Neural Network を使った波形入出力による雑音抑圧

Noise Reduction Using Neural Networks

田村 震一　アレックス ワイベル

Shin' ichi Tamura , Alex Waibel

ATR自動翻訳電話研究所

ATR Interpreting Telephony Research Laboratories

あらまし　　neural network を使った波形入出力による雑音抑圧について述べる。4層のfeed forward neural network を使って雑音が加わった信号の空間から雑音のない信号の空間への写像を実現する。neural network の学習アルゴリズムはバックプロパゲーションアルゴリズムを用いる。サンプリングレートが12kHzの日本語の単語音声と定常、非定常雑音を用いてコンピュータ実験を行い、本手法の有効性、有望性を確認した。

Abstract　　In this paper, we describe a method for noise reduction using neural networks.With the back propagation network learning algorithm, a four-layered feed-forward network is trained on learning samples to realize a mapping from the set of noisy signals to the set of noise-free signals. Computer experiments were carried out on 12kHz-sampled Japanese speech data and using stationary and non-stationary noise. Our experiments showed that the network can indeed learn to perform noise reduction. Even for noisy speech signals that had not been part of the training data, the network successfully produced noise-supressed output signals.

1. INTRODUCTION

Especially from the invention of James Watt's steam engine, machines have greatly enlarged human beings' abilities. In the mid of 20th century, this tendency that machines enlarge our abilities got accellated owing to the exciting invention of electric digital computers. Using digital computers, we could even go to the moon. . . . .

And today, our labs, ATR Interpreting Telephony Research Laboratories are aiming at the realization of an automatic telephone interpretation system using digital computers. This will enable a person speaking in one language to communicate readily by telephone with someone speaking in a different language. To construct such a system, speech recognition, machine translation, and speech synthesis are the three major technologies required. Being Done our daily communication by telephone in noisy environments, reduction of additivenoise, or noise reduction is one of the key technologies for such a system, especially for the speech recognition part.

Most noise reduction methods to-date fall into two major categories. One of them is based on mathematical models. Such an approach uses a priori mathematical knowledge of speech and noise in the form of a mathematical model. So in practice, detailed information is required for successful application. For example, a typical approach of this category is to model speech production dynamics using an all pole time-invariant filter [1]. First, the parameters of the filter are estimated based on the short time segment of noisy speech data. Then that part of noisy speech is filtered using the speech production model with the estimated parameters. This approach therefore relies heavily on parameter estimates, that are difficult to obtain in practice, given noisy speech. It is also based on linear speech production models, or all pole models and they are only first order approximations of speech production dynamics.

The second approach uses examples [2,3,4,5]of speech and noise and performs noise reduction using rather intuitional methods. Power Spectrum Subtraction is a typical example[2]. This method is based on the assumption that the phase of short time speech spectra is less important than the magnitude and that peaks in the power spectrum are more important than valleys. The short time power spectrum of speech is estimated by subtracting the estimated short time power spectrum of noise from noisy speech. Then the estimated short time power spectrum is combined with the short time phase of the noisy speech and the spectrum is transformed into the time domain signal. Even though the short time phase information is less important than the short time magnitude information, it would be preferable to exploit phase information for better noise reduction. Spectral subtraction also makes simplifying assumption about the shape of the noise and it's combination with the original speech signal. More complex interactions between noise and speech signal, as well as non-stationary noises can not be captured easily.

As an approach that might overcome some of these limitations, we propose a new noise reduction method using neural networks. Noise reduction can be viewed as a mapping from the set of noisy signals to the set of noise-free signals. Let *f* be such a mapping. The problem is how to find *f*. Neural networks are attractive as mapping definition for the following reasons.

(1) An arbitrary decision surface can be formed in a multi-layered neural network[6]. So any complex mapping from the set of noisy speech signals to the set of noise-free speech signals can in principle be realized .

(2) Simple learning algorithms exist to construct a suitable mapping function using training samples.[7].

(3) Neural networks have attractive generalizing properties [7].

In the following, we first describe a neural network for mapping noisy to noise-free speech signals and then show the effectiveness of this approach by computer experiments.

2. NEURAL NETWORKS FOR MAPPING

As a framework for representing an arbitrary mapping function, a network of interconnected simple computing elements is considered.

2.1 Network Architecture

A four-layer feed-forward network was chosen as an architecture for it can realize in principle any mapping function[6]. Each layer has 60 units and is fully interconnected with its next higher layer (Fig. 1).

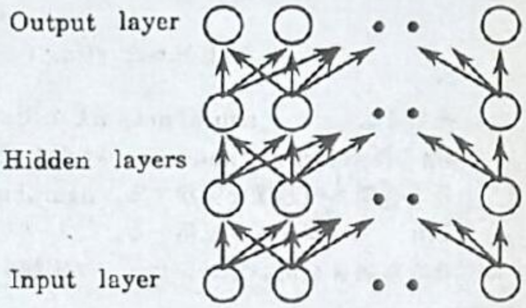


Fig 1 Network for noise reduction

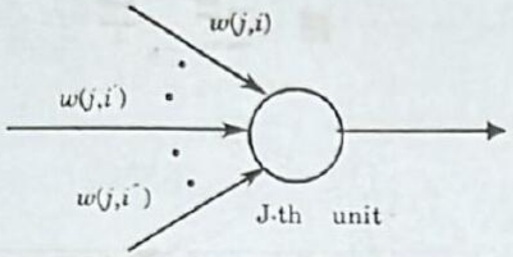
The network's state, or the units' outputs are updated synchronously on each layer and signals flow upwards from the input layer to the output layer. For the network to use as much information on speech and noise as possible, the input and output of the network is given by the waveform itself, the units on the output and input layers are all linear units, i.e., are not passed through a non-linear output function.

2.2 Unit Element

A unit element is one of many simple processors that make up the network. It first computes the weighted sum of all its inputs (including a bias input) and then deforms this sum by passing it through a nonlinear function, in our case the sigmoid function [7](Fig. 2).

2.3 Learning by Error Back-Propagation

Using the training input and output data, the back-propagation learning procedure adjusts the network's link weights to realize the noise reduction mapping[7]. The back-propagation algorithm defines a square error measure between a desired target output and the actual network's output given its current input and network connection strengths. On every presentation of learning samples, each link weight is updated in an attempt to decrease this output measure[7].



J-th unit's output = $f(\Sigma\, w(j,i)o(i) + \theta(j))$,

where
$f(x) = 1/(1+exp(-x))$ is the sigmoid function,
$\theta(j)$, the bias value of j-th unit and
$w(j,i)$, the link weight from the i-th unit to the j-th unit

Fig 2 A unit element

3. EXPERIMENT

In the following, we present experimental results from using our model for noise reduction.

3.1 Data

The speech database used in our experiments consists of 5000 common Japanese words uttered in isolation by several male speakers (professional announcers). The data was digitized (16 bits) at a 20kHz rate and then down-sampled to 12 kHz. A subset of 216 phoneme balanced words from this database was used for our experiments.
Computer room noise was chosen as non-stationary noise. This noise was first recorded using an analog tape recorder and then digitized to 16bit data at a 12kHz-sampling rate. Noisy speech data was generated artificially by adding the computer room noise to the speech data. The resulting S/N ratio was about -20db.

3.2 Learning

Using the waveforms of the 216 phoneme-balanced words as target output and their noise added versions as the training input, the network scans each training utterance from beginning to end at a rate of 60 data points per input frame. When the network reaches the end of the training data, it returns to the beginning for additional learning passes. This procedure is repeated until the network's squared error rate converges to a sufficiently small value.

During this phase, the back-propagation learning procedure repeatedly adjusts the network's internal link weights in an attempt to find an optimal mapping between noisy and noise-free signals (Fig. 3).

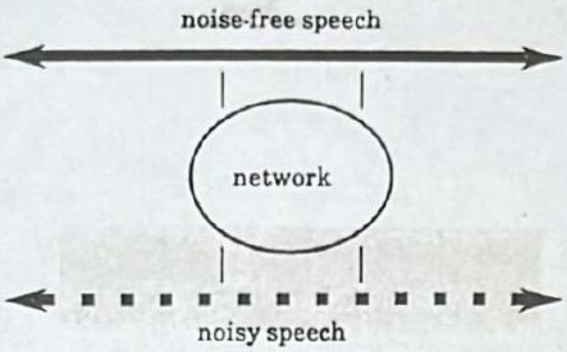


Fig. 3 Network learning

3.3 Results

Fig. 4 shows the squared error of the network during learning. It illustrates that learning was done successfully and demonstrates the convergence of the network's output to the desired target output.

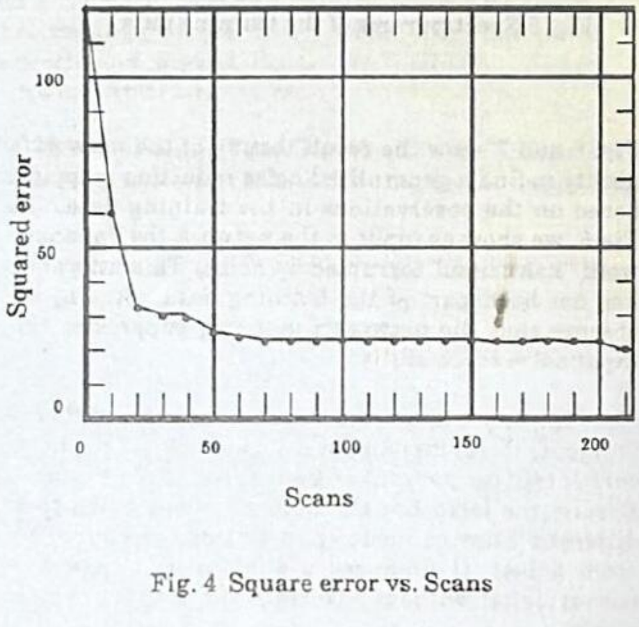


Fig. 4 Square error vs. Scans

The tests reported in the following were performed on networks that were trained on about 200 scans through the traininig utterances. Learning the noise suppression mapping for this data took about three weeks on an Alliant super computer. Fig. 5 shows the result of training after about 200 scans. The input to the network is the noisy Japanese word "ikioi" from the training data. As can be seen, the noise has been reduced significantly, while the speech spectrum is preserved.

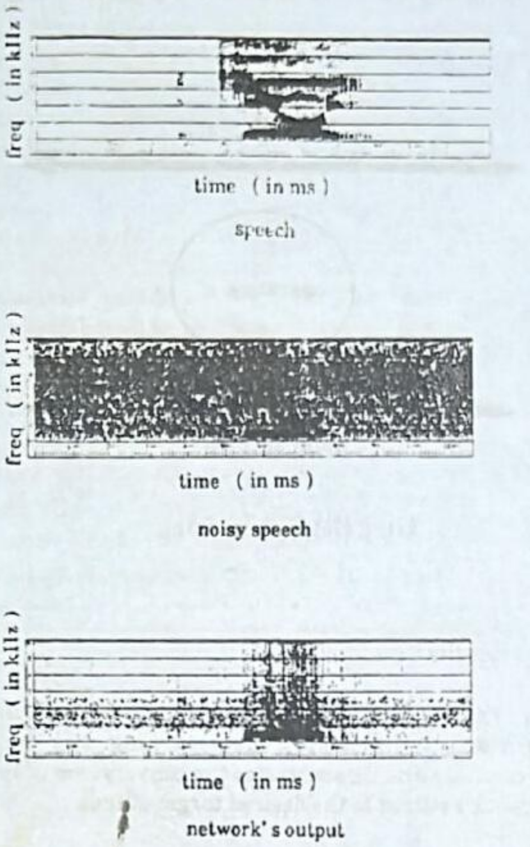


Fig. 5 Spectrograms of the training data

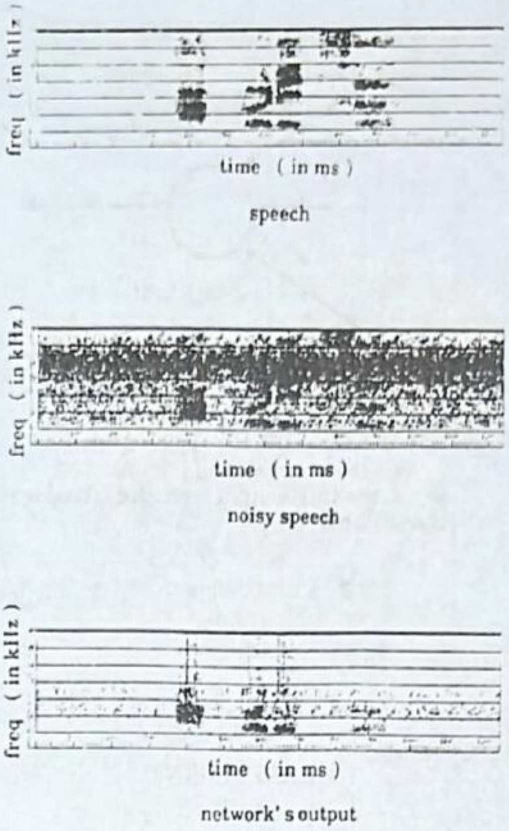


Fig. 6 Spectrograms of non training data

Method	Score
Power spectrum subtraction	43.4%
Neural Network	56.6%

Fig. 8 Result of auditory preference test

Fig. 6 and 7 show the result testing of the network's ability to find a generalized noise reduction mapping based on the observations in the training data. In Fig.6, we show as input to the network the Japanese word "kakuritsu" corrupted by noise. This utterance has not been part of the training data. Again, we observe that the network's mapping suppresses the input noise successfully.

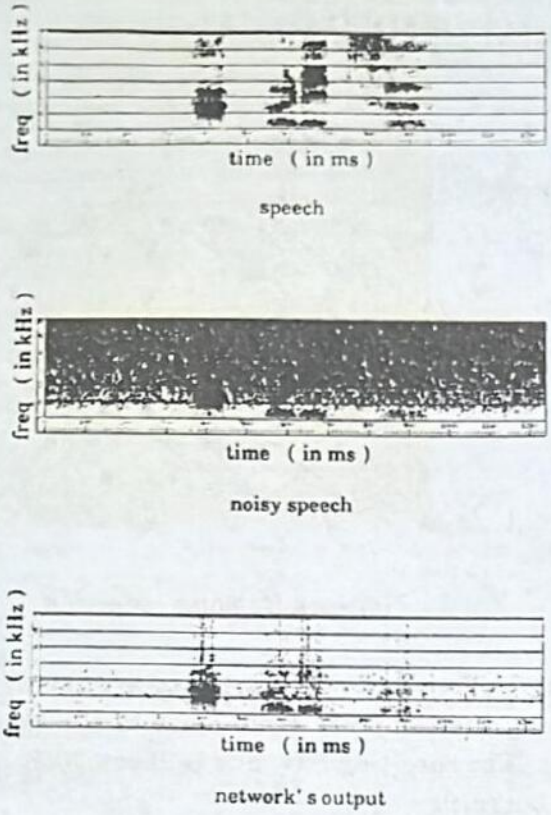
In Fig. 7, we show the result of a more difficult problem. Here, the same word, "kakuritsu" has been corrupted by computer-generated white noise. Despite the fact, that the network was trained on a different kind of noise (non-stationary computer room noise), it produces a substantially cleaner output signal, without adversely affecting the speech signal.

Fig. 8 is the result of an auditory comparison with the conventional power spectrum subtraction method. In this method, the short-time spectral magnitude of speech is estimated by

$$\begin{cases} |Y(\omega)|^2 - E\,|N(\omega)|^2 & \text{for } |Y(\omega)|^2 > E\,|N(\omega)|^2 \\ 0 & \text{otherwise,} \end{cases}$$

where
$Y(\omega)$ is the short-time spectrum of noisy speech, $N(\omega)$ is the short-time spectrum of noise, and E is the operation of the ensemble mean. The frame length is 64 points long and the shift is also 64 points long.

Noise suppressed speech was presented to listeners in pairs and subjects were asked to mark the prefered speech sample. Subjects' responses indicate that our noise reduction method yields a noise free speech signal that is comparable to or better than the conventional power spectrum subtraction method. Although our connectionist model produced a cleaner signal than power spectrum subtraction, it does, however, not appear to yield greater intelligibility.



speech

noisy speech

network' s output

Fig. 7 Spectrograms of non training data and different noise

We believe that more focused learning of acoustic-phonetically important parts of the speech signal might lead to further improvements in intelligibility.

4 . CONCLUSION

In this paper, we have described a noise reduction method using neural networks. In a series of computer experiments we have shown that connectionist models can learn the mapping between the set of noisy signals and the set of noise-free signals correctly. We have shown that the network produces noise-suppressed signals even for signals that differed from the training data in both the original speech input as well as the type of environmental noise.

ACKNOWLEDGEMENT

The authors are deeply grateful to Dr. Kurematsu, the president of ATR Interpreting Telephony Research Laboratories, Dr. Shikano, the head of Speech Processing Department, Dr. Kuwabara, the supervisor of Speech Processing Department, and all the members of Speech Processing Department for their constant help and encouragement.

REFERENCES

[1] J. S. Lim and A. V. Oppenheim, " All-pole modeling of degraded speech, " IEEE Trans. on Acoust., Speech, Signal Processing, vol. ASSP-26, pp. 197-210, June 1978.

[2].S. F. Boll "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. on Acoustics, Speech and Signal Proc., vol. ASSP-29, pp. 113-120, April 1979.

[3] H. Drucher, " Speech processing in a High Ambient Noise Environment, " IEEE Trans. on Audio and Electroacoustics, vol. AU-16, pp. 165-168, June, 1968.

[4] R. H. Frazier, S. Samsam, L. D. Braida, A. V. Oppenheim, " Enhancement of Speech by Adaptive Filtering", Proceedings of the Int. Conf. on Acoustics, Speech and Signal Proc., pp. 251-253, Philadelphia, PA, April 12-14, 1976.

[5] Y. Ariki, K. Kajimoto and T. Sakai, " ACOUSTIC NOISE REDUCTION BY TWO DIMENSIONAL SPECTRAL SMOOTHING AND SPECTRAL AMPLITUDE TRANSFORM, " Proceedings of the Int. Conf. on Acoustics, Speech and Signal Proc. pp. 97-100, Tokyo, Japan, April 7-11, 1986.

[6] R.P.Lippmann, "An Introduction to Computing with Neural Nets," IEEE ASSP magazine, pp. 4-22, April 1987.

[7] D.E.Rumelhart, J.L.McClelland and the PDP Research Group, *Parallel Distributed Processing*, Vol.1, Chap.8, MIT Press, 1986.

盛光印刷所