SPEECH SEPARATION BY SIMULATING THE COCKTAIL PARTY EFFECT WITH A NEURAL NETWORK CONTROLLED WIENER FILTER

Yuchang Cao, Sridha Sridharan and Miles Moody

Signal Processing Research Centre
School of Electrical and Electronic Systems Engineering
Queensland University of Technology, GPO Box 2434, Brisbane Q4001, Australia
email: s.sridharan@qut.edu.au

ABSTRACT

A novel speech separation structure which simulates the cocktail party effect using a modified iterative Wiener filter and a multi-layer perceptron neural network is presented. The neural network is used as a speaker recognition system to control the iterative Wiener filter. The neural network is a modified perceptron with a hidden layer using feature data extracted from LPC cepstral analysis. The proposed technique has been successfully used for speech separation when the interference is competing speech or broad band noise.

1. INTRODUCTION

Human beings have the ability to concentrate on speech of interest while suppressing sound from other sources. This emphasis on a particular sound source and rejection of the other sound sources has been referred to as the cocktail party effect or "attentional selectivity". In 1953, Cherry reported his pioneering research work on this problem [1]. He found that the difference between the signals arriving at the two ears is essential for the cocktail party effect. Cherry mentioned several factors that he thought would have contributed to separate the desired voice from others. Some of those factors were raw physical qualities of the speech, but others had to do with the ability of the listener to predict the next moment of speech from the previous one. It is still not clear as to whether this ability to predict is really used in segregating desired speech from other sources. However, there is no doubt that the listener's knowledge of the voice and language governs the process of concentrating on a particular speech [2].

Several theories were invoked to account for the cocktail party effect demonstrated by Cherry. However, they were all similar in many ways. They all postulated that the physical properties of one of the voices could be used by the listener's attention to select that voice, and those properties include the locations of sources, the quality of voices, the pitch, timbre as well as loudness. Further theoretical study on this problem is being conducted.

Experiments have been carried out to understand and duplicate the ability of attentional selectivity for speech acquisition, enhancement and recognition [3]. These methods use dual microphones or a microphone array and are derived from "the binaural perception scene" which is related to the phase difference between the signal arriving at the two ears. Under some conditions this processing has exceeded our natural ability to discriminate a desired speech signal from a distracting background. Most of these processes, however, introduce distortion into the resulting processed signal [4]. These methods have found limited applications in real speech enhancement and recognition because they were only based upon the phase information and failed to take many other factors mentioned above into account in the simulation of the cocktail party effect.

In this paper, we use more information related to the cocktail party effect to simulate it for speech enhancement based on two channel observations in which the relative levels of the desired speech and interference are different. Our key idea is to use a single or a number of modified multi-layer perceptron neural networks as a speaker discrimination unit (speech classifier) to control an iterative filter. At the beginning of the iterative filtering for each block, a step factor α in the modified Wiener filter is chosen according to the difference between the desired speaker recognition rate of the main channel input and that of the reference channel input. The iterative filtering then continues until the recognition rate receives its first maximum (or the optimal) value. It has been found that the iterative process uniformly converges towards the first maximum (or the optimal) recognition rate and best quality simultaneously under several conditions that were tested.

2. METHODOLOGY

The proposed system has a structure depicted in Fig.1. The speech data acquisition system supplies both main channel and reference channel signals to the modified iterative Wiener filter. The data acquisition system may consist of two microphones or alternatively a dual beamformer based on a microphone array to obtain the two channels of input. Each of the two channels will contain both the desired speech and interference but with different signal-to-interference (or signal-to-noise) ratios. In the case where the data acquisition is carried out by a single microphone system and the interference is broad band noise (which can be modelled as a wide-sense stationary random process), the reference channel signal can be extracted from the non-speech segments of the acquired signal. The iterative Wiener filter attenuates the interference components con-

tinuously while maintaining the desired speech at a certain level of magnitude for each frame of the main channel signal until the neural network decides that the average quality of the filter output at the current iteration is the best for the frames of a block.

2.1. Iterative Wiener filter

An enhancement system based on an iterative Wiener filter using the estimation of all-pole speech parameters was investigated by Lim, Oppenheim, Hansen and Clements [4, 5]. This approach attempts to solve for the maximum *a posteriori* (MAP) estimate of a speech waveform in additive Gaussian noise, with the requirement that the signal be the response from an all-pole process. The frequency response of the non-causal Wiener filter is

$$H(\omega) = \frac{P_{s_1}(\omega)}{P_{s_1}(\omega) + P_{s_2}(\omega)} \tag{1}$$

where $P_{s_1}(\omega)$ and $P_{s_2}(\omega)$ are power spectral densities of the desired signal $s_1(t)$ and the noise $s_2(t)$, respectively. Obviously, the Wiener filter of Eq.1 cannot be applied directly to estimate the desired signal since the spectrum $P_{s_1}(\omega)$ cannot be assumed known, even if $s_2(t)$ has been assumed to be a stationary process with *a priori* Gaussian probability density function and can be extracted from non-speech segments. A traditional iterative approach takes the spectrum of noisy speech $Y(\omega)$ or the spectral subtraction estimator $\hat{S}_1(\omega)$ as an estimated spectrum at the beginning of the loop and uses it to form an iterative Wiener filter. Although successful in a mathematical sense, these techniques have received little application in practice. The main drawback of these techniques is that no effective procedures exist to create a convergence criterion in environments requiring automatic speech enhancement [5].

Recently Hansen, Clements and Nandkumar have developed a constrained Wiener filtering system for which the convergence criterion is based on objective speech quality measures [5]. Such measures are formed by a weighted comparison of actual and resulting estimated LPC predictor coefficients found during enhancement. The obvious problem with such a criterion for practical applications is that, the actual speech is unknown during the procedure. Hansen *et al* have found some experiential termination point for the iteration. Note that the constrained Wiener filters studied by Hansen *et al* in both single and dual channel systems are based on an *a priori* assumption: The interference is a non-speech-like additive background noise. The methods, including the convergence criterion employed in the systems, cannot enhance a noisy speech which has been corrupted by competing speech.

In our approach, the disadvantage of the traditional Wiener filter has been overcome successfully by use of a neural network controller. Our system can work in both situations where speech has been corrupted by stationary non-speech-like noise or by competing speech.

For a dual channel system shown with a coupling function

$$H(\omega) = \begin{bmatrix} H_{11}(\omega) & H_{12}(\omega) \\ H_{21}(\omega) & H_{22}(\omega) \end{bmatrix} \tag{2}$$

the form of the modified Wiener filter and the control procedure with the neural network is achieved as follows: Firstly, to obtain an effective and smooth iterative filtering, the ith iteration of the Wiener filter of Eq.1 is modified by adding a time varying parameter α_i which controls the "step" of filter as follows:

$$H^{(i)}(\omega) = \frac{(1+\alpha_i)P_{\hat{s}_1^{(i)}}(\omega)}{P_{\hat{s}_1^{(i)}}(\omega) + \alpha_i(|\hat{H}_{11}(\omega)|/|\hat{H}_{21}(\omega)|)^2 P_{y_2}(\omega)} \tag{3}$$

and

$$\hat{S}_1^{(i+1)}(\omega) = \hat{S}_1^{(i)}(\omega)H^{(i)}(\omega) \qquad for \;\; i = 0,\, 1,\, \cdots \tag{4}$$

The ratio $(|\hat{H}_{11}(\omega)|/|\hat{H}_{21}(\omega)|)^2$ can be calculated from the non-interference segments. $\hat{S}_1^{(i)}(\omega)$ is the spectrum of the filtered (at the ith iteration) speech signal $\hat{s}_1^{(i)}(t)$. At the beginning of the iterative loop for each block, $\hat{s}_1^{(0)}(t)$ is replaced by $y_1(t)$.

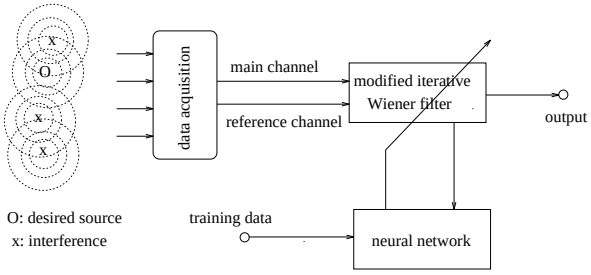

Fig.1. Structure of the proposed speech enhancement system. The data acquisition system could be a pair of microphones, a microphone array or a single microphone.

The coefficients of the modified Wiener filter change frame by frame (short term) while the factor α_i is updated block by block (long term). For each block, α_0 may be chosen according to the difference between the signal-to-interference ratio of the main channel input $y_1(t)$ and that of the reference channel input $y_2(t)$. An alternative method which we use in our system for the factor α_0 is based on the difference between the recognition rates of the two channels. Here, the recognition rate is defined as the proportion of the number of frames, within a block, which are recognised as the desired speech. For example, the recognition rate of the single binary classifier model (SBCM) is defined as

$$R_{rec} = M_0/M \tag{5}$$

where M is the number of frames in a block, M_0 is the number of frames within the block which are recognised as the desired speech.

After the main channel input $y_1(t)$ is substituted for $\hat{s}_1^{(0)}(t)$ at the beginning of the iterative loop, the iterative filtering loop (Equations 3 and 4) continues until that it is terminated when the neural network decides that the output signal is the closest to the desired voice, i.e., when the recognition rate of $\hat{s}_1^{(i)}(t)$ receives its first maximum value.

For the single channel system, the modified Wiener filter Eq. 3 is simplified as

$$H^{(i)}(\omega) = \frac{P_{\hat{s}_1^{(i)}}(\omega)}{P_{\hat{s}_1^{(i)}}(\omega) + \alpha_i P_{\hat{s}_2}(\omega)} \tag{6}$$

where $P_{\hat{s}_2}(\omega)$ is the power spectrum of the reference signal $\hat{s}_2(t)$ which may be extracted from the non-speech segments. The value of α_0 in this single channel system can be chosen according to the recognition rate of the main channel input.

2.2. Artificial Neural network (ANN)

The ANN adopted in this study is the Logicon Projection Network (LPN) distributed by Neuralware Inc. [6, 9]. The basic structure of LPN is shown in Fig. 2. This network model projects the N_{in} dimensional input vectors of a standard feed forward neural network onto a hypersphere in a higher dimension $(N_{in}+1)$ space before implementing a modified back-propagation algorithm. The Logicon Projection Network utilises a modified back-propagation algorithm for training.

Input vectors for neural network training are the parameterised speech as well as alternative data. The alternative data may consist of parameterised speech from competing speaker(s) and reference speech signals, or background noise. The parametrisation is created by use of Linear Predictive Coding (LPC) cepstral analysis [7].

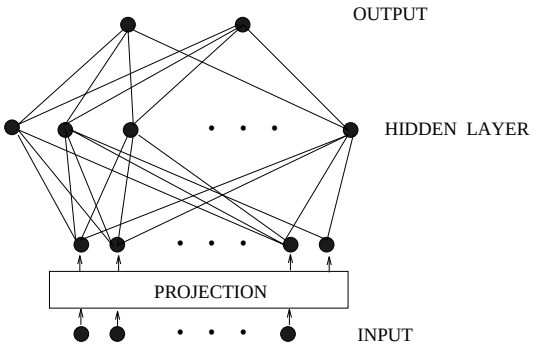

Fig.2. Logicon Projection Network. The input vectors are projected to vectors in one higher dimension

2.3. Architecture of the classifier

2.3.1. The single binary classifier model

An SBCM, which consists of a single LPN, used to calculate the recognition rates of the enhanced speech for each iteration is shown in Fig. 3. In this model, the speech of the desired speaker and the speech of the competing speaker (for co-talker separation) or noise (for background noise cancellation) is parameterised and concatenated to form training vectors. At the test stage, the enhanced speech is parameterised before it is input into the classifier. It is found that recognition rate increases iteration by iteration before the output of the Wiener filter receives the best quality. After the output receives the best quality, the enhanced speech will be over-filtered and its quality will be degraded if the iterative process goes too far.

It has been found after a large amount of simulation tests that there is a high correlation between the quality of enhanced speech signal and the recognition rate for each iteration. However, even though the over-filtering degrades the quality of the desired speech, it depresses the competing speech further at the same time. This may cause the SBCM to select an incorrect point for termination of the iterations in some cases for co-talker separation.

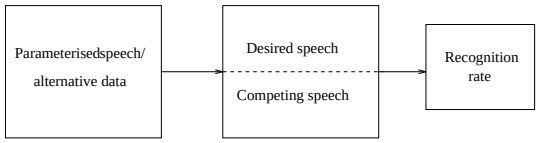

Fig.3.Single binary classifier model with a single LPN.

2.3.2. The multiple binary classifier model

To improve the accuracy of classification and the robustness of the system for co-talker separation, the SBCM can be replaced by the multiple binary classifier model (MBCM). The motivation for developing the MBCM in the proposed speech enhancement system is derived from statistical analysis. The architecture of MBCM as shown in Fig.4, consists of a set of LPNs. In this model, the top LPN is trained with the parameterised data formed by the speech of the desired speaker and the speech of the competing speaker. The rest of the LPNs are trained with the data formed by the speech of the desired speaker and that of one of the reference speakers, respectively. The reference speakers are chosen arbitrarily from a database.

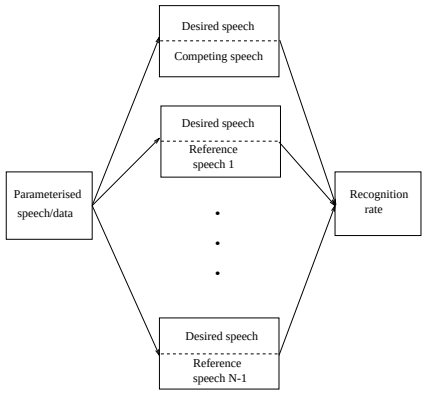

Fig.4. Multiple binary classifier model with N LPNs.

The enhanced speech after each iteration is parameterised and then fed to all trained LPNs. If M_0 frames among M frames are classified by the top LPN as the desired speech and $M_1 \cdots M_{N-1}$ frames are classified by the other LPNs as the desired speech, respectively, the recognition rate of the MBCM for the output at this iteration is defined as

$$R_{rec} \stackrel{\text{def}}{=} \frac{2}{N+1}\left(\frac{M_0}{M}+\frac{1}{M}\left(\sum_{i=1}^{N-1} M_i-(N-1)\frac{M}{2}\right)\right) \tag{7}$$

or

$$R_{rec} \stackrel{\text{def}}{=} \frac{2}{(N+1)M}\sum_{i=0}^{N-1} M_i-\frac{N-1}{N+1} \tag{8}$$

where N is the number of LPNs and M is the frames of the block (or utterance). The last term in Eq.7 (as well as in Eq.8) is an unbiasing factor.

3. SIMULATION TESTS

A large number of tests have been carried out to verify the performance of the proposed system. The relationship between the convergence of iterative processing and the improvement of quality has been examined. The improvement in speech quality for the desired speech was measured by both objective measurements and informal listening tests. The objective measurements for the speech quality used in the study include traditional objective speech quality measures such as the logarithmic area ratio (LAR), the logarithmic spectral distance (LSD), Itakura ratio (IR) and the segmental signal-to-noise ratio (SNRseg). Details about these measures may be found in [8].

A set of experiment for separation of co-talkers was arranged as follows: All utterances of a total of 40 speakers were extracted from the *TIMIT* database (female and male speakers from dr1 and dr2, with sampling rate of 16kHz). For each speaker, all utterances were concatenated together to form a speech signal. The first 14.4 seconds (i.e., 230,400 samples) of each speech signal were used to create the training data and the corresponding test data was selected from the rest of the signal. Each ten speech signals were grouped together for a test: one as the desired speech signal, one as the competing speech signal and the other 8 speech signals as the reference signals.

(A) Single Binary Classifier Model):

Iteration	i_M-2	i_M-1	i_M	i_M+1	i_M+2
LAR	1.5	20.9	75.6	2.0	0
LSD	1.5	20.9	76.1	1.5	0
IR	2.0	22.4	75.1	0.5	0
SNRseg	1.5	18.9	77.1	2.5	0

(B) Multiple Binary Classifier Model):

Iteration	i_M-2	i_M-1	i_M	i_M+1	i_M+2
LAR	0	3.5	95.0	1.5	0
LSD	0	3.5	95.5	1.0	0
IR	0	6.0	93.5	0.5	0
SNRseg	0	2.0	95.0	3.0	0

Table 1. Percentage coincidence between the first maximum recognition rate and the quality measures.

For each quality measure the percentage of tests which obtained the best quality at iterations centred around i_M is seen in the columns of Table 1. The top one (A) shows the results for the single binary classifier model and the bottom one (B) shows the results when the multiple binary classifier model was used. Note that i_M is the iteration at which the output of the enhanced speech received its the first maximum recognition rate.

For cases when the SBCM is used, it can be seen from the table that the percentage coincidence between the first maximum recognition rate and the best quality measures is about 76%. It has been found that the SBCM recognition rate may receive its first maximum value one iteration after the quality measures had got their best values in some tests. This error is eliminated by use of the MBCM. In fact, the percentage coincidence between the maximum recognition rate and the quality measures increased to 95% (average for the all quality measures listed in Table 1B at iteration i_M).

4. CONCLUSION

We have described the principles and structure of a speech enhancement system in which a neural network simulates the cocktail party effect of the human auditory perception system. The neural network controls an iterative Wiener filter which enhances the corrupted speech signal from observations. Both objective and subjective measures (informal listening) have been utilized to evaluate the performance of the proposed system. The test results have shown that the proposed system is powerful and reliable for speech enhancement because of the use of the neural network to mimic the human auditory perception system. A limitation of the method is that sufficient quantity of the desired speakers' speech samples must be available to train the system.

Reference:

[1] E. Colin Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears", *J. of The Acoust. Soc. of Amer.*, Vol.25, No.5, pp975-9, Sept. 1953.

[2] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organisation of Sound"*, The MIT Press, Cambridge, London, 1990.

[3] O. M. Mracek Mitchell, C. A. Ross, and G. H. Yates "Signal Processing for a Cocktail Party Effect", *J. of The Acoust. Soc. of Amer.*, Vol.50, No.2 (Pt2), pp656-60, Aug. 1971.

[4] J.S. Lim, *Speech Enhancement.* Prentice Hall Inc., Englewood Cliffs, New Jersey, 1983.

[5] J.H.L. Hansen and M.A. Clements, " Constrained Iterative Speech Enhancement with Application to Speech Recognition", in *IEEE Trans. Signal Processing*, Vol. 39, No. 4, pp795-805, April 1991.

[6] Neuralware Inc. "Neural Computing - A Technology Handbook for Professional II/Plus and Neural work Explorer", Technical Report, Technical Publication Group, Suite 227, Pann Centre West, Pittsburge PA15276, 1993.

[7] Furui, S., *Digital Speech Processing, Synthesis, and Recognition*, Marcel Dekker, Inc. New York, 1989

[8] S.R. Quackenbush, T.P. Barnwell and M.A Clements, *Objective Measures of Speech Quality.* Englewood Cliffs, N.J. Prentice Hall, 1988.

[9] P. Castellano and S. Sridharan, "Speaker Identification with Concomitant Open and Closed Decision Boundaries", *Aust. J. Intel. Inf. Proc. Syst.*, pp47-53, Vol.2(2), 1995.