Speech Coding Based on a Multi-layer Neural Network

Shigeo Morishima, Hiroshi Harashima† and Yasuo Katayama††

Seikei University, †The University of Tokyo and
††Graphics Communication Technologies , Japan

Abstract

We present a speech compression scheme based on a 3-layer perceptron in which we reduce the number of units in the hidden layer. Input and output layers have the same number of units in order to achieve Identity Mapping.

Speech coding is realized by scalar or vector quantization of hidden layer outputs. By analyzing the weighting coefficients, it can be shown that speech coding based on a 3-layer neural network is studied speaker-independent.

We studied the relation between compression ratio and SNR at first. We give the bit allocation and optimum number of hidden layer units necessary to realize a specific bit rate. According to the analysis of weighting coefficients, speech coding based on neural network is transform coding similar to KL transformation.

Finally, we studied a special feature of mapping with 5-layer neural network which is more efficient than 3-layer neural network.

1. Introduction

This paper presents a trial to realize speech compression or speech coding based on a 3-layer perceptron at first. The learning process is performed by back propagation and a nonlinear function is sigmoid.

The three layers consist of input, output and hidden layers. The input and output layers have the same number of perceptron units to enable Identity Mapping. The number of units in the hidden layer is less than that of the input or output layers.

Speech coding is carried out by quantizing the output of the hidden layer. The transformation from input layer to hidden layer is known as coding process and that from hidden layer to output layer is known as decoding process.

Some trials of image compression and speech analysis based on a neural network have been already reported[2][3]. This paper shows the possibility of carrying out speech coding based on a neural network. We have decided an optimum compression ratio and bit allocation to obtain good speech coding at some bit rates.

Speech coding based on a 3-layer neural network is studied as a transform coding by analysis of weighting coefficients and this is only linear mapping from input to output. Non-linear mapping can be realized by 5-layer neural network. We show the special feature of mapping using 5-layer network, secondly.

2. System Feature

Figure 1 shows the structure of a speech compression system based on a neural network. This system is composed of a 3-layer perceptron and there are no links between the units in the same layer. Each unit is connected to the units in other layers by weighting coefficients. Input and output layers have the same number of units. This number indicates the frame length.

Speech signals from several male and female speakers are sampled at 8KHz and linearly quantized at 12bits. These speech samples are used as the input data or training data frame by frame. X is the input speech frame and Y is the output speech frame. n is frame length.

$$X=(x_1,x_2,x_3, \cdots\cdots\cdots,x_n)$$
$$Y=(y_1,y_2,y_3, \cdots\cdots\cdots,y_n)$$

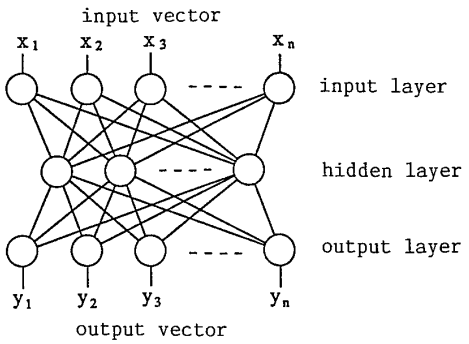


Figure 1 3-layer Neural Network

CH2829-0/90/0000-0429 $1.00 © 1990 IEEE

3. Learning

Learning is achieved by back propagation[1]. Figure 2 shows convergence. A training sequence is composed of a total of 5440 patterns which are divided from continuous speech samples from a male speaker. The ordinate indicates SNR and the abscissa indicates the number of logarithmic training iterations. The number of units in the input and output layers is 8 and the number of units in the hidden layer changes from 1 to 8.

As the number of units in the hidden layer increases, the better SNR becomes and the more training CPU time it needs to converge.

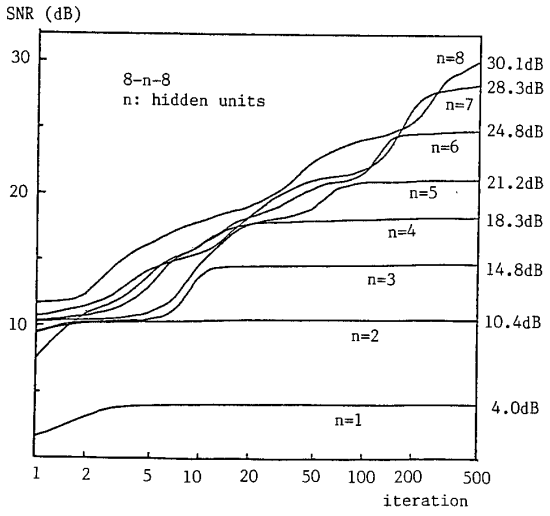


Figure 2 Convergence

4. Speech Compression

We studied the relationship between compression ratio and SNR when the number of units in the input and output layers is fixed.

Figure 3 shows SNR at several compression ratios when the number of input units is 8. The abscissa indicates the number of units in the hidden layer.

Figure 4 shows SNR when the number of input units is 32. The number of training iterations is 500 for 8 units and 250 for 32 units.

The solid line indicates the average SNR of playback speech by the same speaker as for the training sample. The other lines are of different speakers.

There is little audible difference between the three kinds of playback speech. It is clear that SNR increases linearly in proportion to the number of hidden layer units in all of the three cases, and so this system proves to be speaker independent.

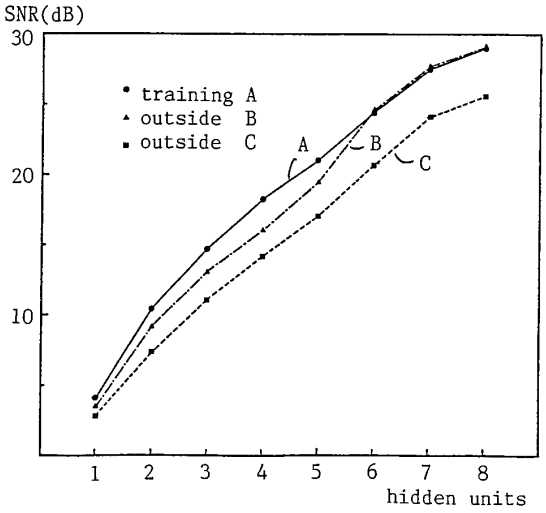


Figure 3 SNR and Compression Ratio (1) (frame length = 8)

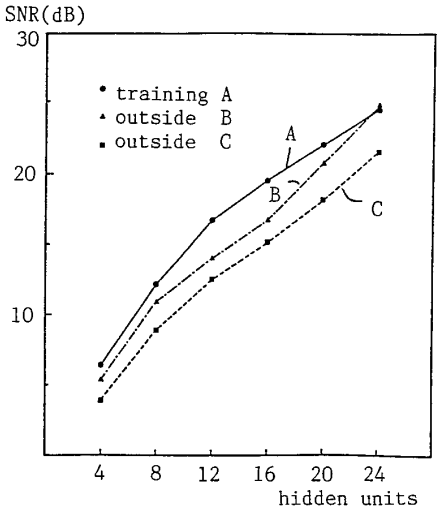


Figure 4 SNR and Compression Ratio (2) (frame length = 32)

5. Speech Coding

Speech coding can be performed by quantizing the outputs of the hidden layer units. The transformation from input layer to hidden layer is the speech coder and that from hidden layer to output layer is the decoder. Chapter 4 shows the linearity between SNR and compression ratio. So, it is necessary to decide the number of hidden layer units and a bit allocation for each unit in order to obtain the desired bit rate. There is a trade-off between these two factors. Vector quantization and scalar quantization can be used in quantizing the output of hidden layer.

Figures 5 and 6 show the relationship between SNR and bit rates of two types of system. The frame lengths of these systems are 8 and 32 respectively. The abscissas represent the compression ratio and the ordinates represent SNR of playback speech.

Each point indicates the bit allocation number for each hidden layer unit. The broken lines show the coding performance at the same bit rates. The solid line means the unquantized performance.

From these results, there is an optimum value for the compression ratio and an optimum number of hidden layer units to realize a specific bit rate. For example, when the frame length is 8, and the desired bit rate is 16 kbps, the optimum number of hidden layer units is 4 and the optimum quantization level is 4 bits.

Non-linear scalar quantization is used and the quantization level for each hidden layer unit is decided by the 1-dimensional LBG algorithm using 5440 training patterns.

Table 1 shows the coding performance of vector quantization of the hidden layer output for 3 speakers. The frame length is 8, so the number of codebook size multiplied by 1000 equals to bit rate number.

This table shows SNR of two types of vector quantization. The "direct" represents the performance in case of vector quantization of speech signal without compression and "hidden" represents the performance of VQ of hidden layer output.

There is little difference in SNR and speech intelligibility between two types of vector quantization. So speech compression based on neural network is effective to reduce the dimension number of vector quantization.

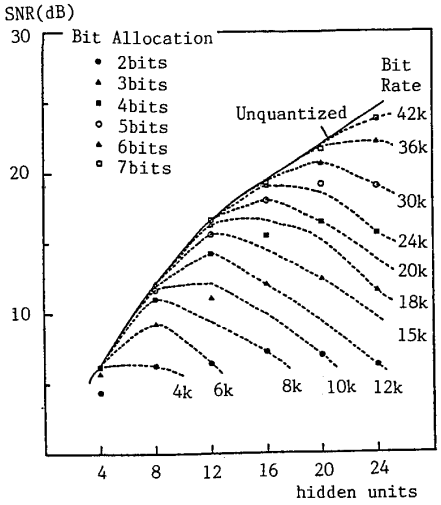


Figure 6 SNR and Bit Rate (2) (I/O units = 32)

(dB)

code book size	Training A		Outside B		Outside C	
	direct	hidden	direct	hidden	direct	hidden
5bit	9.28	9.32	7.44	6.88	6.32	6.00
6bit	11.08	11.15	8.57	8.22	7.36	7.17
7bit	12.67	12.67	9.71	9.31	8.45	7.97

Table 1 Comparison of two types of VQ

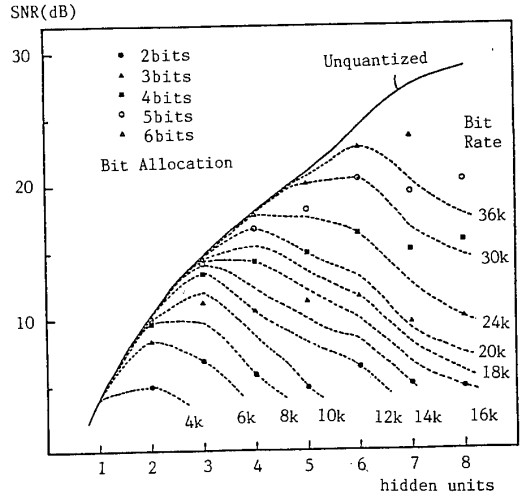


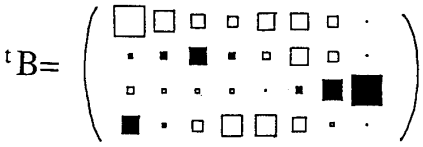
Figure 5 SNR and Bit Rate (1) (I/O units = 8)

6. Analysis of Weighting Coefficients

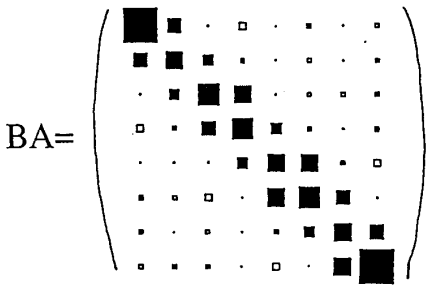
Figure 7 shows the matrix of weighting coefficients trained by back-propagation with a frame length of 8 and the number of hidden units equal to 4. Matrix A, whose size is 4*8, is the transformation from the input layer to the hidden layer, if a linear function is assumed. Matrix B, whose size is 8*4, is the transformation from the hidden layer to the output layer. A white square means a positive value, a black square means a negative value and the size of square means the absolute weighting value. BA represents the transformation from the input layer to the output layer if a linear function is assumed. Hidden layer outputs take values close to zero, so sigmoid function can be approximated to the linear function. Figure 8 shows the matrix for 8 hidden units.

These results show that A and B are very similar, and that BA is an almost diagonal matrix. Matrix A shows features of an orthogonal matrix. Speech coding based on a neural network is a transform coding method similar to KL transform.

a) Weight from Input to Hidden layer



b) Weight from Hidden to Output layer



c) Transform from Input to Output

Figure 7 Matrix of Weighting Coefficients (1)
(number of hidden units = 4)

7. Analysis of 5-layer Neural Network

We studied 5-layer neural network in this section. In case of 3 layer network, the transform from the input layer to the hidden layer and that from the hidden layer to the output layer are almost expressed as a linear and orthogonal matrix. So the nonlinearity of sigmoid can not be applied.

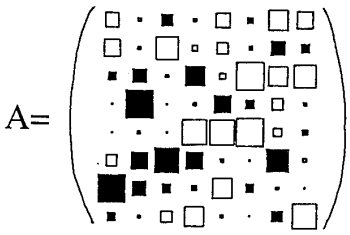
In this 5-layer network, 1st, 3rd and 5th layers are composed of linear units and 2nd and 4th layers have non-linear units.

An example of 5-layer network is indicated in Figure 9. The dimension of input and output planes is 2 and 3rd layer has only one unit in order to make it possible to check visibly the feature of mapping in the plane.

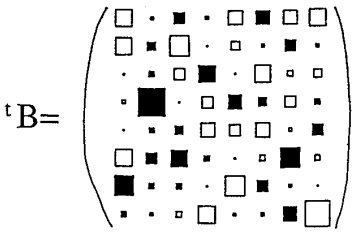
In the mapping from 1st layer to 3rd layer, all of the points in the 2D input plane are expressed with 1 dimensional value and that from 3rd to 5th draws a curved line in the 2D output plane. Eight points are selected randomly in the 2D plane to learn. Training iteration is 10000.

Figure 10 indicates the 8 learning points and reproduced points. The curve represents the locus of reproduced point when hidden layer (3rd layer) output takes value from -1 to 1 continuously. All of the training points can be reproduced with only little prediction error by this network.

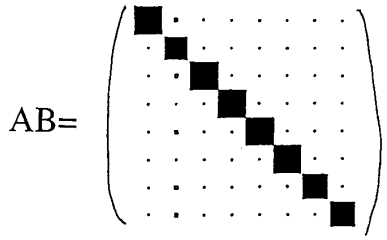
Figure 11 indicates a contour line of input plane. All of the points in the region between any two lines is mapped to the same region on the reproduced locus line in output plane. We can see that non-linear mapping can be realized by 5-layer neural network.



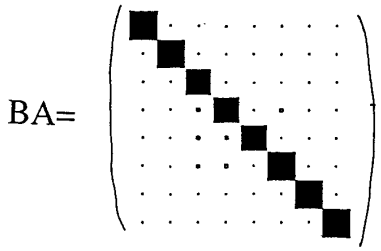
a) Weight from Input to Hidden layer



b) Weight from Hidden to Output layer



c) Transform from Output to Input



d) Transform from Input to Output

Figure 8 Matrix of Weighting Coefficients (2)
(number of hidden units = 8)

Figure 12 and 13 show the reproduced points and locus line, and the contour line of input plane by 3-layer neural network respectively. The number of input and output units are 2. The hidden layer has one unit.

In this case, we can see that the reproduced line is only straight line and mapping is only linear.

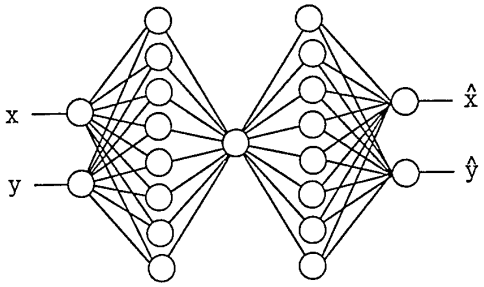


Figure 9 5-layer Neural Network

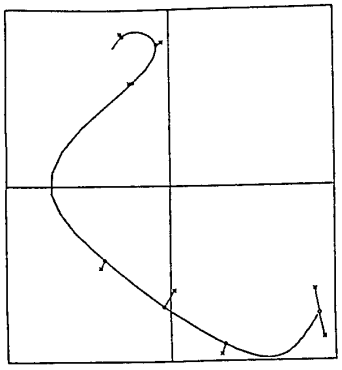


Figure 10 Reproduced Points and Locus Line by 5-layer Neural Network

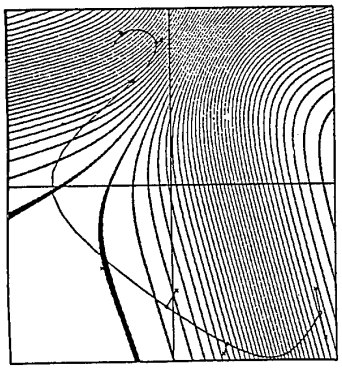


Figure 11 Contour Line of Input Plane in case of 5-layer

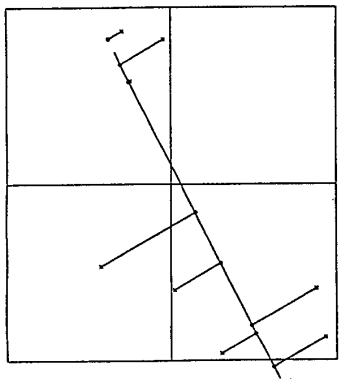


Figure 12 Reproduced Points and Locus Line by 3-layer(2-1-2) Neural Network

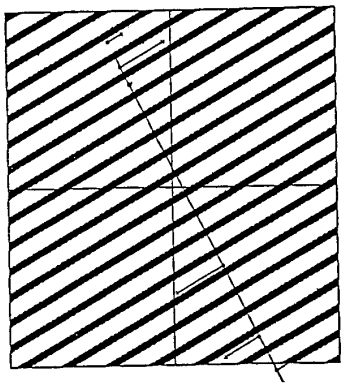


Figure 13 Contour Line of Input Plane in case of 3-layer

8. Conclusion

We have discussed the performance of speech compression or speech coding based on a 3-layer neural network. This system has obtained the feature of transform coding automatically based on back- propagation. The coding performance is independent of speaker or data.

The 5-layer neural network can realize non-linear mapping, so it is more effective than 3-layer network. We are examining the ways of improving coder by 5-layer network.

References

[1] J.L.JcCell and D.E.Rumelhart, "Parallel Distributed Processing", MIT Press, 1986.
[2] J.L.Elman and D.Zipser, "Learning the Hidden Structure of Speech", ICS Report 8701, Feb. 1987.
[3] G.W.Cottrell, P.Munro and D.Zipser, "Image Compression by Back Propagation", ICS Report 8702, Feb. 1987.
[4] Y.Katayama, "Image Coding based on Neural Network", PCS-J88, 3.5, 1988.