NOISE ESTIMATION FOR ENHANCEMENT AND RECOGNITION WITHIN AN AUTOREGRESSIVE HIDDEN-MARKOV-MODEL FRAMEWORK

B. T. Logan and A. J. Robinson

Speech, Vision and Robotics Group,
Department of Engineering,
University of Cambridge

ABSTRACT - This paper describes a new algorithm to enhance and recognise noisy speech when only the noisy signal is available. The system uses autoregressive hidden Markov models (HMMs) to model the clean speech and noise and combines these to form a model for the noisy speech. The combined model is used to determine the likelihood of each observation being just noise. These likelihoods are used to weight each observation to form a new estimate of the noise and the process is repeated. Enhancement is performed using Wiener filters formed from the clean speech and noise models. Results are presented for additive stationary Gaussian and coloured noise.

INTRODUCTION

Speech enhancement systems attempt to improve the perceptual aspects (e.g. quality, intelligibility) of noisy speech. Many researchers have looked at this task (Ephraim October 1992, Lim 1979). Much of this work requires estimates of the statistics of the clean speech and the interfering noise. While training databases are available to make models of clean speech, the noise may only be available as part of the noisy signal. A popular technique is to estimate the noise from periods of non-speech, reducing the problem to one of determining whether a given frame is speech or noise.

This work models the noisy speech using an Autoregressive Hidden Markov Model (Juang 1984) framework. A combined speech and noise model is built and used to recognise the noisy speech. This model is used to estimate the likelihood of each observation vector being noise. These likelihoods are used to weight the observation vectors to form a new estimate of the noise.

Autoregressive HMMs are used because they segment the speech into clusters of signals with similar autocorrelation parameters which are used to form Wiener filters to enhance the speech. Also, speech recognition in unknown noise is possible with this approach. Additionally, the technique is potentially extendible to non-stationary noise.

This paper describes the theory of the enhancement system and details the results of experiments conducted on speech degraded by additive, stationary Gaussian and coloured noise. These show that the algorithm can effectively enhance the speech and improve the recognition in noise.

THE ENHANCEMENT SYSTEM

The enhancement system is based on a system by Ephraim (April 1992) but modified to use noise estimates from the noisy speech. The basic algorithm is given below.

set noise statistics to zero
loop
  perform recognition using a combination of clean speech models and noise statistics
  estimate noise
end
form Wiener filters using models given by recognition and estimated noise
enhance speech

There are three main components to the system: noise estimation, recognition in noise and enhancement. These are described in the following sections.

Recognition in Noise

The clean speech is modelled by HMMs with pdfs given by Equation 1.

$$p(\mathbf{y}_0^T) = \sum_{x_0^T} \prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau} b_{x_\tau}(\mathbf{y}_\tau) \tag{1}$$

Here $\mathbf{y}_0^T$ is a sequence of clean observations $\{\mathbf{y}_\tau, \tau = 0, \ldots, T\}$, x_0^T is a sequence of states $\{x_\tau, \tau = 0, \ldots, T\}$, $a_{x_{\tau-1}x_\tau}$ is the transition probability from state $x_{\tau-1}$ to state x_τ and $b_{x_\tau}(\mathbf{y}_\tau)$ is the pdf of the output vector $\mathbf{y}_\tau$ from the state x_τ.

The pdf $b_{x_\tau}(\mathbf{y}_\tau)$ is assumed Gaussian with zero mean and covariance matrix S_{x_τ}. If the further assumption is made that the observations are from an autoregressive process of order P, then this covariance matrix is dependent on only $P+1$ parameters (Juang 1984). This assumption leads to a mathematically tractable system.

The noise process is also modelled by an autoregressive HMM. Its pdf is given by Equation 2.

$$p(\mathbf{v}_0^T) = \sum_{\tilde{x}_0^T} \prod_{\tau=0}^{T} a_{\tilde{x}_{\tau-1}\tilde{x}_\tau} b_{\tilde{x}_\tau}(\mathbf{v}_\tau) \tag{2}$$

Here $\mathbf{v}_0^T$ is a sequence of noise observations $\{\mathbf{v}_\tau, \tau = 0, \ldots, T\}$ and $\tilde{x}_0^T$ is a sequence of noise states $\{\tilde{x}_\tau, \tau = 0, \ldots, T\}$. Once again again, $b(\cdot)$ is the pdf of a Gaussian autoregressive process.

For additive noise, these models can be combined to yield a model for the noisy signal. The pdf for this model is given by Equation 3.

$$p(\mathbf{z}_0^T) = \sum_{\bar{x}_0^T} \prod_{\tau=0}^{T} a_{\bar{x}_{\tau-1}\bar{x}_\tau} b_{\bar{x}_\tau}(\mathbf{z}_\tau) \tag{3}$$

Here $\mathbf{z}_0^T$ is a sequence of noisy observations $\{\mathbf{z}_\tau, \tau = 0, \ldots, T\}$, $\bar{x}_0^T$ is a sequence of composite noisy states $\{(x_\tau, \tilde{x}_\tau), \tau = 0, \ldots, T\}$, $a_{\bar{x}_{\tau-1}\bar{x}_\tau}$ is the transition probability from state $\bar{x}_{\tau-1}$ to state $\bar{x}_\tau$ and $b_{\bar{x}_\tau}(\mathbf{z}_\tau)$ is the pdf of the output vector $\mathbf{z}_\tau$ from the state $\bar{x}_\tau$.

For additive noise, the following equations hold.

$$\begin{aligned}
\mathbf{z}_0^T &= \mathbf{y}_0^T + \mathbf{v}_0^T && (4)\\
a_{\bar{x}_{\tau-1}\bar{x}_\tau} &= a_{x_{\tau-1}x_\tau} a_{\tilde{x}_{\tau-1}\tilde{x}_\tau} && (5)\\
b_{\bar{x}_\tau}(\mathbf{z}_\tau) &= \int b(\mathbf{z}_\tau - \mathbf{y}_\tau) b(\mathbf{y}_\tau) d\mathbf{y}_\tau && (6)
\end{aligned}$$

The pdf $b_{\bar{x}_\tau}(\mathbf{z}_\tau)$ is Gaussian with zero mean and covariance matrix $S_{\bar{x}}$ given by:

$$S_{\bar{x}} = g^2 S_x + S_{\tilde{x}} \tag{7}$$

Here g^2 is a gain to take into account the mismatch between training data (for the clean speech models) and testing data. The calculation of g^2 and a mathematically tractable technique to calculate the determinant and inverse of $S_{\bar{x}}$ are described by Ephraim (June 1992).

Noise Estimation

An estimate of the noise is required for both recognition of the noisy speech and enhancement. The required statistics are the noise covariance matrix $S_{\tilde{x}}$ used for recognition and an estimate of the noise spectrum used for enhancement. Both of these parameters can be easily estimated from the noise autocorrelation function.

The noise autocorrelation function is estimated from a weighted sum of the autocorrelation functions of all of the observation vectors. Each weight reflects the likelihood of the observation being noise. The likelihoods are determined using the autoregressive HMM framework as follows.

Using the noisy model built from the current noise estimate and the clean speech models, recognition is performed yielding the composite state-frame alignment. For the stationary noise considered here, the likelihood of interest is that of the composite state $\bar{x}_\tau$ being $(x_\tau \equiv \text{silence}, \tilde{x}_\tau)$. That is, the likelihood of there being no speech present.

In this work, a simple formula was used to approximate this likelihood where transition probabilities between states have been ignored and only the observation at time τ is considered. Specifically, the the weight $w(\tau)$ used for the observation at time τ is given by:

$$w(\tau) = \frac{b_{\bar{x}_\tau^{\mathsf{s}}}(\mathbf{z}_\tau)}{\sum_{t=0}^{T} b_{\bar{x}_t^{\mathsf{s}}}(\mathbf{z}_t)} \tag{8}$$

where $\bar{x}_\tau^{\mathsf{s}} = (x_\tau \equiv \text{silence}, \tilde{x}_\tau)$.

Wiener Filtering

Once the most likely state alignment has been obtained using the noisy model for recognition, non-causal frequency domain Wiener filters are formed. The frequency response of the Wiener filter at time τ is:

$$W_\tau(\theta) = \frac{g_\tau^2 f_{x_\tau}(\theta)}{g_\tau^2 f_{x_\tau}(\theta) + f_{\tilde{x}_\tau}(\theta)} \tag{9}$$

Here $f_{x_\tau}(\theta)$ and $f_{\tilde{x_\tau}}(\theta)$ are the power spectral densities of the clean speech state and noise state and g_τ^2 is the gain term described in the recognition section. The power spectral densities can be determined from the autoregressive parameters for these states using the following:

$$f_{x_\tau}(\theta) = \frac{\sigma_{x_\tau}^2}{\mid A_{x_\tau}(\theta)\mid^2} \tag{10}$$

$$f_{\tilde{x}_\tau}(\theta) = \frac{\sigma_{\tilde{x}_\tau}^2}{\mid A_{\tilde{x}_\tau}(\theta)\mid^2} \tag{11}$$

Here $A_{x_\tau}(\theta)$ and $A_{\tilde{x}_\tau}(\theta)$ are the Fourier transforms of the autoregressive coefficients for the states and x_τ and $\tilde{x}_\tau$ respectively and σ_{x_τ} and $\sigma_{\tilde{x}_\tau}$ are the gains of these autoregressive models.

This technique assumes that one state sequence dominates the pdf in (3). For the experiments conducted, little to no improvement in enhancement was observed by relaxing this assumption and forming the weighted sum of Wiener filters.

Noise Source	Clean Models (% correct)	Noisy Models (% correct)
None (Clean)	100.0	-
10dB Gaussian	25.0	100.0
6dB Gaussian	12.5	95.0
6dB Coloured[1]	22.5	82.5

Table 1: Recognition in Noise

RESULTS

The basic noise estimation algorithm was tested on a simple single speaker isolated digit recognition task. The data was taken from the Noisex database (Varga 1992). The speech is sampled at 16kHz and observation vectors are formed by applying a Hamming window to 32ms frames at a frame rate of 16ms. The order of the autoregressive models was 20. One 8-state HMM was trained for each of the ten digits and a 1-state HMM was trained for the separating silence.

In these experiments, the effects of adding Gaussian noise and coloured noise were studied. Only stationary additive noise was considered. The clean models were trained on 100 utterances of digits (10 of each digit). The digits were grouped into files containing 20 each for training and testing purposes.

Tests were conducted on 40 different utterances (4 of each digit). Results for recognition in noise using the clean models and the the noisy models determined by the algorithm are given in Table 1. It is seen that the noise estimation is sufficiently good to improve recognition in noise for this task. Four iterations of the noise estimation algorithm were used.

Figure 1 shows the real and estimated (power) spectrum of the 6dB coloured noise on successive iterations of the algorithm. It is seen that although no formal convergence has been proved, the estimated spectrum does approach the true spectrum in this case.

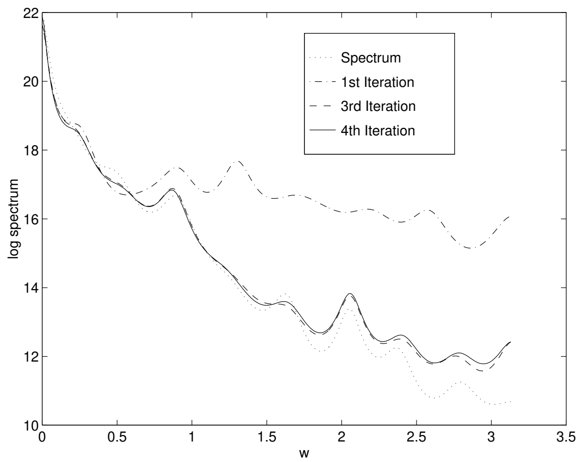

Figure 1: Estimated and Real Power Spectrum of 6dB Coloured Noise

The quality of the enhanced speech was quite high, particularly for the Gaussian noise sources. For these utterances, the main distortions were some residual noise due to gain estimation errors during some consonants. This was more pronounced for the coloured-noise speech. Figures 2, 3 and 4 show the clean, noisy and enhanced spectrums for the first four digits of the speech distorted by 10dB

Gaussian noise.

CONCLUSIONS

A new algorithm that performs enhancement and recognition and operates when only the noisy signal is available has been presented. It iteratively uses autoregressive HMMs to model the clean speech and combines these with noise estimates to form a combined model. This is used to recognise the speech and estimate the likelihood of each observation being noise. A new noise estimate is determined using these likelihoods to weight each observation. The enhancement is performed by Wiener filters formed from the speech and noise estimates. Results presented for additive stationary Gaussian and coloured noise show the algorithm to be effective. The algorithm is potentially extendible to non-stationary noise.

ACKNOWLEDGEMENTS

B. T. Logan gratefully acknowledges funding from the Cambridge Commonwealth Trust.

NOTES

(1) Noise Type 12 ('Lynx') in the NOISEX-92 database

REFERENCES

Ephraim, Y. (April 1992) *A Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models*, IEEE Transactions on Signal Processing 40, 725–735.

Ephraim, Y. (June 1992) *Gain-Adapted Hidden Markov Models for Recognition of Clean and Noisy Speech*, IEEE Transactions on Signal Processing 40, 1303–1316.

Ephraim, Y. (October 1992) *Statistical–Model–Based Speech Enhancement Systems*, Proceedings of the IEEE 80, 1562–1555.

Juang, B. (1984) *On the Hidden Markov Model and Dynamic Time Warping for Speech Recognition - A Unified View*, AT&T Bell Laboratories Technical Journal 63, 1213–1243.

Lim, J & Oppenheim, A (1979) *Enhancement and Bandwidth Compression of Noisy Speech*, Proceedings of the IEEE 67, 1586–1604.

Varga, A.P., Steeneken, H.J.M, Tomlinson, M. & Jones, D. (1992) *The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition*, Tech. Report, DRA Speech Research Unit.

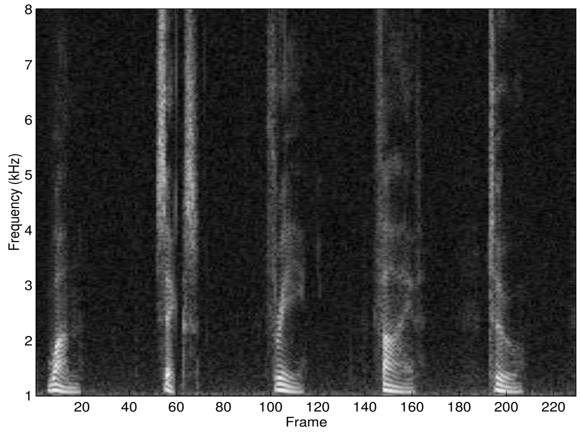

Figure 2: Clean Speech ("1 6 3 5 2")

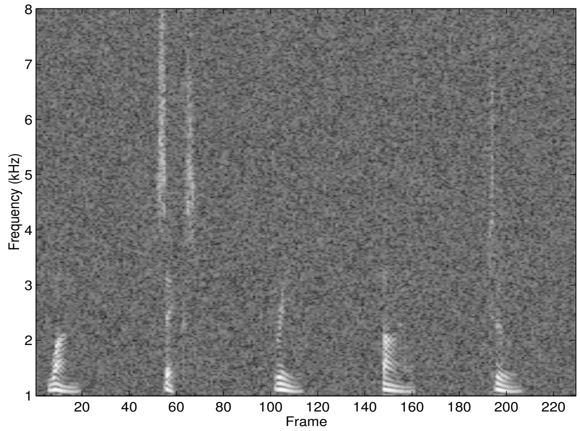

Figure 3: Noisy Speech ("1 6 3 5 2") with 6dB Gaussian Noise

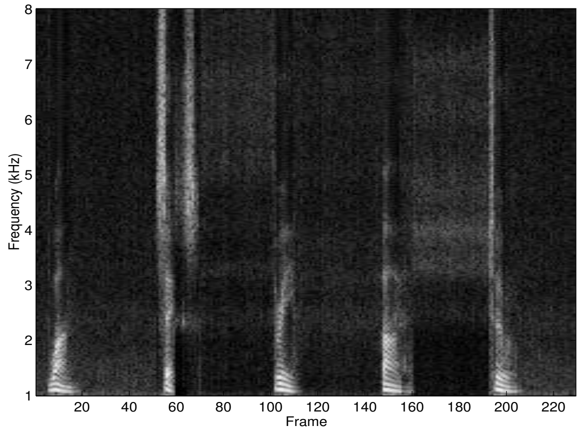

Figure 4: Enhanced Speech ("1 6 3 5 2") from 6dB Gaussian Noisy Speech