Enhancement Artificial Neural Networks for Low-Bit Rate Speech Compression system

J. Srinonchat
Signal Processing Research Laboratory
Faculty of Engineering
Rajamangala University of Technology Thunyaburi, Phathumtani, Thailand, 12110
jakkree@rmut.ac.th

Abstract— An Artificial Neural Networks (ANNs) is the intelligent system which has been recently exploited in linear and non-linear system such as image and speech processing. In this work, there are two types of neural networks, namely Kohonen Self Organizing Feature Maps (KSOFM) and Probabilistic Neural Networks (PNNs), which are investigated to use in CELP speech coding system. KSOFM is used to classify the repetitiveness of speech signal and create the optimal codebook and PNNs is applied to predict the codebook index by using the knowledge of training system. The results show that the neural index prediction can reduce the number of bit rate approximately 25% while maintains the quality of the synthesized speech as similar as the original speech.

1. INTRODUCTION

An Artificial Neural Networks (ANNs) [1] is the artificial intelligent system which operates as parallel distributed information processing system. They are composed of a large number of simple processing units, namely neuron, which each neuron is connected and created to be a network. The ANNs has been developed and applied to perform complex computational tasks such as in linear and nonlinear system [2]. The speech processing is a nonlinear system which the ANNs has been applied for many proposes such as word recognition system or compression speech signals.

Speech coding or compression is essentially technique for communication system which obviously uses in many application such as public switched telephone network (PSTN) which uses pulse code modulator (PCM) technique to operate at 64 kbit/s, the GSM half rate which uses vector sum excited linear predictive (VSELP) codec technique to operate at 6.7 kbit/s. The ANNs has also been used in some of speech coding system areas such as non-linear prediction based on ANN in speech coding [3] which applied General Radical Basic Function (GERBF) neural network to the speech coding system. The results provide the high segmental SNR for the speech coding system. The another work on neural network based arithmetic coding for real time audio transmission on the TMS320C6000 DSP platform [4] also achieved to function the ANNs in the real time DSP chip.

The goal of speech coding is to represent speech in digital form with as few bits as possible while maintaining the intelligibility and quality required for the particular application. This work is interested to use the ANNs to accomplish the goal of speech coding system. There are two ANNs types, namely Kohonen Self Organizing Feature Maps (KSOFM) and Probabilistic Neural Networks (PNNs) are applied to classify and predict the speech signal coefficients based on the code excited linear prediction (CELP) coder [5]. The paper is organized as follow: the speech signal is discuss in section 2; the ANNs are introduced in section 3; the speech coding based on ANNs is presented in section 4; the experimental and results are shown in section 5 and the conclusion is in section 6.

2. SPEECH SIGNAL

The requirements for the compression of a speech signal have been sought after in most main speech coding research centers, and as a result many different strategies for the suitable compression of speech for bandwidth-restricted applications have been developed. The exploitation of bit rate speech coders have been standardized in many international and national communication systems [6]. However these applications of speech compression systems are in the field of Speaker Independent Coding Systems (SICS), which does not specifically operate for any individual speaker. A SICS can be used by any speaker. The disadvantage of SICS is that the compression of the speech signal has to be made general for the voice characteristics of different speakers. This work is mainly concerned with Speaker Dependent Coding Systems (SDCS), which have the ability to gain knowledge about a particular speaker by exploiting their speech waveform characteristics to gain a reduced bit rate. Thus speech signal is the majority part to be considered for design the compression algorithm.

Speech is the continuous and non-linear signal which is basically produced by the physical nature of the human body and is substantially shaped by the vocal tract. Thus the properties of speech signal are the positions and movements of the vocal tract over time and also depend on the characteristic of the speaker. Different speech signals are produced from both changes in the excitation signal and the movement of positions of the vocal tract. This movement is individual to each speaker. There is the opportunity for the trajectory of

vocal tract positions to repeat. Speech coding system is required to model the speech signal in the format of parameters which, in this paper, the line spectral pairs (LSP) technique had been used to model the speech signal.

LSP is the mathematical model [7, 8] which represents of natural resonant frequencies of vocal tract. If $H(z)$ is the transfer function of a digital filter of vocal tract which can be defined as

$$H(Z) = \frac{1}{A(Z)} = \frac{1}{1 - \sum_{k=1}^{p} a_p Z^{-k}} \qquad \text{.....(1)}$$

which $H(z)$ is the transfer function of a digital filter as refers to all-pole system. However, a general transfer function of a real vocal has both poles and zeros. By this reason, LSP technique is map the $A(z)$ into other equivalent polynomials to represent the all pole and zero in that transfer function. $A(z)$ is decomposed into both symmetric and an antisymmetric polynomial by adding and subtracting the time reversed system function as follows:

$$P(z) = A(z) + z^{-(k+1)} A(z^{-1}) \qquad (2)$$

$$Q(z) = A(z) - z^{-(k+1)} A(z^{-1}) \qquad (3)$$

The roots of two polynomials (2) and (3) are the LSP. If $A(z)$ is minimum-phase, then the zeros of $P(z)$ and $Q(z)$ are interlaced each with other. This property allows verification of minimum-phase status and hence the stability of synthesis filters.

3. ARTIFICIAL NEURAL NETWORK

ANNs is a massively parallel distributed processor made up of simple processing units known as neuron. These neurons are organized in layers and every neuron in each layer is connected by weight to each neuron in the adjacent layers. There are two types of ANNs, namely Kohonen Self Organizing Feature Maps (KSOFM) and Probabilistic Neural Networks (PNNs), which investigated to use in this work.

A. KSOFM

The KSOFM algorithm [9] basically belongs to the class of neural networks, utilizing unsupervised learning algorithm. It is able to learn to classify input codevectors into groups. The block diagram of the KSOFM process is shown in Fig. 1.

In the training process, the number of neurons is designed to give a network in which one neuron is expected to represent one of the data group. Once input codevectors are added to the neural networks, the distance or weight between the input and all neurons are calculated. The neuron closest to that input codevectors is selected as the "wining neuron" to represent the input codevector. The input is then categorized and located in the topology pattern. Thus similar input codevectors are placed in similar locations in which they might use the same neuron or its neighbor to be represented; this depends on how close the neuron is to that input in the training process. The training process in KSOFM is an iteration process. When the new input codevector is provided to the system, the training phase is started over again. Also, the weights of the winning neuron are adjusted or updated. The winning neuron then aligns its own weight vector with the training input and hence provides maximum response to the network again after training. Once the winning neuron has its weight updated, the weights of the neighboring neuron are also updated. The neighboring neuron is defined as the neuron most close to the winning neuron. The weights of the neighborhood neurons are updated weight in terms of their distance away from the winning neuron.

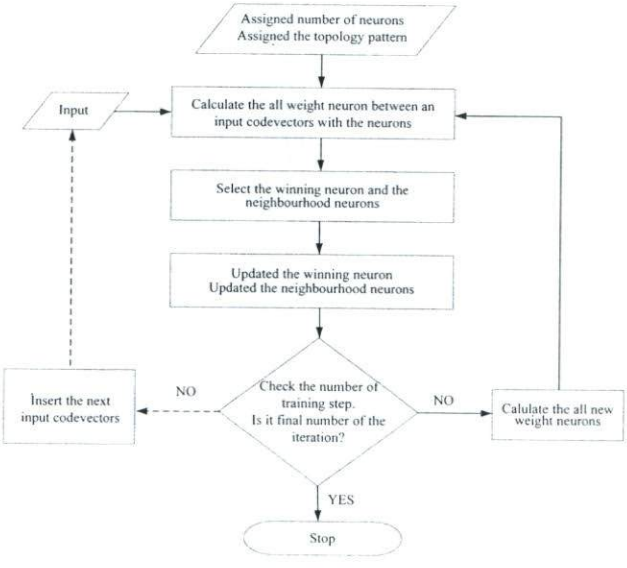

Fig. 1 Process of KSOFM

B. PNNs

The PNNs proposed by Specht [10] combines the characteristics of a feed-forward architecture and utilizes a supervised training algorithm similar to back propagation. Each training input pattern is used as the connection weights to a new hidden unit. In effect, each input pattern is incorporated into the PNNs architecture. The PNNs architecture is presented in Fig 2.

The network consists of four layers; an input layer, a pattern layer, a summation layer, and an output layer. When an input is presented, the neurons in the input layer allocate the inputs to the pattern neurons and then compute distances between the input and the training pattern. Then a vector is produced whose elements indicate how close the input is to a training input. These elements are then multiplied, element by element, by the bias which is the spread value. This spread determines the width of an area in the input space to which each neuron responds. However the only condition required

is to make sure that the spread is large enough such that the active input regions of the neuron overlap enough; several neurons always have fairly large outputs at any given moment. This makes the network function more smoothly and results in better generalization for new input vectors. However, the spread should not be so large that each neuron is effectively responding to the same large area of the input space.

The summation layer has one neuron for each class. This layer sums these contributions for each class of inputs to produce as its net output a pattern of probabilities. Each summation neuron that is dedicated to a single class, sums the pattern layer neuron corresponding to the numbers of that summation neuron's class. Finally, a transfer function on the output layer picks the maximum of these probabilities and produces the pattern for that input.

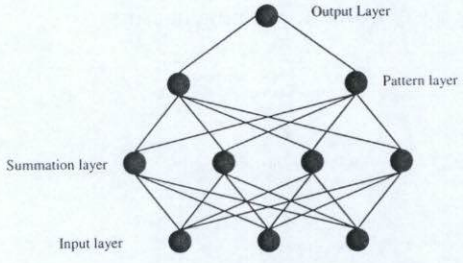

Fig. 2 the PNN architecture

4. SPEECH CODING BASED ON ANNs

The speech signal is separated into frame and each frame converted to the LSP parameters. Then the KSOFM is exploited to measure and classify the similar LSP characteristic into a codebook. This is because the system would like to reduce the number of the similar characteristic of speech signal and create in new sequence of the speech signal. Thus speech signal is now the sequence of the index codebooks.

The PNNs has been employed to predict the index codebook, exploiting the repetition of the combination indexes. Information on popular combination indexes are provided for the training process. This allows the PNNs to gain knowledge and then the PNNs is used as the neural index predictor to predict the incoming index.

The purpose of the neural index predictor is to reduce the number of actual codebook indexes transmitted, by exploiting previous and present codebook indexes to predict the forthcoming index, as shown in Fig 3. This can be achieved if the most common positions of consecutive codebook indexes can be learned.

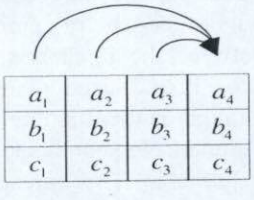

Fig. 3 The general structure of a prediction codebook indexes

In Fig 3, if a_1 to a_3 are defined as the training indexes and a_4 is the target index. The PNNs is trained to predict the accuracy target index. Subsequently, in the transmission, the a_1 to a_3 are transmitted to the receiver without the target index. Then the receiver uses the neural index predictor to predict the target index. Thus the neural index prediction can reduce the bit rate around 25 % according to the prediction process.

5. EXPERIMENT AND RESULTS

The speech data input is generated from four speakers: two males and two females. Each speaker contributed a total of 90 minutes of speech, of which 60 minutes was used to be initial learning system data and the final 30 minutes was used to be the testing system data. This experiment is tested in the CELP speech coding system which normally operates at 4.8 kbps. Thus it is expected to reduce the bit rate to be less than 4.8 kbps.

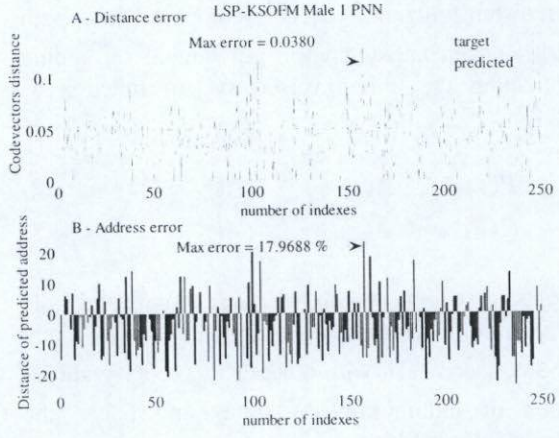

Fig. 4 Performance of Neural Index Prediction using PNN in Male 1

Fig 4 shows the results using the neural index predictor to predict the codebook indexes. This figure is examples of 250 predicted indexes in the performance of the different neural predictors using the LSP coefficients with the KSOFM. Each figure is divided into two parts; (A) the distance error and (B) the codebook index error. In A, the target distance is shown by the solid line and the predicted distance by the dotted line. It can be seen that using the PNN algorithm to be neural index predictor provided the max error of 0.038 in the LSP-KSOFM- Male 1.

The efficiency of performance of the neural index predictors was tested for the accuracy of prediction using the 30 minutes of test speech data. The testing process is similar to the initial training process, but the neural networks will not adjust weights for the coming input. Here, the neural index predictor will use knowledge from the training process to predict the target values. The results using the neural index predictors in Fig 5 and 6 show how close the synthetic speech is to the original speech • waveform, and also the error of speech

waveforms in frequency domain can be compared.

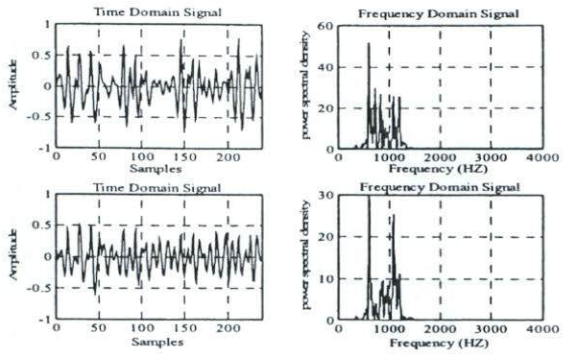

Fig. 5 Neural Index Predictor with LSP - KSOFM - PNN - Male 1

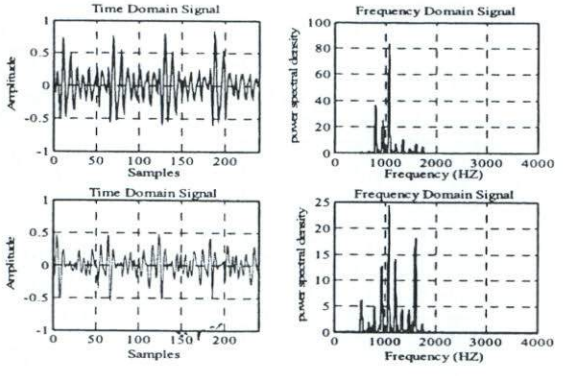

Fig. 6 Neural Index Predictor with LSP - KSOFM - PNN - Male 2

The results in Fig 5 and 6 show the performance of the neural index predictors in terms of the reconstruction of speech. Each figure contains of four small figures. Parts (A) and (B) show the synthesized speech from the CELP coder without the neural index predictor, and parts (C) and (D) show the synthesized speech from the neural index predictor based on the CELP coder.

It can be noticed that the synthetic speech using the neural predictor contains some frequency mistakes when compared to the CELP waveform. This error is distributed in the lower frequency range more than the higher frequencies. However the synthesized speech using neural predictors still maintained the majority of frequencies and pitch waveform characteristics of the original speech. Also, the energy in the synthetic speech is lower than in the original speech. This means that the synthesized speech has decreased loudness when compared to the CELP speech.

6. CONCLUSIN

It can be seen that the predicted indexes from the neural networks can be used to reconstruct the speech waveforms to be similar to the original speech. This means that the neural index predictor can possibly be used in the speech coding system. Therefore, it can be concluded that the neural index predictor can be used to reduce the number of bits required to transmit the codebook indexes by predicting every fourth index. This cancels the requirement to transmit this fourth index, and this is accomplished by having a duplicate set of weights in both the receiver and transmitter components of the system

ACKNOWLEDGMENT

I would like to thank to Dr. S. Danaher and Mr. J.I.H Allen for their suggestions and also Northumbria University, UK, for our research cooperation.

REFERENCE

[1] S. Haykin, *Neural Networks: Acomprehensive Foundation*, 2nd ed, Prentice-Hall, 1999, pp.1-256.

[2] M. bnkahla, "Applications of neuralnetworks to digital communications: a survey," *Signal Processing*, 2000, pp. 1185-1215.

[3] L. Li, Z. Sun, A. Wang and Z. Li, "Non-linear predictor based on ANN in speech coding," *International Conference on Control, Automation, Robots and Vision*, Vol 2, 2004, pp. 1098 - 1103

[4] E. Pasero and A. Montuori, "Neural network based arithmetic coding for real-time audio transmission on the TMS320C6000 DSP platform," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol 2, 2003, pp. 761-764.

[5] J. Srinonchat, S. Danaher, J.H. Allan, and A. Murray, "New Bit Rate CELP coder for Speaker Dependent Coding System," in *International Conference on Artificial Intelligence and Applications*. 2004.

[6] J. D. Gibson, "Speech coding methods standards and applications," IEEE circuits and systems magazine, Vol. 5, 2005, pp. 30 – 49.

[7] Hasegawa-Johnson, M., "Line spectral frequencies are poles and zeros of the glottal driving-point impedance of a discrete matched-impedance vocal tract model," *Journal of Acoustic Society of America*, 2000. 108: p. 457-460

[8] takula, F., "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signal," *Journal of Acoustic Society of America*, 1975. **57**: p. 535(A)

[9] Haykin, S., *Neural Networks: A comprehensive foundation*. 1999: Prentice Hall, pp. 1 -379.

[10] H. Deshuang and M. Songde, "A new radial basis probabilistic neural network model" *International Conference of Signal Processing*, 1996, pp. 1449 – 1452.