INTERSPEECH 2009 BRIGHTON

Enhancement of Binaural Speech Using Codebook Constrained Iterative Binaural Wiener Filter

Nadir Cazi and T.V. Sreenivas

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore, India 560012
nadir@ee.iisc.ernet.in, tvsree@ece.iisc.ernet.in

10.21437/Interspeech.2009-415

Abstract

A clean speech VQ codebook has been shown to be effective in providing intraframe constraints and hence better convergence of the iterative Wiener filtering scheme for single channel speech enhancement. Here we present an extension of the single channel CCIWF scheme to binaural speech input by incorporating a speech distortion weighted multi-channel Wiener filter. The new algorithm shows considerable improvement over single channel CCIWF in each channel, in a diffuse noise field environment, in terms of *aposteriori* SNR and speech intelligibility measure. Next, considering a moving speech source, a good tracking performance is seen, upto a certain resolution.
Index Terms: binaural speech, iterative Wiener filtering, codebook constraint, time delay of arrival, source tracking

1. Introduction

Binaural hearing aids use inputs from both the left and right hearing aid to generate an output for each ear as compared to a monaural hearing aid which uses only monaural input to generate output for the specific ear. It is well known that binaural hearing provides better noise immunity than monaural, as evidenced by the binaural masking level difference (BMLD) in psycho-acoustics. A binaural algorithm can exploit the two channel advantage and provide better enhancement performance as compared to a monaural algorithm. This would be useful for not only hearing aids, but also other speech applications in the presence of noise. Typically, at a given time, one of the two channels would have a higher signal-to-noise ratio (SNR) which can be exploited by the enhancement algorithm. Compared to monaural systems, a binaural system provides improved capability in suppressing the effect of noise on speech intelligibility. In [1], an efficient Voice Activity Detector (VAD) is designed based on a simplified binaural model in order to detect the speech pauses. This is combined with psychoacoustically motivated spectral subtraction applied on each channel independently. Here, the interaural differences between the two channels are utilized in order to find out the dominant source direction for each frame, and thus, better detect speech pauses. In [2], they have explored the integration of a Adaptive Noise Cancellation (ANC) technique with the binaural model based VAD, where an intermittent ANC is utilized to cancel the noise estimated during the speech pauses. In [3], a multichannel Wiener filter with interaural transfer function extension has been proposed. The cost function in the Wiener filter formulation is modified to give a relative weightage to the speech distortion and noise reduction separately. The relative weightage parameter provides for an effective design of the noise reduction algorithm that does not introduce any adverse processing artefacts, such as distortion of the speech signal itself or the interaural cues which are needed by the user to correctly localize the sounds.

To enhance single channel speech in additive noise, the technique of Iterative Wiener Filtering (IWF), introduced by Lim and Oppenheim [4], is a sequential maximization of the a posteriori probability (MAP) of the speech signal and its all-pole parameters. The IWF is investigated further by Hansen and Clements [5] with incorporation of auditory motivated spectral constraints. It is found that two types of constraints are needed: (a) the *inter-frame* constraint which helps in preserving the speech spectral continuity, and (b) the *intra-frame* constraint which uses the correlation in the formant frequencies of speech signal. Rule based schemes are used to incorporate these constraints to enhance speech spectral parameters. To use the intra-frame constraint, Sreenivas and Kirnapure [6] have shown that a VQ codebook approach is successful in utilizing the redundancy between the spectral parameters. In this codebook constrained IWF (CCIWF) method, the enhanced speech is constrained to belong to a codebook of clean speech spectra based on minimizing a perceptually relevant distance measure.

In this paper, we develop an iterative *binaural* Wiener filtering algorithm (CCI***B***WF) which makes use of the codebook constrained approach to estimate the clean speech of each channel; the noise is assumed to be diffuse. Separate VQ codebooks are designed for each channel and for each quantized level of the time delay of arrival (TDOA) between the two channels. We show the advantage of using a joint binaural IWF over single channel IWF in each channel. We then explore the trade-off between the speech distortion and noise reduction. We also obtain an objective measure for the accuracy of localization of the source in terms of the difference between TDOA for the enhanced output and the desired TDOA. Performance of the new algorithm for a moving speech source is studied in terms of estimating the TDOA required in selecting the VQ codebook pair.

2. Binaural Wiener Filtering

The overall scheme of the new algorithm is shown in Figure 1.

Let us consider a typical situation for the binaural enhancement model. There is one microphone at each ear. The speaker is positioned at a certain direction with respect to the listener. We consider the background noise to be the result of a diffuse noise field. A diffuse noise field is when the resulting noise at the two ears comes from all directions, with no particular dominant direction. The received signals at the two ears can be expressed in frequency domain as below:

$$\begin{bmatrix} X_1(\omega) \\ X_2(\omega) \end{bmatrix} = \begin{bmatrix} S_1(\omega) \\ S_2(\omega) \end{bmatrix} + \begin{bmatrix} N_1(\omega) \\ N_2(\omega) \end{bmatrix} \quad (1)$$

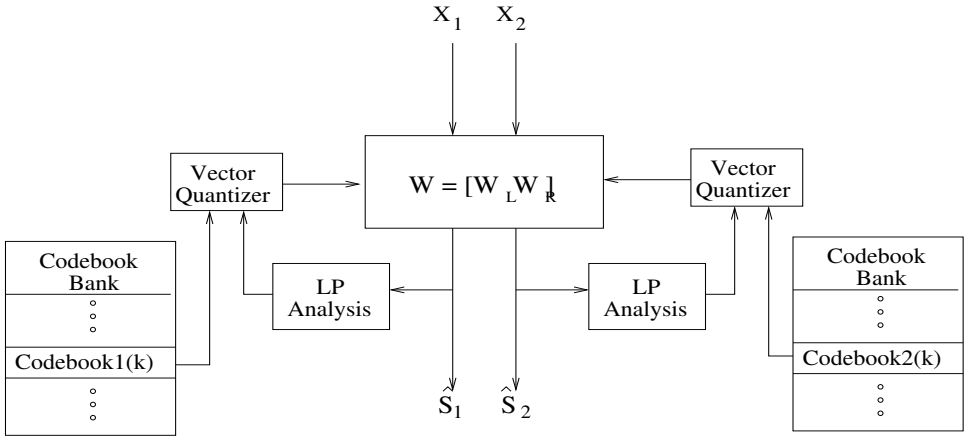

Figure 1: *Binaural CCIWF scheme with an optimum two-channel Wiener filter.*

In the above equations, $\mathbf{S}(\omega) = [S_1(\omega)\ S_2(\omega)]^T$ represents the vector of the speech components in the left and right channels respectively. Similarly, $\mathbf{N}(\omega) = [N_1(\omega)\ N_2(\omega)]^T$ denote the noise components at the respective ears, which are assumed diffuse and uncorrelated.

Our aim is to find out the weight vectors $\mathbf{W}_L(\omega) = [W_{11}(\omega)\ W_{12}(\omega)]^T$ and $\mathbf{W}_R(\omega) = [W_{21}(\omega)\ W_{22}(\omega)]^T$ for the left and right channel such that a combined cost measure is minimized. For convenience, we define $\mathbf{W}(\omega)^H = \left[\mathbf{W}_L(\omega)^H\ \mathbf{W}_R(\omega)^H\right]$ and $\mathbf{X}(\omega) = [X_1(\omega)\ X_2(\omega)]^T$. We have already defined $\mathbf{S}(\omega)$ and $\mathbf{N}(\omega)$ as similar 2×1 vectors.

In the frequency domain, each frequency component is processed independently. Hence we may omit the variable ω.

The conventional Wiener filter cost function for signal estimation, $J(\mathbf{W}) = \mathcal{E}\{\|\mathbf{S} - \hat{\mathbf{S}}\|^2\}$:

$$\text{Thus, } J(\mathbf{W}) = \mathcal{E}\left\{\left\|\mathbf{S}^T - \mathbf{W}^H \begin{bmatrix} \mathbf{X} & \mathbf{0_{2\times 1}} \\ \mathbf{0_{2\times 1}} & \mathbf{X} \end{bmatrix}\right\|^2\right\}$$
$$= \mathcal{E}\left\{\left\|\mathbf{S}^T - \mathbf{W}^H \begin{bmatrix} \mathbf{S} & \mathbf{0_{2\times 1}} \\ \mathbf{0_{2\times 1}} & \mathbf{S} \end{bmatrix}\right\|^2 + \left\|\mathbf{W}^H \begin{bmatrix} \mathbf{N} & \mathbf{0_{2\times 1}} \\ \mathbf{0_{2\times 1}} & \mathbf{N} \end{bmatrix}\right\|^2\right\} \quad (2)$$

The cross-terms are 0 because the speech and noise is assumed zero mean and uncorrelated. We recognize the first term represents the speech distortion energy while the second term represents the residual noise energy. We consider the following modified cost function, which is the weighted sum of the residual noise energy and the speech distortion energy:

$$\text{Thus, } C(\mathbf{W}) = \mathcal{E}\left\{\left\|\mathbf{S}^T - \mathbf{W}^H \begin{bmatrix} \mathbf{S} & \mathbf{0_{2\times 1}} \\ \mathbf{0_{2\times 1}} & \mathbf{S} \end{bmatrix}\right\|^2\right\}$$
$$+\mu\,\mathcal{E}\left\{\left\|\mathbf{W}^H \begin{bmatrix} \mathbf{N} & \mathbf{0_{2\times 1}} \\ \mathbf{0_{2\times 1}} & \mathbf{N} \end{bmatrix}\right\|^2\right\} \quad (3)$$

By controlling the parameter μ we are able to give different weightage to the speech distortion and the residual noise as a tradeoff between enhanced speech quality and intelligibility.

Minimizing the above cost function over $\mathbf{W}$ assuming the speech signal to be uncorrelated with the noise in either channel and the noises in each channel being uncorrelated with each other, we get

$$\mathbf{W}_{opt} = \begin{bmatrix} \mathbf{R_S} + \mu\mathbf{R_N} & \mathbf{0} \\ \mathbf{0} & \mathbf{R_S} + \mu\mathbf{R_N} \end{bmatrix}^{-1}_{\mathbf{4\times 4}} \begin{bmatrix} P_{S_1} \\ P_{S_{21}} \\ P_{S_{12}} \\ P_{S_2} \end{bmatrix}_{\mathbf{4\times 1}}$$

where, (4)

$$P_{S_1} = \mathcal{E}\{S_1 S_1^*\} \quad P_{S_{21}} = \mathcal{E}\{S_2 S_1^*\}$$
$$P_{S_{12}} = \mathcal{E}\{S_1 S_2^*\} \quad P_{S_2} = \mathcal{E}\{S_2 S_2^*\}$$
$$P_{N_1} = \mathcal{E}\{N_1 N_1^*\} \quad P_{N_2} = \mathcal{E}\{N_2 N_2^*\}$$

also $\mathbf{R_S} = \mathcal{E}\{\mathbf{S}\mathbf{S}^*\} = \begin{bmatrix} P_{S_1} & P_{S_{12}} \\ P_{S_{21}} & P_{S_2} \end{bmatrix}_{\mathbf{2\times 2}}$ and $\mathbf{R_N} = \mathcal{E}\{\mathbf{N}\mathbf{N}^*\} = \begin{bmatrix} P_{N_1} & 0 \\ 0 & P_{N_2} \end{bmatrix}_{\mathbf{2\times 2}}$

We estimate P_{S_1} and P_{S_2} from the codebook constrained iterative estimation of speech parameters for each channel. However the estimate of $P_{S_{12}}$ involves both phase and magnitude; the magnitude is equal to $\sqrt{P_{S_1} P_{S_2}}$ but, the phase is equal to that of $S_1 S_2^*$ averaged over a large number of frames. Since linear prediction (LP) is a minimum phase model, LP codebook based P_{S_1} and P_{S_2} does not provide for the phase difference between S_1 and S_2 exactly. However the averaging over several frames is expected to capture the significant phase differences between the two channels.

$$\text{Thus, } P_{S_{12}} = \sqrt{P_{S_1} P_{S_2}} \cdot \mathrm{e}^{j\phi_{12}} = P_{S_{21}}^*$$

If we denote,

$$H_1 = \frac{P_{S_1}}{P_{S_1} + \mu P_{N_1}}$$
$$H_2 = \frac{P_{S_2}}{P_{S_2} + \mu P_{N_2}}$$

we can simplify equation (3) to get signal estimates as:

$$\hat{S}_1 = \mathbf{W}_L^H \mathbf{X}$$
$$= \frac{H_1(1 - H_2)X_1 + \sqrt{\frac{P_{S_1}}{P_{S_2}}}\mathrm{e}^{j\phi_{12}} H_2(1 - H_1)X_2}{1 - H_1 H_2} \quad (5)$$

Similarly, we get an expression for $\hat{S}_2$ also.

From equation (4), we can see that $\hat{S}_1$ is dependent on both input channels. H_1 and H_2 are directly dependent on the SNR of the 2 channels. If the SNR in one channel is much lower than that in the other channel, say if $H_2 \ll H_1$ and also $H_2 \ll 1$, then we get from (4)

$$\hat{S}_1 \approx H_1 X_1\ ,$$

since $(1 - H_1 H_2) \approx 1$, $H_2 \approx 0$ and $(1 - H_2) \approx 1$. This is equivalent to the case of using CCIWF in the left channel independently.

But for the same case, the S_2 estimate is given by:

Table 1: *Average segmental SNR ($\overline{SSNR}$) and Average LLR ($\overline{LLR}$) distance for enhanced speech at 5,0 and -5 dB input SNR in left channel for increasing number of iterations of CCI**B**WF and for monaural CCIWF for each channel;speech source directly in front of listener;$\mu = 2$.*

Speech Type	5 dB		0 dB		-5 dB	
	$SSNR$(dB)	LLR	$SSNR$(dB)	LLR	$SSNR$(dB)	LLR
Noisy	-6.801	0.521	-11.784	0.676	-16.770	0.814
monaural CCIWF	2.507	0.414	0.314	0.564	-1.651	0.731
CCI**B**WF						
+1 iteration	2.844	0.376	0.360	0.532	-1.842	0.695
+2 iteration	**3.183**	**0.371**	**0.721**	**0.527**	**-1.639**	**0.693**
+3 iterations	3.287	0.400	0.842	0.558	-1.545	0.717
+4 iterations	3.322	0.423	0.873	0.580	-1.513	0.734
+6 iterations	3.316	0.427	0.861	0.585	-1.535	0.733

$$\hat{S}_2 \approx \sqrt{\frac{P_{S_2}}{P_{S_1}}} \mathrm{e}^{-j\phi_{12}} H_1 X_1,$$

That is, the signal in the channel having very low SNR is estimated almost entirely from the other channel with high SNR.

3. Codebook Constrained Iterations

For single channel speech enhancement, VQ codebook approach has been effective in imposing intraframe constraints by providing better convergence along-with increasing the naturalness of speech. We perform binaural Wiener filtering as described in section 2 in an iterative fashion. At each iteration, we find the all-pole parameters of speech for each enhanced output by LP analysis and then search the respective VQ codebooks for the clean speech vectors with least distortion. These codebook constrained all-pole parameters are then used to update the coefficients of the binaural Wiener filter $\mathbf{W}$, for joint filtering leading to next iteration.

We design different pairs of VQ codebooks depending upon the interaural time difference between the two channels, which would be a result of different azimuth angles at which the speech source is located about the listener. The observed interaural time delay (ITD) is quantized in order to choose from this finite set (bank) of codebooks. The proposed CCI**B**WF algorithm needs to have a knowledge of the initial TDOA in order to select the corresponding codebook pairs needed for the constrained iterative filtering. In real life scenarios, the source might not be always in the same direction throughout a conversation. Hence in case of a gradually moving source, we use the enhanced speech outputs to estimate the gradually changing ITD. ITD is calculated using cross-correlation between the two channel signals. Thus we can track the moving source upto a certain resolution in its lateral position in terms of the time difference of arrival (TDOA).

4. Experiments and Results

4.1. Experimental Setup

The speech data has been obtained from the Indian Language Database (ILDB), IISc. We have used data amounting to 30 male and 30 female speakers, each providing about 60 sec of speech, resulting in about 3600 seconds of speech, sampled at 8kHz. Of these, 10 speakers totaling 600 seconds have been reserved for testing and the rest is used for training. The speech source is simulated to be positioned at different azimuthal angles. The speech signal at both ears is obtained by convolving

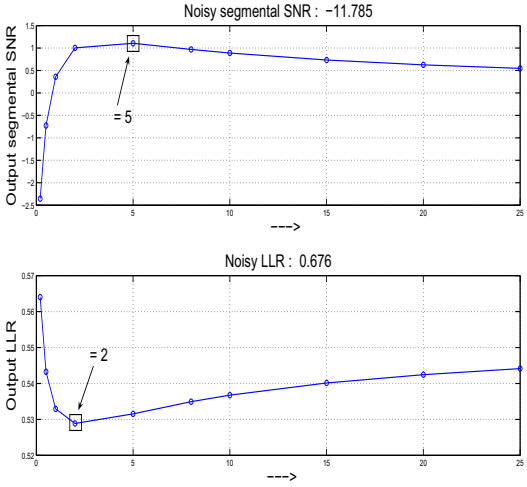

Figure 2: *Binaural speech enhancement performance as a function of the cost function parameter μ;optimum μ is shown.*

the speech with the corresponding head related transfer functions (HRTF) [7]. In order to simulate the diffuse noise environment, we add white Gaussian noise at various SNRs to the left and right channels. The noise signals in the two channels are uncorrelated. Feature vectors of clean speech are derived through LP analysis of 20ms frames. The speech source position is varied over a set of azimuth angles and the resulting binaural data is used to design the bank of LP parameter VQ codebook pairs. For the purpose of codebook design, we quantize the TDOA in steps of 1 sampling period duration, i.e. 125μsec. We consider TDOA ranging from -7 to $+7$ sample delay i.e -875μsec to $+875\mu$sec, resulting in 15 codebook pairs.

-

4.2. Results and Discussion

The performance of codebook constrained iterative binaural Wiener filter (CCI**B**WF) is shown in Table 1, for increasing number of iterations,and SNRs of -5dB, 0dB and +5dB global SNR. The speech source is simulated to be positioned directly in front of the listener (i.e. at 0^o azimuth). The comparison has been made with monaural enhancement method of using independent CCIWF for each channel (monaural CCIWF). We have shown results only for the left channel because at 0^o azimuth, both channels showed similar performance characteristics. The CCI**B**WF shows a consistent improvement over

Table 2: *Absolute ITD error for different angles of speech source to the left of the listener*

Azimuth (degree)	ITD error (μsec)
0	15.2
10	24.5
20	36.6
30	22.7
40	17.9
50	16.2
60	66.9
70	107.4
80	37.5
90	38.3

monaural CCIWF at all SNR, in terms of both average segmental SNR measure as well as average log-likelihood ratio measure. This shows that for each channel, the additional information obtained from the other channel is indeed beneficial for improving its quality, since there is a bidirectional linear association between the speech components present in each channel. Also, we observe that as we increase the maximum number of iterations allowed, we obtain best noise performance for 2 iterations itself. For more number of iterations, the performance decreases. Hence we can say that though the iterative binaural algorithm does not show a fast codebook convergence, we are able to achieve best performance in a very small number of iterations.

In Figure 2, we observe the noise performance as we vary the parameter μ, which provides a trade-off between the speech distortion energy and the residual noise energy. We see that the performance in terms of both Avg. SNR and Avg. LLR improves rapidly with increasing μ, reaches an optima, and then starts reducing gradually. Looking at the graph we see that a suitable range for μ is, $2 \leq \mu \leq 5$. We choose $\mu = 2$ in all results reported.

The binaural cues such as ITD and interaural level difference (ILD) are important for sound localization. In order to judge the enhancement algorithm in terms of degradation of the localization information, we use an absolute ITD error metric. That is, for a given source direction, we interpolate the enhanced binaural output to a higher sampling frequency (by a factor of 4) and calculate the ITD by cross-correlation and define:

$$\textbf{ITD Error} = |ITD_{enhanced} - ITD_{clean}|$$

averaged over a few seconds. We consider different azimuth angles ranging from 0^o to 90^o in steps of 10^o. We can see from Table 2 that, for a fixed direction of the source location, the CCI***B***WF algorithm is able to estimate the TDOA well within an absolute ITD error of 1 sample period i.e.125μsec; the error for small azimuth angles i.e. -40^o to $+40^o$ being less than 40μsec.

In order to study the performance of the algorithm for a slowly moving source, we simulated a speech source being initially positioned directly in front of the listener which is then moved gradually to the left of the user and then to the back, over a period of about 10 seconds. In Figure 3, we plot the actual TDOAs for the simulation along with the TDOA estimated successively throughout the speech utterance. As we can see, the algorithm is able to track the source TDOA reasonably well except for the tracking delay. This delay is the result of using long-term averaging in order to make a better estimate of the TDOA .

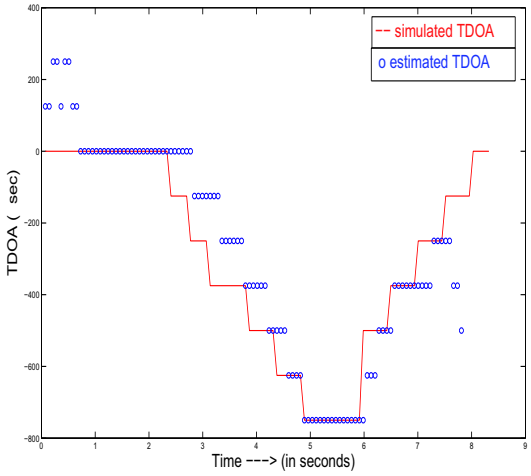

Figure 3: *Moving source tracking in terms of interaural time delay: simulated and estimated*

5. CONCLUSION

The proposed CCI***B***WF (codebook constrained iterative binaural Wiener filter) algorithm helps to improve the speech quality in terms of both noise reduction as well as speech intelligibility. We are able to achieve the optimum performance within just 2 iterations of the CCI***B***WF. We show that the algorithm does not introduce much distortion in the interaural time delay cues, thereby preserving the localization information in the source. The algorithm is able to track the TDOA in cases where the source may be moving slowly, thereby ensuring that the performance does not degrade due to selection of wrong codebooks.

6. References

[1] H. Huang and Kyriakakis, C., "Binaural noise reduction combining binaural analysis and psychoacoustically motivated spectral subtraction," *Conf. Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Comp.* , Vol.2, no., pp. 2260-2264 Vol.2, 7-10 Nov. 2004

[2] H. Huang and Kyriakakis, C., "Binaural Model Based Adaptive Binaural Noise Reduction," *Signals, Systems and Computers*, 2006. ACSSC '06. Fortieth Asilomar Conference on , vol., no., pp.1110-1113, Oct. 29 2006-Nov. 1 2006

[3] Klasen, T.J., Doclo, S., Van den Bogaert, T., Moonen, M. and Wouters, J., "Binaural Multi-Channel Wiener Filtering for Hearing Aids: Preserving Interaural Time and Level Differences," *Inter. Conf. on Acous., Speech and Signal Processing*, Vol.5, no., pp.V-V, 14-19 May 2006

[4] J.S. Lim and A. Oppenheim, "All-pole Modeling of Degraded Speech", *IEEE Trans on Acoust, Speech and Signal Proc* , vol ASSP-6, no 3,pp 197-220, June 1978.

[5] J.H.L Hansen and M.A.Clements,"Constrained Iterative Speech Enhancement with application to Speech Recognition",*IEEE Trans on Acoust, Speech and Signal Proc*, vol 39, no. 4, pp 795-805, Apr 1991.

[6] Sreenivas, T.V. and Kirnapure, P., "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech and Audio Proc.* , vol.4, no.5, pp.383-389, Sep 1996.

[7] B. Gardner and K. Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone", Technical Report #280, MIT Media Lab, 1994.