QUANTIZATION NOISE REDUCTION OF LINEAR PCM SOUND BY USING DENOISING AUTO-ENCODER

Shohei Oouchi, and Kazunori Mano

Division of Systems Engineering and Science

Graduate School of Engineering and Science, Shibaura Institute of Technology

Contact Email: mf15010@shibaura-it.ac.jp, mano@sic.shibaura-it.ac.jp

ABSTRACT

Recently, it is becoming popular to record meeting discussions or daily conversations as life log speech data. The recorded speech is highly degraded by quantization noise if the amplitude of the input speech level is very low. To improve the degraded speech quality, speech enhancement techniques are used. This paper proposes a quantization noise reduction technique for low bit linear PCM speech sound by using a denoising auto-encoder (DAE). The proposed system is composed of two parts, a training part and a noise reduction part. In the training part, firstly, clean sound databases and corresponding quantized sound databases with low bit quantization are prepared as parallel data. Secondly, the data sounds are transformed into spectral sequences by short-time Fourier transform. The denoising auto-encoder is a kind of neural networks. In this research, three layers of networks are organized. Input data are the connected vectors of amplitude spectral sequences of consecutive five frames in order to obtain inter-frame changes of speech. The output for each frame is obtained as a result of each connected vector input. As a result, comparing the processed speech with the original low bit sound, the signal-to-noise ratio of the sound by using DAE improves the objective quality by around 2 dB. Informal subjective listening test also shows the effectiveness of the proposed method.

1. INTRODUCTION

The linear PCM is recorded sound faithfully as possible, and the sound is high quality. However, it has disadvantage that sound quality suddenly decreases when quantization noise occur by lowing bit rate. Usually, the sound is converted as quantization accuracy 16bit as IC recorder. The converted sound stores in memory. The sound data is unfavorable in this way at a point of view of the memory capacity, because the memory capacity becomes large. The error with the original analog signal appears conspicuously by using IC recorder in meetings and lectures because the sound is relatively recorded in a low bit when the distance between the talker and microphone is long. In sound communications, transmitted information can be reduced to deal with encoding speech by quantizing at low bit rate. In this paper, a quantization noise reduction technique is proposed for low bit linear PCM sound by using a denoising auto-encoder (DAE), which is one of neural network systems.

2. SOUND PROCESSING

2.1 PCM Transmission

The PCM transmission system is shown in Fig.1. In the system, the analog signal input by a microphone is passed through a low-pass filter. Then, the analog signal is converted into PCM signal by sampling, quantization, and encoding. The PCM signal is transmitted from transmission side. On the receiving end, the "decoder" transforms the PCM signal into a regenerated analog signal. In this research, the "quantization" mechanism at A/D conversion is studied.

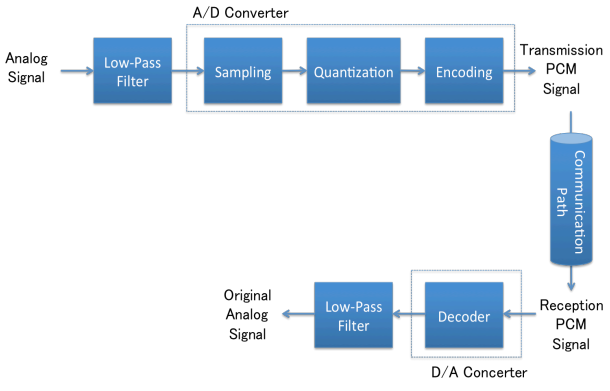

Fig.1 PCM transmission scheme

2.2 Quantization Noise

The quantization process is to convert an analog signal into a digital signal. In basic PCM quantization, signals are quantized with uniform step-width. This is called "the uniform quantization". The sound data before the quantization is defined as $x(n)$. When the data is quantized by B-bit by using the same quantization, the quantization step width δ is defined as follows.

$$\delta = \frac{2R}{2^B}, \tag{1}$$

where, R is the absolute maximum value of $x(n)$. The sound data is divided by the step width. Then, the sound data are encoded in integer values as follows by using a rounding off function.

$$c(n) = round\left[\frac{x(n)}{\delta}\right] \tag{2}$$

Quantization noises degrade the waveform patterns, when the bit rate of quantization is lowered. Fig. 2 compares 16-bit sound with 6-bit sound. The quantized sound in low bit rate is a stepped signal.

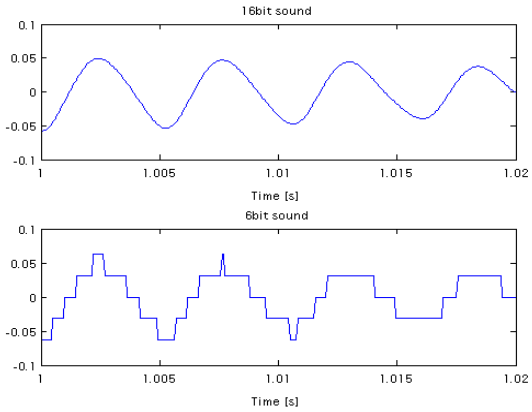

Fig.2 Comparison of 16-bit sound with 6-bit sound

3. LEARNING BY DENOISING AUTO-ENCODER

3.1 Auto-encoder (AE)

AE is used for dimension compression and feature extraction of signals. An input layer x, a hidden layer y, and an output layer z are expressed by the following formula.

$$\text{Encoder: } y = f(Wx + b_0) \tag{3}$$

$$\text{Decoder: } z = f(W'y + b_1) \tag{4}$$

where, W is $m \times n$ matrix of weights for the layers, b_0 and b_1 express each bias. b_0 is an m dimensional vector and b_1 is an n dimensional vector. f is an activation function. AE defines a loss function which corresponding to the error between the target output and the output from the hidden layer. The parameters of the function are minimized by using a stochastic gradient descent method [2].

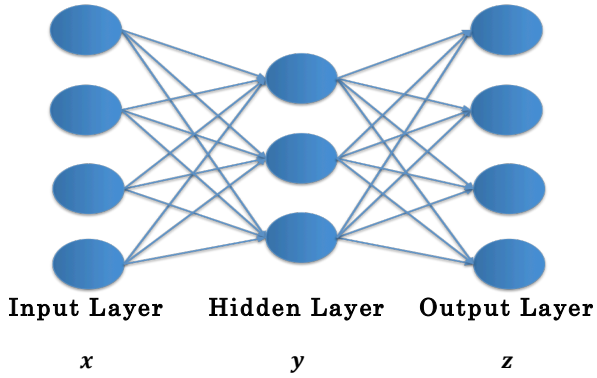

Fig.3 AE scheme

3.2 Denoising Auto-Encoder

DAE converts certain noisy signals into the clean signals. The basic structure is almost the same as AE. As a difference with AE, in DAE system, the input data are the noisy signals, and the noises are removed by training. In DAE, the training is performed using the input layer x in the equation (3) as the signal mixed with noise. DAE defines the error between the target output and the output from the hidden layer as a loss function. Each parameter is trained to minimize the loss function based on a stochastic gradient descent method. DAE can be effectively applied to remove additive noises, multiplicative noises, and reverberation noises [3].

3.3 Stochastic Gradient Descent Method

In this paper, a square error is used as loss function. Here, the signal z_n is output by DAE and the clean signal d_n is the target output. The error is expressed by the following formula.

$$E(w) = \sum_{n=1}^{N} \|d_n - z_n\|^2 \tag{5}$$

Stochastic gradient descent is the method that updates the parameter sample-by-sample. The parameters w, b_0, b_1 are obtained by minimizing the selected $E(w)$. From the current weight $w^{(t)}$, the updated weight $w^{(t+1)}$ are calculated as follows.

$$w^{(t+1)} = w^{(t)} - \varepsilon \frac{\partial E_n}{\partial w} \tag{6}$$

In addition, The biases b_0, b_1 are updated in the same way.

$$b_0^{(t+1)} = b_0^{(t)} - \varepsilon \frac{\partial E_n}{\partial b_0} \tag{7}$$

$$b_1^{(t+1)} = b_1^{(t)} - \varepsilon \frac{\partial E_n}{\partial b_1} \tag{8}$$

where, ε is a learning coeffeicient. It is a constant to determine the update quantity. Then, the error with the target output can be decreased.

4. EXPERIMENT

4.1 Experimental Condition

The experiment was conducted by using ATR phoneme balance sentence uttered by a woman speaker to show the effectiveness of the quantization noise reduction by using DAE. Parallel training data of 450 sentences with 6-bit quantization and 450 sentences with 16-bit quantization are prepared. Other 53 sentences with 6-bit quantization were prepared as test data. The sampling frequency of the data is 16 [kHz]. Spectra of each sound data are calculated by using STFT of 512-sample FFT size and 256-sample shift size. Therefore, the number of input data is $257 \times$ "frame-number". Phase of the testdata is prepared. When a sound is restored to the original state, these filter is used. To catch temporal changes of sound the input to DAE is a connected three frames, which are previous, current and next frames. From the preliminary experiment, the structure of DAE is decided that the number of hidden layer is 200, and the number of updates is 1000. The encoder uses non-linear function and the decoder uses identity function.

As evaluation experiment, the S/N ratios of DAE output sound and the 6-bit sound in equations (9) and (10) are compared.

$$SNR_{6bit} = 10\log_{10}\frac{\sum {x_{16bit}}^2(n)}{\sum(x_{16bit}(n) - x_{6bit}(n))^2} \quad (9)$$

$$SNR_{dae} = 10\log_{10}\frac{\sum {x_{16bit}}^2(n)}{\sum(x_{16bit}(n) - x_{dae}(n))^2} \quad (10)$$

where, x_{16bit} is original, x_{6bit} is 6-bit sound, and x_{dae} is the DAE sound.

4.2 Experiment Procedure

The proposed method consists of two parts, the training phase (as in Fig. 4) and the test phase (as in Fig.5). The summary is shown below.

Training phase:

(1) The 16-bit sound spectrograms of the training data are the target output. Therefore, the 6-bit sound spectrogram is input data. Every five frames of each sound are trained by the DAE.

(2) The square error between the output of frames by DAE and those of the target is calculated, and the paramters,W, b_0, $b_1$are updated by stochastic gradient descent method.

Test phase:

(1) The 6-bit sound spectrum of test data is input to DAE with the parameters obtained in the training phase.

(2)The spectrum made by DAE and the original 6-bit sound spectrum are converted into time domain.

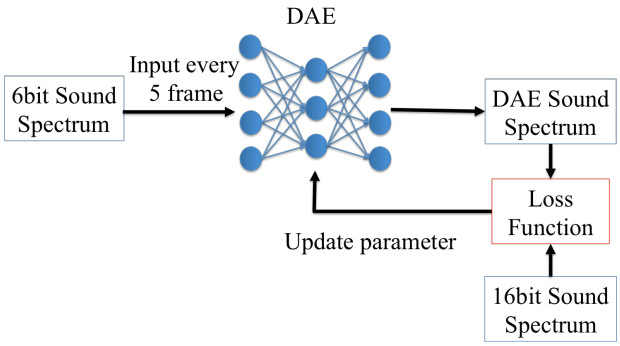

Fig.4 Training phase

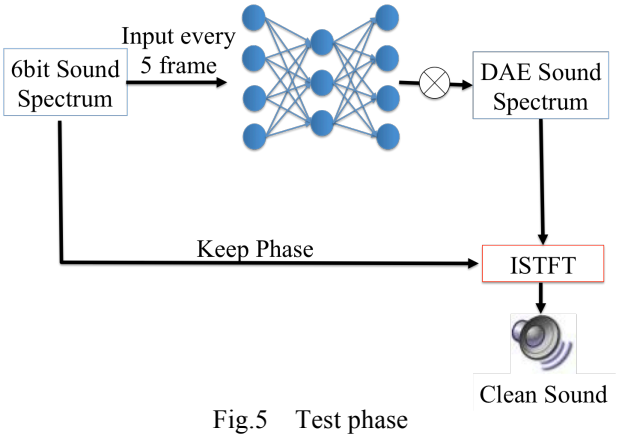

Fig.5 Test phase

4.3 Experimental Result

The 16-bit sound wave pattern of the target sound and the 6-bit sound wave pattern of the input sound, which is the output sound made by DAE are shown in Fig. 8. Compared to target sound and input sound, the input sound is degraded and the information of the wave pattern is lost by quantization. On the other hand, compared to the output sound and the target sound, two wave patterns are similar than those of target and input patterns. The lost Information by quantization may be recovered.

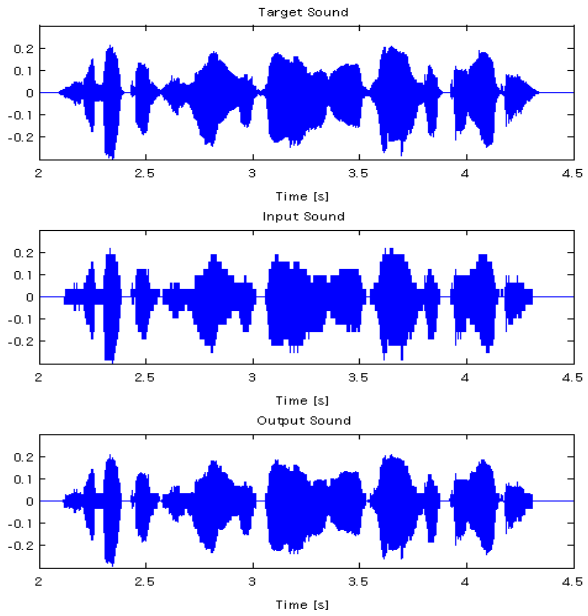

Fig.6 Wave pattern of each sound

The spectrograms are shown in Fig. 7.

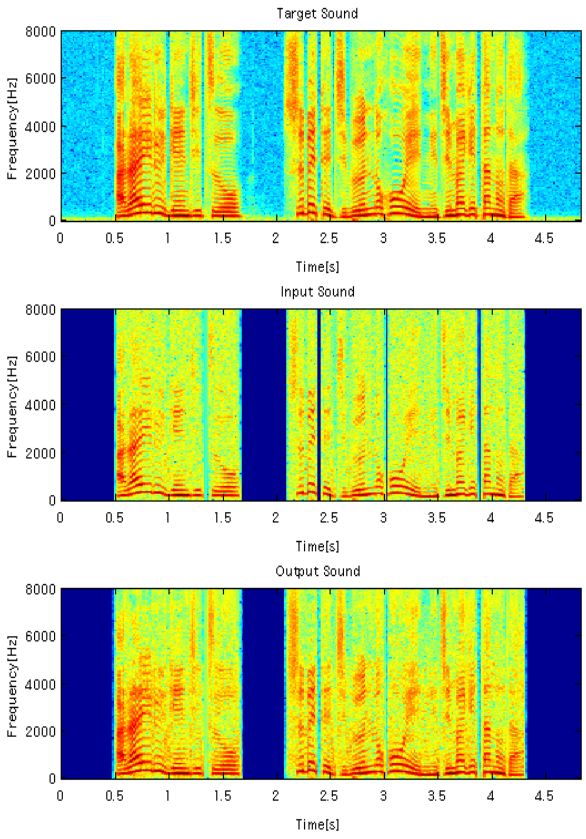

Fig.7 Spectrogram of each sound

Compared the target sound with the input sound, quantization errors are not found in the low level. From middle level to high level, some information is lost. On the other hand, compared the target sound with the output sound, the quantization error may be recovered. The wave pattern is a little bit different from middle level to high level frequencies, but the DAE sound value is almost the same as the target sound.

4.4 Evaluation experiment

The S/N ratios are shown in Table 1 obtained by equations (9) and (10).

Table 1. The S/N ratios with the target sound

Sound	S/N ratio [dB]
Input sound	18.65
Output DAE sound	20.67

The bigger, the S/N ratio is referred to a better quality sound. The output DAE sound has better quality than the input sound. Table 2 shows the square errors.

Table2. Square error with the purpose sound

Sound	Error
Input Sound	4.082
Output DAE Sound	2.561

The smaller, the square error expresses fewer errors with target output. As a result, the DAE sound quality is better than that of input quantized sound.

5. CONCLUSION

In this paper, a quantization noise reduction method based on DAE is proposed. The clean sound component is recovered from degraded 6-bit sound with the DAE. As current problem, the estimated spectrum still contained quantization noises. It is due to the size of the training data, the size of input layer, and the number of hidden layers. Another problem is that the data restored to the original wave pattern is limited when the number of the quantization bit is below 4-bit.

REFERENCES

[1] P.Vincent et al., “Extracting and composing robust features with denoising autoencoder,” ICML , pp. 1096-1103, 2008

[2] G. E. Hinton and R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science 28, vol. 313, no. 5786, pp. 504-507, 2006.

[3] Ishii, T., Komiyama, H., Shinozaki, T., Horiuchi, Y. and Kuroiwa, S.: “Reverberant speech recognition based on de- noising autoencoder, " INTERSPEECH 2013, pp. 3512–3516, 2013.

Shohei Oouchi received B.E. from Shibaura Institute of Technology in 2015. He is admitted to Graduate School of Shibaura Institute of Technology in 2015. His current research theme is speech processing.

Kazunori Mano received the B.E., M.E. and Dr. Eng. degrees in electrical engineering from Waseda University, Japan, in 1982, 1984 and 1987, respectively. From 1987 to 2008, he engaged in research on speech coding and synthesis at NTT laboratories. Since 2008, he has been a Professor, Department of Electronic Information Systems, Shibaura Institute of Technology, Japan. His current interests include speech processing, media coding, and communication systems.