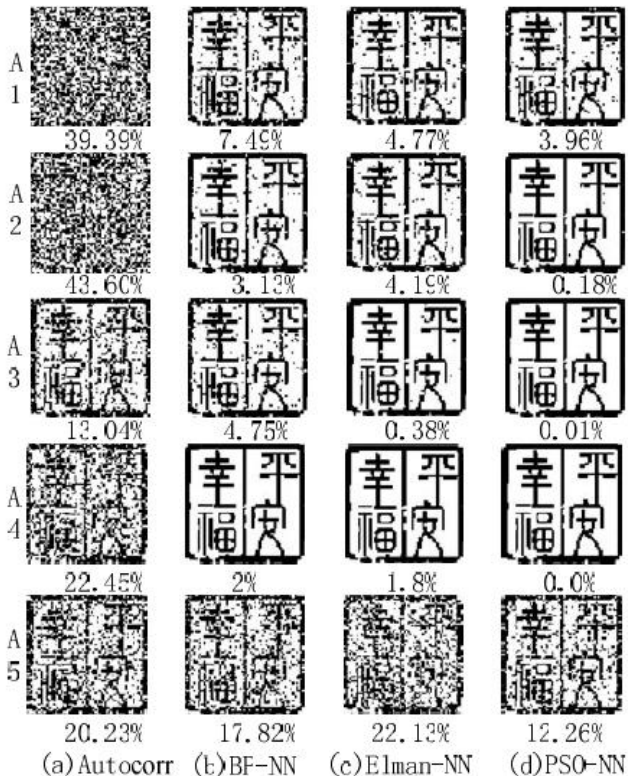

Figure 4 : diagram of anti-attack effect

TABLE 1 : Comparison of decoding error rate under different attacks

Types	Decoding methods (%)			
	Autocorr	BP	Elman	PSO
Original	10.78	4.46	5.64	1.59
Filter	44.84	6.89	5.85	0.62
Sampling	17.71	7.40	8.71	4.07
Noise	43.81	7.65	7.02	0.18
Compression	21.13	14.36	4.62	0.24
Tensile	18.30	26.82	21.45	22.66

Performance comparison of pre-training and post-training

In this paper, a speech hidden data extraction method of post-training was used, that is, redundant data was added at the encoder end while performed training and extracting at the receiver end. Another way was to implement training while coding, and the receiver just input the received data into trained neural networks to get the hidden data. Considering the effects of transmission channel on data, Pre-training methods usually failed, so the decoding efficiency was often affected. Figure 5 was the performance comparison of two decoding methods. From the figure, the encoding error rate of post-training was 10% lower than that of pre-training.

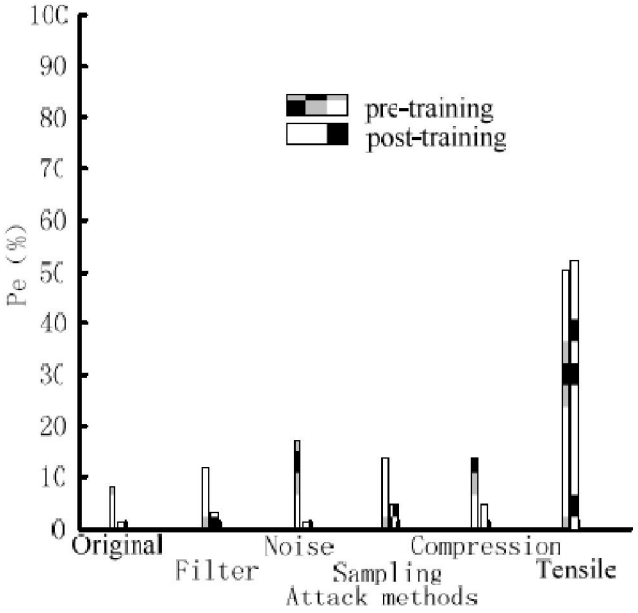

Figure 5 : Performance comparison of pre-training and post-training

CONCLUSION

This paper had a research on the speech hiding and decoding algorithms based on neural networks, and then designed a speech data hiding and extraction algorithm based on PSO-NN. To improve the performance of anti-channel interference, the algorithm added redundant data into the hidden data and then trained at the decoding end. The data features were optimized to improve the efficiency of training and PSO-NN was used to improve the decoding efficiency. In the comparison experiment of BP neural networks, Elman neural networks and the autocorrelation algorithm, the proposed algorithm under different attacks like filter attack, noise attack, sampling attack and compression attack had the best robustness. However the improvements against tensile attack were less effective, which needed to be

further studied. Besides, in this paper only the embedding method of binary images was considered. Next, grayscale images and other large data embedding method and decoding method should be studied.

ACKNOWLEDGMENTS

This project was supported by China Postdoctoral Science Foundation (No. 2012M520973), the Scientific Research Funds of Nanjing Institute of Technology (No. ZKJ201202) and the underwater acoustic signal processing open research fund of Southeast University of Key Laboratory of Ministry of Education (B) (No. UASP1202).

REFERENCES

[1] W.Jia, F.P.Tso, Z.Ling, X.Fu, D.Xuan,W.Yu; Blind detection of spread spectrum flow watermarks[J]. Security and Communication Networks, **6(3)**, 257-274 **(2013)**.

[2] P.T.Yu, H.H.Tsai, J.S.Lin; Digital watermarking based on neural networks for color images[J]. Signal processing, **81(3)**, 663-671 **(2001)**.

[3] H.H.Tsai, C.C.Liu; Wavelet-based image watermarking with visibility range estimation based on HVS and neural networks[J]. Pattern Recognition, **44(4)**, 751-763 **(2011)**.

[4] Hu Jinyan,Zhang Taiyi,Lu Congde,Zhang Chunmei; Audio Watermarking with Neural Networks in the Wavelet Domain [J]. Journal of XiAn Jiaotong University, **37(4)**, 355-358 **(2003)**.

[5] Y.C.Fan, W.L.Mao,H.W.Tsao; An artificial neural network-based scheme for fragile watermarking. Consumer Electronics, 2003. ICCE. 2003 IEEE International Conference on[C].IEEE, **(2003)**.

[6] H.H.Tsai, J.S.Cheng,P.T.Yu; Audio watermarking based on HAS and neural networks in DCT domain[J]. EURASIP Journal on Applied Signal Processing, **3**, 252-263 **(2003)**.

[7] H.H.Tsai, J.S.Cheng; Adaptive signal-dependent audio watermarking based on human auditory system and neural networks[J]. Applied Intelligence, **23(3)**, 191-206 **(2005)**.

[8] B.Soudan, M.Saad; An evolutionary dynamic population size PSO implementation.Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on[C].IEEE, **(2008)**.

[9] Yu Hua, Huang Chengwei, Jin Yun, Zhao Li; Speech Emotion Recognition Based on Particle Swarm Optimizer Neural Network [J]. Journal of Data Acquisition & Processing, **26(001)**, 57-62 **(2011)**.

[10] T.H.Lin, H.T.Li, K.C.Tsai; Implementing the Fisher's Discriminant Ratio in ak-Means Clustering Algorithm for Feature Selection and Data Set Trimming[J]. Journal of chemical information and computer sciences, **44(1)**, 76-87 **(2004)**.

[11] C.Clavel, I.Vasilescu, L.Devillers, G.Richard, T.Ehrette; Fear-type emotion recognition for future audio-based surveillance systems[J]. Speech Communication, **50(6)**, 487-503 **(2008)**.