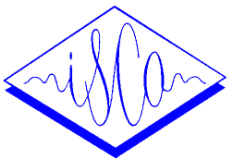


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

STATE BASED SUB-BAND WIENER FILTERS FOR SPEECH ENHANCEMENT IN CAR ENVIRONMENTS

Aimin. Chen Saeed. Vaseghi

Brunel University, England.
E-mails : (aimin.chen, saeed.vaseghi)@brunel.ac.uk

ABSTRACT

The performance of Wiener filters in restoring the quality and intelligibility of noisy speech depends on: (i) the accuracy of the estimates of the power spectra or the correlation values of the noise and the speech processes, and (ii) on the Wiener filter structure. In this paper a Bayesian method is proposed where model combination and model decomposition are employed for the estimation of parameters required to implement subband Wiener filters. The use of subband Wiener filters provides advantages in terms of improved parameter estimates and also in restoring the temporal-spectral composition of speech. The method is evaluated, and compared with the parallel model combination, using the TIMIT continuous speech database with BMW and VOLVO car noise databases.

1. INTRODUCTION

Speech communication from a moving car or train over a mobile phone can be severely degraded by the ambient noise. With the increasing deployment of speech recognition and voice-based systems across a wide range of voice-based services, it is important for the users and providers of mobile phones that access to voice recognition systems is not impaired by the noise. Noise degrades the accuracy of automatic speech recognition even for such modest tasks as voice dialling, automatic directory enquiry, or voice control of the accessories in a moving car. The ambient noise in a vehicle is a time-varying process and may be from a number of sources such as; noise from the engine and the revolving mechanical parts of the car, the vibration noise from the surface contact of the wheels and the roads, the noise from air flow into the car through the ducts or open windows, noise from passing/overtaking vehicles, clicks from the left/right indicators etc.

In this paper we consider a sub-band implementation of Wiener filters based on the linear prediction models of speech and noise signals, and hidden Markov models of the time-varying statistics of the speech and noise LP features. The performance of Wiener filters in restoring the quality and intelligibility of speech depends mainly on two factors; the ability of the method to accurately estimate the time-varying correlation (or power spectrum) coefficients of the signal and the noise processes, and the Wiener filter structure. The structure of the Wiener filter (full-band vs sub-band, time vs frequency implementation) can also have a significant bearing on the accuracy of the estimates of the correlation coefficients and on the ability of the filter to restore the temporal-spectral characteristics of speech.

The sub-band linear prediction models considered in this work have the following advantages:

(1) The signal and noise within each band are modelled by relatively low order LP models requiring fewer predictor coefficients and correlation values than a full-band LP model, which may be better estimated.

(2) The prediction order in different bands can be varied according to the expected correlation structure of the signal and the noise process in the sub-bands.

(3) Te effects of noise and distortion in each sub-band are confined to the speech features extracted from that band and do not effect speech features in other bands.

The remainder of the paper is organised as follows. In section 2 the system outline is described. Section 3 presents a sub-band LP model for noise reduction. Section 4 describes implementation of sub-band linear prediction Wiener filters and in section 5 evaluation results on BMW and VOLVO car noise are presented.

2. SYSTEM OUTLINE

Fig. (1) illustrates the outline of the proposed HMM-based sub-band noisy speech recognition/enhancement system. The system performs the following functions:

1- Combination of speech and noise HMMs to form noisy speech models.
2- Estimation of the best combined noisy speech model given the current input and HMMs of speech and noise.
3- State decomposition; separation of speech and noise probability density functions.
4- State-based Wiener filtering.

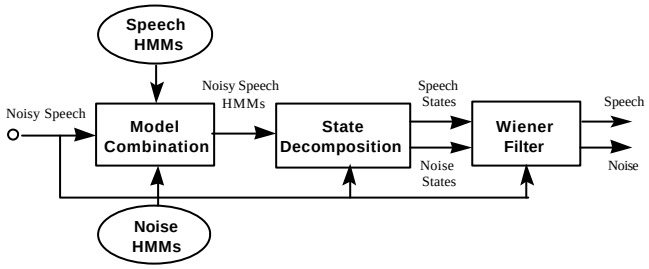
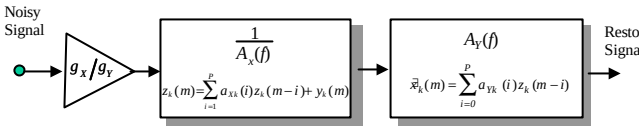


Figure 1 - Outline configuration of HMM-based noisy speech recognition/enhancement.

Clean speech HMMs and noise HMMs are combined to form noisy speech HMMs for the current noise environment [1]. Model combination also requires an estimate of the current signal to noise ratio. The noisy speech decoder estimates the most likely speech and noise models and states and then derives a set of Wiener filters for speech enhancement.

Assuming that the recognition system correctly labels the noisy speech and the noise states, the quality of a noisy speech enhancement system depends on the accuracy of the linear prediction model of speech. The subband linear prediction modelling provides a number of advantages over the full band method including; (a) the signal within each band can be well modelled by a smaller number of correlation lags, (b) the Wiener filtering in each band can be achieved by filtering through a cascade of a linear prediction model of the noisy speech and an inverse prediction model of the clean speech. The perceptual quality of the restoration achieved through HMM-based subband linear prediction outperforms the full band system.

3. SUBBAND LINEAR PREDICTION MODELS



In sub-band linear prediction the signal $x(m)$ is passed through a bank of N band pass filters, and split into N subband signals $x_k(m)$ k=1, …, N. The kth subband signal is modelled using a low-order linear prediction model

$$x_k(m) = \sum_{i=1}^{P_k} a_k(i)\, x_k(m-i) + g_k e_k(m) \qquad (1)$$

where [a_k, g_k] are the coefficients and the gain of the predictor model for the k^{th} subband. The choice of the model order P_k depends on the signal corelation structure within each subband and the sub-band width

Each subband signal is demodulated and then down-sampled by α.

$$\alpha = \frac{whole\ band}{Subband\ bandwidth} \qquad (2)$$

After LP Wiener filtering the signals are re-sampled by 1/α and then transformed back to original band.

4. IMPLEMENTATION OF SUBBAND LINEAR PREDICTION WIENER FILTERS

Assuming that the noise is additive, the noisy signal in each subband is modelled as

$$y_k(m) = x_k(m) + n_k(m) \qquad (3)$$

The Wiener filter in the frequency domain can be expressed in terms of the power spectra, or in terms of LP model frequency responses, of the signal and noise process as

$$\begin{aligned} W_k(f) &= \frac{P_{X,k}(f)}{P_{Y,k}(f)} \\ &= \frac{g_{X,k}^2}{\left|A_{X,k}(f)\right|^2} \frac{\left|A_{Y,k}(f)\right|^2}{g_{Y,k}^2} \end{aligned} \qquad (4)$$

Where $P_{X,k}(f)$ and $P_{Y,k}(f)$ are the power spectra of the clean signal and the noisy signal respectively.
The linear prediction Wiener filter can be implemented in the time domain with a cascade of a linear predictor of the clean signal, followed by an inverse predictor filter of the noisy signal as expressed by the following relations

Figure 2 - A cascade implementation of LP Wiener filter.

$$z_k(m) = \sum_{i=1}^{P} a_{Xk}(i) z_k(m-i) + \frac{g_X}{g_Y} y_k(m) \quad (5)$$

$$\hat{x}_k(m) = \sum_{i=0}^{P} a_{Yk}(i) z_k(m-i) \quad (6)$$

5. EXPERIMENTS AND EVALUATIONS

The evaluations are based on the TIMIT continuous speech database and two noise databases of BMW and VOLVO car noise. The speech was transcribed using a set of 39 phonemes. The statistics of the spectral and temporal structure of each phone where modelled with a 3-state left-right HMM, and the distributions of speech feature vectors within each state were modelled with a multivariate Gaussian mixture distribution.
Speech was sampled at a rate of 8 kHz and then split into a number of subband signals using a bank of linear phase FIR filters. The subband speech signals were then segmented into frames of 25 ms with a frame rate of 10 ms (segment overlap of 15 ms).

5.1 Car Noise

Car noise as shown in figures 3 and 4 is a predominantly low frequency noise with most of the energy below a frequency of 1 kHz. Although as the examples in figures 3 and 4 demonstrate there are significant variations in the noise spectra of different cars. In a car with the windows shut most of the noise emanates from the engine, the revolving mechanical parts of the car, the air flow from the air ducts, winds, road surface noise, windscreen viper, indicators, and noise from passing vehicles. The noise from engine and mechanical parts of the car is a relatively slowly varying process where as the noise from passing cars is usually very nonstationary and transient.

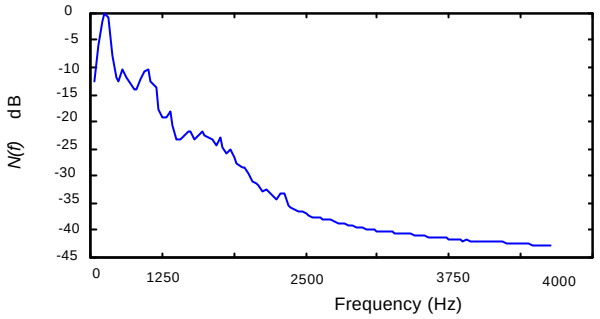


Figure 3 - Power spectral of noise in a BMW at 70 mph.

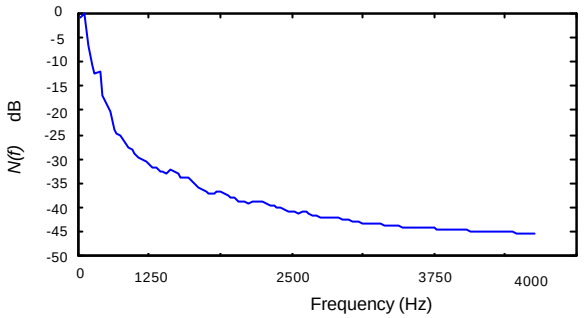


Figure 4 - Power spectral of noise in a Volvo at 70 mph.

5.2 Car Noise Model

In noisy speech processing often well-trained models of speech are available, however the availability of adequate and well-trained noise models can be an obstacle to noisy speech processing. In a car environment over time there will be plenty of noise data and conditions that can be use to train noise models. The difficulty is that car noise has a large variety of states and unlike speech does not have a well-constrained temporal structure of the kind that could be modelled say by a 3-state left right HMM. The question of noise model accuracy has a bearing not only on the accuracy of the combined noisy speech model but also on the computational practicality of the system.
An adequate model for noise may be a Gaussian mixture model (GMM), equivalent to a form of single state HMM, with a large number of mixtures.

5.3 Recognition with enhanced speech

Recognition experiments involved BMW and VOLVO car noise. The MFCCs features are extracted from the frames of 25ms at a frame rate of 10ms in log mel-filter bank. The sub-band MFCCs are augmented by delta and delta-delta features. The clean speech models are used to test enhanced speech in mismatch training conditions. These

compared the performance of subband Wiener filters based on the linear prediction (LP) models of speech and noise using PMC method and noisy speech in mismatch training conditions. The results from these experiments are given in tables 1-6 for two types of car noise at different SNRs. The noise used for PMC adaptation is from a simple single mixture averaging model.

Results in table 1- 6 show that subband Wiener filters based on the linear prediction (LP) models give good performance for speech enhancement. For BMW noise the maximum accuracy of 44.79%, 38.86% and 35.07% under training mismatch condition improves on the 35.63%, 29.75% and 21.02% achieved by sub-band PMC models in SNRs equivalent to 10dB, 5dB and 0dB respectively.

No. Mixture	Enhanced Speech	Sub-band PMC	Noisy
1 mix	53.32/38.39	47.17/25.19	35.55/26.54
5 mixs	58.29/41.23	55.89/31.16	38.86/28.20
12 mixs	57.11/40.05	58.80/34.98	36.97/26.07
18 mixs	60.19/44.31	59.53/35.45	36.73/26.07
20 mixs	60.19/44.79	59.78/35.63	37.68/27.01

Table 1: Effect of enhancement comparing with sub-band PMC in BMW SNR=10dB

No. Mixture	Enhanced Speech	Sub-band PMC	Noisy
1 mix	52.37/35.78	45.39/22.73	25.83/19.43
5 mixs	53.08/33.89	53.56/27.82	27.73/18.48
12 mixs	53.79/33.41	55.69/29.86	24.64/17.06
18 mixs	55.69/38.63	56.72/29.31	25.83/17.54
20 mixs	55.45/38.86	56.94/29.75	27.73/19.43

Table 2: Effect of enhancement comparing with sub-band PMC in BMW SNR=5dB

No. Mixture	Enhanced Speech	Sub-band PMC	Noisy
1 mix	49.29/29.15	42.58/18.33	19.91/14.45
5 mixs	53.55/33.89	49.08/20.92	17.30/13.98
12 mixs	52.13/34.36	50.70/21.51	18.96/12.09
18 mixs	52.13/34.60	51.95/20.86	18.72/14.69
20 mixs	52.37/35.07	52.18/21.02	19.67/15.17

Table 3: Effect of enhancement comparing with sub-band PMC in BMW SNR=0dB

No. Mixture	Enhanced Speech	Sub-band PMC	Noisy
1 mix	44.55/25.85	55.92/37.91	45.57/31.48
5 mixs	56.30/33.06	57.35/41.47	52.40/35.20
12 mixs	59.76/37.44	60.90/42.89	54.05/35.99
18 mixs	60.46/38.61	65.17/50.47	55.23/37.56
20 mixs	61.29/39.02	67.54/50.47	55.35/35.94

Table 4: Effect of enhancement comparing with sub-band PMC in VOLVO SNR=10dB

No. Mixture	Enhanced Speech	Sub-band PMC	Noisy
1 mix	44.81/25.70	55.92/36.26	41.35/28.72
5 mixs	55.63/32.57	56.87/39.34	47.91/31.65
12 mixs	58.85/35.63	57.58/37.68	47.79/29.98
18 mixs	59.87/36.63	60.19/42.89	47.92/29.07
20 mixs	60.44/37.00	61.37/44.55	48.17/28.85

Table 5: Effect of enhancement comparing with sub-band PMC in VOLVO SNR=5dB

No. Mixture	Enhanced Speech	Sub-band PMC	Noisy
1 mix	45.22/25.34	54.27/37.44	34.36/24.17
5 mixs	54.33/29.78	54.98/35.78	40.13/25.64
12 mixs	56.84/32.13	55.92/37.20	37.75/22.25
18 mixs	58.29/31.61	55.45/37.20	36.06/17.20
20 mixs	58.82/32.41	59.00/37.91	36.05/17.85

Table 6: Effect of enhancement comparing with sub-band PMC in VOLVO SNR=0dB

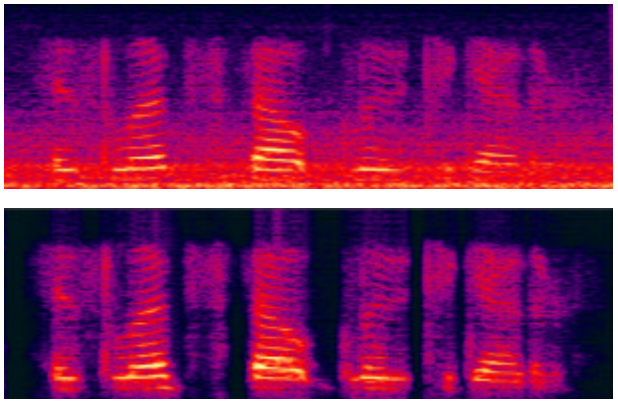


Figure 5 – Spectra of noisy and enhanced speech signal in a BMW at 70 mph SNR=5dB

6. CONCLUSIONS

The performance of state sub-band LP Wiener filters in restoring the quality of noisy speech rely on the accuracy of the estimates of the power spectra or the correlation values of the noise and the speech, and on the Wiener filter structure. Results show that the method of state sub-band LP Wiener filters is effective, and the performance is impacted in terms of the spectral characteristics of the car noise.

References

[1] M. Gales and S. Young, "HMM recognition in noise using parallel model combination", EUROSPEECH-93, pages 837-840, 1993.

[2] A. Varga and R. Moore, "Hidden Markov model decomposition of speech and noise", ICASSP-90, pages 845-848, 1990.

[3] J. Lim and A. Oppenheim, "All-pole modelling of degraded speech", IEEE Trans. ASSP, Vol. 26, No. 3, pages 197-210, 1978.

[4] Young, P. Woodland (1996), HTK-Hidden Markov Model Toolkit, Entropic, Cambridge

[5] Y. Ephraim (1992), "Statistical-model-based speech enhancement systems", Proc. IEEE, Vol. 80, Pages 1555-1562.

[6] A. Y. Minami, S. Furui, (1995),"A maximum likelihood procedure for a universal adaptation method based on HMM recognition", ICASSP-95, Vol. 1, pages 129-133.

[7] S. J. Godsill and P.J.W. Rayner. ,"A Bayesian approach to the restoration of degraded audio signals", IEEE Trans. on Speech and Audio Processing, 3(4): 267-278, July1995.

[8] H. Hermansky *et al* (1992), "RASTA-PLP speech analysis technique", ICASSP-92, pages 121-124.