NEURAL NETWORK BASED SPEECH ENHANCEMENT

Jaroslav TLUČÁK, Jozef JUHÁR
Ľubomír DOBOŠ, Anton ČIŽMÁR
Dept. of Electronics and Multimedia Communications
Technical University of Kosice
Park Komenskeho 13, 040 01 Kosice
Slovak Republic

Abstract

This paper deals with methods of speech enhancement with particular focus on neural speech enhancement. Speech enhancement is concerned with the neural processing of noisy speech to improve the quality and intelligibility of the speech signal. The goal of this paper is to describe an experiment with implementation of two channel adaptive noise canceler via direct time domain mapping approach.

Keywords

Adaptive noise canceling, artificial neural networks, direct time-domain mapping, noise reduction, robust speech recognition, speech enhancement,

1. Introduction

In many communication settings the presence of background noise and channel interference causes the quality or intelligibility of speech to degrade. The effect of speech degradation becomes real problem in systems of Automatic Speech Recognition (ASR) too. The first step in most of the current ASR systems is to convert the incoming speech signal into series of short-term vectors. Each element of the vector describes some part of the information carried by the signal. Some of the elements of the short-term vector contain corrupted information. In the current ASR the entire feature vector is used as one entity and even a single corrupted spectral element can severely degrade the performance of the recogniser. The purpose of many enhancement algorithms is to reduce background noise, improve speech quality, or suppress channel or speaker interference. In this paper we describe approaches of speech enhancement with particular focus on speech enhancement with the use of artificial neural networks (ANNs).

2. Speech enhancement methods

There are a number of ways in which speech enhancement systems can be classified. Systems based on stochastic process models rely on a given mathematical criterion. Systems based on perceptual criteria attempt to improve aspects important in human perception. Enhancement algorithms can also be partitioned depending on whether a single-channel or dual-channel (or multichannel) approach is used. There are four broad classes of speech enhancement techniques.

The first class concentrates on the short-term spectral domain. These techniques suppress noise by subtracting an estimated noise bias found during nonspeech activity in single-microphone cases, or from a reference microphone in a dual-channel setting. The enhancement procedure is performed over frames by obtaining the short-term magnitude and phase of the noisy speech spectrum, subtracting an estimated noise magnitude spectrum from the speech magnitude spectrum, and inverse transforming this spectral amplitude using the phase of the original degraded speech.

The second class of enhancement techniques are those, which is based on speech modeling using iterative methods. These systems focus on estimating model parameters that characterize the speech signal, followed by resynthesis of the noise-free signal based on noncausual Wiener filtering. This class of enhancement techniques requires a priory knowledge of noise and speech statistics and therefore must also adapt to changing characteristics.

The third class of systems is based on adaptive noise canceling (ANC). Traditional ANC is formulated using a dual-channel time or frequency domain environment based on the "least mean square" (LMS) algorithm. Although other enhancement algorithms can benefit from a reference channel, successful ANC requires one.

The last area of enhancement methods is based on the periodicity of voiced speech. These methods employ fundamental frequency tracking using either single channel ANC (a special application) or adaptive comb filtering of the harmonic magnitude spectrum.

3. Neural Network Based Speech Enhancement Methods

We can describe the neural network techniques for speech enhancement as follows:

Time-Domain Filtering: The direct mapping approach trains a neural network using noisy inputs and clean targets. The implicit assumptions are that the statistics of both the speech and noise encountered will be the same as those of the training set, and that the statistics are constant (stationary) throughout. The extended Kalman filtering (EKF) approach uses a predictive neural model (trained on clean speech) in a state-space framework in order to produce approximate maximum likelihood estimates of the speech. The assumptions are that the signals are stationary, and the statistics of speech to be enhanced will be the same as those of the training set.

Transform-Domain Mapping: Neural networks are trained from noisy input features in a transformed domain. Domains are usually chosen for their desirable perceptual or recognition properties. The SNR or some other measure of the joint signal-noise statistics is typically used as an additional input to the network.

State-Depend Model Switching Methods: These methods use a variety of different models trained on different classes of speech and noise signals in an attempt to better reflect the non-stationarity of the data. Two switching methods were discussed. The first was a classifier-based approach for choosing the appropriate neural mapping for the current signal. The second was a hidden Markov model approach which allows for modeling of state transition probabilities, and which is typically used in conjunction with the extended Kalman filtering method of speech enhancement. The assumption is that the number of states used is sufficient to model all the different regions of speech and noise statistics.

On-line Iterative methods: These methods adapt on-line to the specific noisy signal of interest. No assumptions are made about generalization because they do not rely on a training set. On the other hand, they do not make use of any possible prior knovledge available from a training set, and require significantly more computation during actual enhancment.

The simplest approach to training a speech enhancer on-line is to build an adaptive predictor of the speech. It is assumed that the correlation length of the noise is less that of the speech signal.

The second approach uses the EKF speech estimator in conjunction with a second EKF parameter estimator. Using short windows of the noisy speech and running these two estimators in paralel results in the Dual EKF algorithm. Assumptions are the same as in the basic approach, with the exception that stationarity is assumed only over short-term windows.

4. Noise Canceling via Neural Network

The general technique of adaptive noise canceling (ANC) has been aplied succesfully to a number of problems that include speech, aspects of electrocardiography, elimination of periodic interference, elimination of echoes on long-distance telephone transmission lines, and adaptive antenna theory. The initial work on ANC began in the 1960s. Adaptive noise canceling refers to a class of a primary input source and a secondary reference source. The primary input source *d(n)* is assumed to contain speech *s(n)* plus additive noise $n_0(n)$:

$$d(n) = s(n) + n_0(n)$$

where, as usual, these sequences are realizations of stochastic processes **d**, **s** and $\mathbf{n_0}$. The secondary or reference channel receives an input *n(n)* that may be correlated with $n_0(n)$ but not with *s(n)* (see Fig. 1). All random processes are assumed wide sense stationary (WSS) and appropriately ergodic so that time waveforms can be used in the following analysis.

The adaptive noise canceler (Fig. 1) consists of an adaptive filter that acts on the reference signal to produce an estimate of the noise, which is then subtracted from the primary input. The role of this method is to estimate signal *s(n)* by using of filtered signal n(n). It will be done by minimising of mean square error (MSE) $E[e_c^2(n)]$ where $e_c(n) = d(n)$. The filter *H(z)* represents transfer function of an environment. Block 1 is a model of speech degradation process. Block 2 is an adaptive noise canceler.

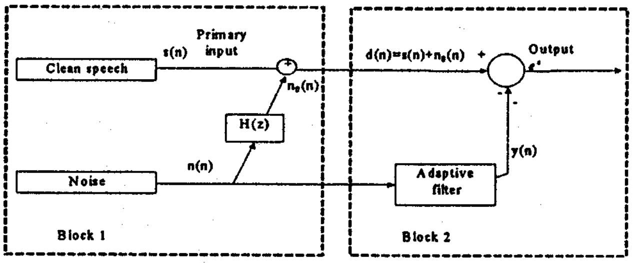

Fig.1 Adaptive noise canceler.

Direct time-domain mapping is most straightforward use of neural networks for speech enhancement. A multi-layer network is used to map a windowed segment of the noisy speech to an estimate of the clean speech. The number of inputs depends on the sampling rate of the speech signal and typically is set to cover 5 to 10 ms of data. To train the network, clean speech is artificially corrupted to create noisy input data. The clean speech signal is used as a target that is time-alignet with the inputs.

Data is presented by sliding the input window across the noisy speech. At each step, the window is shifted by an increment L, between 1 and the window length M. When the increment equals estimation window length, the estimation window moves along without overlap. For increments L<M, the resultant overlapping windows provide redundant estimates. In the case of a single time step increment, L = 1 (which most closely corresponds to our filter implementations), the network topology could be simplified to have only a single output. However, it has generally been recognised that using multiple outputs aids

in the training process. The extra outputs balance the forward and backward flow of signals during training, and allow for a greater number of shared hidden units to be used with improved generalization. After training, the estimate can be taken from one of the center-most outputs, discarding the rest.

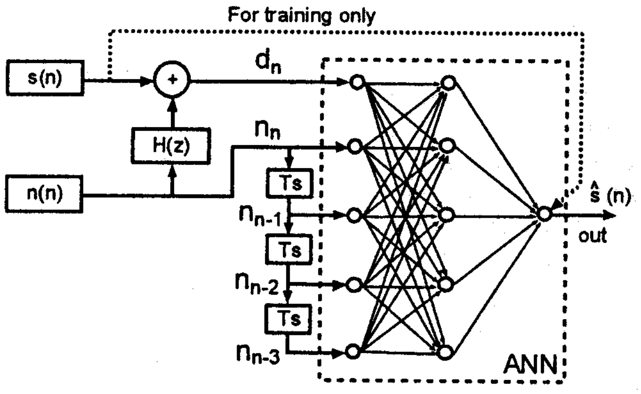

Fig.2 The scheme of implemented method.

In our setting is block 2 substituted by the ANN (Fig.2). We don't assume an adaptive filter. It means the ANN will not vary during the real time process. The ANN is implemented as feedforward fully connected network. It has been trained by back propagation algorithm with momentum and adaptive learning rate.

5. Experiments, results and discussion

Speech vowels were used for training the neural network. These signals was corrupted by a random noise. The magnitude of noise (noise coefficient) was gradually increased. The number of epochs (training time) increase with increasing of noise coefficient. During experiments was observed an influence of changing some parameters of the ANN on SNR improvement. Changed parameters was for example: the number of neurons in the hidden layer, order of filter, matrix of input patterns. The ANN contents of five neurons in input layer five (or 2, 8, 20) in one hidden layer and one output neuron. There was used purelin and transig activation functions.

In first case the ANN was trained with different values of noise coefficient K. (For example if noise coefficient is 2 it means, the noise is 2 times bigger as speech signal.)

In Tab. 1 are shown results for speech vowels. The filter of 8-th order was used. The maximum number of input pattern was 3000 and number of rows in input matrix was 10 and number of neurons in hidden layer was eight. N_ep is the number of training epochs and F_err is the final error.

In next experiment the ANN was trained by noised signal K = 10. Trained ANN processed a signal with different values of K. Other parameters of ANN in all consequent experiments were the same. The maximum number of input pattern was 1000. The results are shown in Tab. 2.

The number of neurons in the hidden layer was changing in the last case. S1 is the number of neurons in the hidden layer. Lin is the number of line.

K	Training process				Real time process	
	N_ep	F_err [10^{-6}]	S/N (in) [dB]	S/N (out) [dB]	S/N (in) [dB]	S/N (out) [dB]
0,1	2949	998	30,0032	62,4643	29,3986	51,4414
0,2	675	895	23,2493	62,9371	24,1989	51,9087
0,4	382	926	17,7572	62,7888	18,1192	51,4056
0,6	256	406	13,8675	63,3649	14,9911	53,6371
0,8	291	979	11,2141	62,5459	12,7954	52,3471
1	253	875	9,7077	63,0321	10,3028	53,5064
2	220	535	3,0531	63,1681	3,6282	53,5144
5	560	996	-3,5892	62,4713	-4,3381	53,6373
10	1047	967	-9,9763	62,5986	-9,1820	52,6532

Tab.1

K	S/N (in) [dB]	S/N (out) [dB]
0,1	30,0508	57,7886
1	8,6446	57,7868
10	-11,6511	57,6889
50	-25,6124	57,0015

Tab.2

Lin	S1	N_ep	F_err [10^{-6}]	S/N (in) [dB]	S/N (out) [dB]
1	2	7354	996	-17,9473	57,7136
2	8	1298	953	-10,6257	57,9083
3	20	2782	997	-9,8667	57,7099

Tab.3

K	L1		L2		L3	
	S/N (in) [dB]	S/N (out) [dB]	S/N (in) [dB]	S/N (out) [dB]	S/N (out) [dB]	S/N (in) [dB]
0,1	28,3129	57,7051	28,3129	57,7051	57,9012	28,8603
1	7,3093	57,7042	7,3093	57,7042	57,9041	8,2406
10	-11,3644	57,7182	-11,3644	57,7182	57,6909	-11,6211

Tab.4

Trained ANN processed signals with different K are in Tab. 4. (L1-L3 corresponds with Lin in the Table 3.)

If order of filter exceeds the number of rows in an input matrix, an ANN will saturate and becomes nonadequate. The number of signal patterns which create an input matrix is an important parameter, because influences the processing time.

As we can see above trained ANN is robust to level of degrading noise. It works in case only the noise is the same type in training and in processing mode. It means in one setting the ANN is trained just for one type of noise. Next research should be conducted to new techniques developed which avoid the need for knowledge of the noise statistic.

6. Conclusion

The advantage of time-domain filtering is the ease and efficiency of implementation. Once trained, a single fixed neural network is used to provide speech estimates.

However, this also underscores the disadvantages of the aproach. Effectively, the neural network approximates the conditional expectation $E[x_k|y_k]$, where x_k is the clean speech and y_k is the windowed noisy input. Using a fixed network implies a single, fixed expectation inferred from the entire training set. Assuming the expectation is constant is equivalent to assuming that both x_k and y_k , have constant density functions. In other words, both the noise and the speech signals would have to be stationary processes. This is clearly not the case. While training on a variety of different SNRs and speakers can greatly improve generalisation, this does not explicitly account for the nonstationarity. Some researches have incorporated pitch information as additional inputs to attempt to account somewhat for the nonstationarity of the speech. Variations in the speech and noise statistics can also be addressed by using an estimate of the time-specific SNR as an additional input to the network.

In general, the direct time-domain filtering approach is most applicable for reducing fixed noise types, or for compensating a distortion that is associated with a specific recording or communication channel.

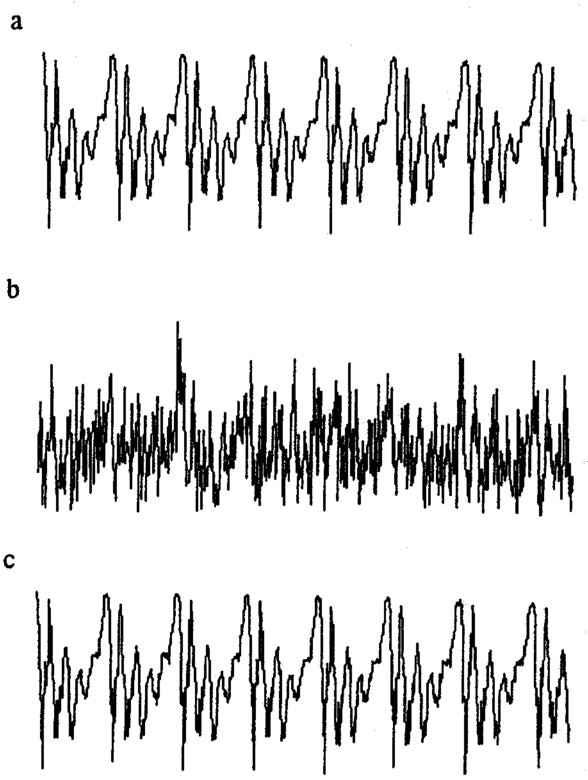

Fig.3 An example of clean (a), noisy (b) and cleaned vowel (c).

References

[1] HAYKIN S.: Neural Network – A Comprehensive Foundation, Macmillian Publishing, 1994.

[2] Neural Network Toolbox User's Guide, The Math Works, Inc., 1989.

[3] DELLER,J.R.-PROAKIS,J.G.-HANSEN,J.H.L.: Discrete-Time Processing of Speech Signals, Prentice Hall, 1993.

[4] WAN,E.A.-NELSON,A.T.: Networks for Speech Enhancement, in Handbook of Neural Networks for Speech Processing, Artech House, 1998.

About authors...

Jaroslav TLUČÁK was born in Brezno, Slovakia 1975. He graduated from the Technical University in Kosice in 1998, then he started Ph.D. study at Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Kosice. His work includes speech enhancement systems.

Jozef JUHÁR was born in Poproč, Slovakia in 1956. He graduated from the Technical University in Kosice in 1980. He received Ph.D. degree in Radioelectronics from Technical University of Košice, in 1991. Now he is Assistant Professor at Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice. His research interests include speech and audio processing, and communications systems.

Ľubomír DOBOŠ was born in Vranov n/T, Slovakia in 1956. He graduated from the Technical University in Kosice in 1980. He received Ph.D. degree in Radioelectronics from Technical University of Kosice, in 1989. Now he is Assistant Professor at Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Kosice. His research interests include adaptive filtering, neural networks, and wireless communications systems.

Anton ČIŽMÁR was born in Michalovce, Slovakia in 1956. He graduated from the Slovak Technical University in Bratislava in 1980. He received Ph.D. degree in Radioelectronics from Technical University of Kosice, in 1986. Now he is Associate Professor at Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Kosice. His research interests include speech processing, data compresion and digital broadband communications.