ENHANCEMENT AND RECOGNITION OF NOISY SPEECH WITHIN AN AUTOREGRESSIVE HIDDEN MARKOV MODEL FRAMEWORK USING NOISE ESTIMATES FROM THE NOISY SIGNAL

B. T. Logan *A. J. Robinson*

Cambridge University Engineering Department, Cambridge, United Kingdom

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 probability framework developed is then used to reestimate the noise models from the corrupted speech waveform and the process is repeated. Enhancement is performed using the Wiener filters formed from the final clean speech models and noise estimates. Results are presented for additive stationary Gaussian and coloured noise.

1. INTRODUCTION

The task of speech enhancement has been investigated by many researchers [1, 2, 3, 4]. 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. Recently, researchers have considered estimating the noise directly from this corrupted signal [2]. Their technique uses hidden filter HMMs [3] to model the clean speech and chooses the noise parameters to give the best possible estimate of the clean signal.

This paper considers estimating the clean speech and noise within an autoregressive HMM framework [5]. Autoregressive HMMs are used to model the speech and noise and a combined model is built and used to recognise the noisy speech. A new noise model is generated by summing the expected value of the noise statistics given each observation and each HMM state, weighted by the likelihood of being in each state. The process is repeated until the total likelihood converges to a maximum.

Autoregressive HMMs are used because they segment the speech into clusters of signals with similar autocorrelation parameters. These are used to form Wiener filters to enhance the speech. A further benefit of this approach is that it provides speech recognition in unknown noise. 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.

Additionally, the quality of the autoregressive parameters determined by the algorithm is investigated by comparing the Itakura distortion measure [6] of the system to that obtained from the iterative Wiener filter system formulated by Lim and Oppenheim [7]. It is seen that the technique of using trained clean speech models yields autoregressive parameters that are better on average in the Itakura sense than those that are estimated from the noisy speech alone as in [7].

2. THE ENHANCEMENT SYSTEM

The enhancement system described in this paper is a version of a system by Ephraim [8] modified to use noise estimates from the noisy speech. The basic algorithm is shown in Figure 1.

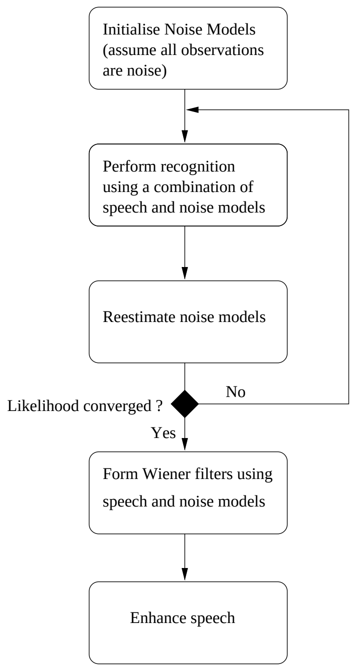

Figure 1. Enhancement Algorithm

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

2.1. Recognition in Noise

The recognition system is similar to that described by Ephraim [8]. It models the clean speech observations y_1^T and noise observations v_1^T by HMMs. These observations are windowed speech samples. For additive noise, the noisy speech is also modelled by an HMM with the pdf given by:

$$p(z_1^T|\lambda) = \sum_{\bar{x}_1^T} \prod_{t=1}^{T} a_{\bar{x}_{t-1}\bar{x}_t} b_{\bar{x}_t}(z_t) \tag{1}$$

Where:

z_1^T = a sequence of noisy observations $\{\mathbf{z}_t, t = 1, \ldots, T\}$

$\bar{x}_1^T$ = a sequence of noisy states $\{\bar{x}_t, t = 1, \ldots, T\}$

$a_{\bar{x}_{t-1}\bar{x}_t}$ = transition probability from state $\bar{x}_{t-1}$ to state $\bar{x}_t$

$b_{\bar{x}_t}(\mathbf{z}_t)$ = pdf of the output vector $\mathbf{z}_t$ from the state $\bar{x}_t$

λ = the model parameters

For the additive noise case, the following equations hold.

$$\mathbf{z}_1^T = \mathbf{y}_1^T + \mathbf{v}_1^T \tag{2}$$

$$a_{\bar{x}_{t-1}\bar{x}_t} = a_{x_{t-1}x_t} a_{\tilde{x}_{t-1}\tilde{x}_t} \tag{3}$$

$$b_{\bar{x}_t}(\mathbf{z}_t) = \int b_{\tilde{x}_t}(\mathbf{z}_t - \mathbf{y}_t) b_{x_t}(\mathbf{y}_t) d\mathbf{y}_t \tag{4}$$

Here, at each time t, the state of the noisy process $\bar{x}_t$ is a combination of the clean state x_t and the noise state $\tilde{x}_t$.

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

$$S_{\bar{x}_t} = g_t^2 S_{x_t} + S_{\tilde{x}_t} \tag{5}$$

Here S_{x_t} and $S_{\tilde{x}_t}$ are the covariance matrices of b_{x_t} and $b_{\tilde{x}_t}$ respectively and g_t^2 is a gain term to take into account the mismatch between training data and testing data for the clean speech models. The calculation of g_t^2 and a mathematically tractable technique to calculate the determinant and inverse of $S_{\bar{x}_t}$ are described by Ephraim [4]. For the experiments described here, the gain was set to one since the training and testing conditions were near-identical.

2.2. Noise Estimation

The noise model parameters are chosen to maximise the likelihood of the noisy model given the observations. The technique used is similar to that of Rose et. al. [9] in which parameters are reestimated from noisy data. In [9] however, speech model parameters were reestimated whereas this paper is concerned with reeestimating the noise model parameters. Also, the models in this case are autoregressive HMMs rather than the Gaussian mixtures used in [9].

The noise parameter reestimation formulas are derived as follows. Consider the model described by Equation 1. The model parameters λ are: $\{a_{x_{t-1}x_t} \forall x_{t-1}x_t\}$, $\{a_{\tilde{x}_{t-1}\tilde{x}_t} \forall \tilde{x}_{t-1}\tilde{x}_t\}$, $\{g_t^2, t = 1, ..., T\}$, $\{S_x \forall x\}$ and $\{S_{\tilde{x}} \forall \tilde{x}\}$. It is required to find a new estimate of λ, λ', which maximises $p(z_1^T|\lambda)$. This can be achieved after Baum et. al. [10] by defining an auxiliary function

$$Q(\lambda, \lambda') = E\{\log(p(z_1^T|\lambda'))\} \tag{6}$$

and maximising $Q(\cdot)$ with respect to λ'. To reestimate the noise parameters, it is only necessary to maximise $Q(\cdot)$ with respect to $\{S_{\tilde{x}} \forall \tilde{x}\}$ and $\{a_{\tilde{x}_{t-1}\tilde{x}_t} \forall \tilde{x}_{t-1}\tilde{x}_t\}$.

Consider first the maximisation of $Q(\cdot)$ with respect to $\{a_{\tilde{x}_{t-1}\tilde{x}_t} \forall \tilde{x}_{t-1}\tilde{x}_t\}$. Applying the method of [10] yields the following equation for a new estimate of $a_{\tilde{x}_{t-1}\tilde{x}_t}$.

$$a'_{\tilde{x}_{t-1}\tilde{x}_t} = \frac{\sum_{\tau=1}^{T} P(\tilde{x}_{\tau-1} = \tilde{x}_{t-1}, \tilde{x}_\tau = \tilde{x}_t, z_1^T|\lambda)}{\sum_{\tau=1}^{T} P(\tilde{x}_\tau = \tilde{x}_t, z_1^T|\lambda)} \tag{7}$$

Now consider the reestimation of $\{S_{\tilde{x}} \forall \tilde{x}\}$. Because the noise is assumed to come from an autoregressive process, each $S_{\tilde{x}}$ can be calculated from the autocorrelation vector of its noise. This is therefore the required statistic and is denoted by $\mathbf{r}_{\tilde{x}}$. Following similar reasoning to [9], it can be reestimated using the following equation for each noise state $\tilde{x}$.

$$\mathbf{r}'_{\tilde{x}} = \frac{\sum_{t=1}^{T} \sum_{\forall x} P(x_t = x, \tilde{x}_t = \tilde{x}|z_t, \lambda) E\{\mathbf{r}_{\tilde{x}}|z_t, x_t = x, \tilde{x}_t = \tilde{x}, \lambda\}}{\sum_{t=1}^{T} \sum_{\forall x} P(x_t = x, \tilde{x}_t = \tilde{x}|z_t, \lambda)} \tag{8}$$

Once each $\mathbf{r}_{\tilde{x}}$ has been reestimated, it is used to form a model of the noise spectrum which is required for Wiener filtering as well as being used to determine $S_{\tilde{x}}$ which is required for the noisy speech model and for gain determination.

Note that the forms of Equations 7 and 8 are reminiscent of the usual parameter reestimation formula for autoregressive HMMs.

For stationary noise, only maximisation with respect to $S_{\tilde{x}}$ is required. Furthermore, Equation 8 can be approximated by the following.

$$\mathbf{r}'_{\tilde{x}} = \frac{\sum_{t=1}^{T} E\{\mathbf{r}_{\tilde{x}}|z_t, x_t = x_t^*, \tilde{x}_t = \tilde{x}, \lambda\}}{T} \tag{9}$$

Here, $x^* = \{x_t^*, t = 1, \ldots, T\}$ is the most likely clean speech state sequence. This can be found by performing a Viterbi alignment on the data.

The term $E\{\mathbf{r}_{\tilde{x}}|z_t, x_t = x_t^*, \tilde{x}_t = \tilde{x}, \lambda\}$ in Equation 9 is evaluated as the inverse Fourier transform of $E\{|\mathbf{V}|^2|z_t, x_t = x_t^*, \tilde{x}_t = \tilde{x}, \lambda\}$. This is calculated using a similar technique to [4]. Each component k of $\mathbf{V}$ is evaluated by the following equation.

$$E\{|V_k|^2|z_t, x_t = x_t^*, \tilde{x}_t, \lambda\} = w_{x_t^*, \tilde{x}_t, k} f_{x_t^*, k} + |w_{x_t^*, \tilde{x}_t, k} Z_{t,k}|^2 \tag{10}$$

Here $w_{x_t, \tilde{x}_t, k}$ is kth component of the Wiener filter for the composite state $(x_t, \tilde{x}_t)$, $f_{x_t, k}$ is the kth component of the Fourier transform of the autoregressive coefficients for clean speech state x_t and $Z_{t,k}$ is the kth component of the Fourier transform of the noisy observation at time t. The Wiener filter in this case is designed to return the MMSE estimator of the noise.

2.3. Wiener Filtering

Once the most likely state alignment has been obtained from recognition using the noisy noisy models, non-causal Wiener filters are formed to return the MMSE estimate of the speech. The filters are formed using $\mathbf{r}_{x_t}$, g_t^2 and $\mathbf{r}_{\tilde{x}_t}$ from the most likely noisy state for each frame. This technique assumes that one state sequence dominates the pdf in (1). For the experiments conducted, little to no improvement in enhancement was observed by relaxing this assumption and forming the weighted sum of Wiener filters.

3. 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 [11]. 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. 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). One 8-state HMM was trained for each of the ten digits and a 1-state HMM was trained for the separating silence. The digits were grouped into files containing 20 each for training and testing purposes. Thus continuous speech recognition was performed. Tests were conducted on 100 different utterances (10 of each digit).

Results for recognition in noise using the clean models and the the combined models determined by the algorithm are given in Table 1. Coloured Noises A and B correspond to Noise Type 06 ('Speech Noise') and Noise Type 12 ('Lynx') from the NOISEX-92 database respectively. The "% Error" figure in Table 1 is derived using the following formula.

$$\% \text{ Error} = \frac{D+S+I}{N} \cdot 100\% \qquad (11)$$

Here D, S and I represent the number of deletions, substitutions and insertions respectively and N is the number of labels in the reference transcription. The number of deletions, substitutions and insertions are also shown explicitly in the table.

For the larger noise sources, recognition using the clean models tended to produce an alignment in which a small number of HMM states fitted most of the data. Therefore, a transcription with a large number of deletions resulted.

The results show that the noise estimation is sufficiently good to improve recognition in noise for this task. Up to five iterations of the noise estimation algorithm were used.

Figure 2 shows the real and estimated (power) spectrum of the 6dB Coloured Noise A on successive iterations of the algorithm. It is seen that the estimated spectrum approaches the true spectrum in this case.

The quality of the enhanced speech was quite high, particularly for the Gaussian noise sources. For these utterances, the main distortion was a slight muffling of the sound. The enhanced coloured-noise speech was less clear, particularly when recognition errors were made and the wrong clean speech model was used for enhancement. Figures 4, 5 and 6 show the clean, noisy and enhanced spectrums for the first few digits of the speech distorted by 6dB Gaussian noise. Note that the silence sections of the isolated digits are actually cleaner in the enhanced version.

To determine the quality of the autoregressive parameters estimated by the algorithm, the Itakura distortion of the system was compared to that from the iterative Wiener

Noise Source	Clean Models % Error (D,S,I)	Combined Models % Error (D,S,I)
None (Clean)	0.0 (0,0,0)	0.0 (0,0,0)
36dB Gaussian	1.0 (0,1,0)	0.0 (0,0,0)
32dB Gaussian	22.0 (0,21,1)	0.0 (0,0,0)
18dB Gaussian	78.0 (48,29,1)	0.0 (0,0,0)
10dB Gaussian	95.0 (95,0,0)	0.0 (0,0,0)
6dB Gaussian	95.0 (95,0,0)	0.0 (0,0,0)
6dB Coloured A	90.0 (85,5,0)	5.0 (0,5,0)
6dB Coloured B	90.0 (87,3,0)	9.0 (0,6,3)

Table 1. Recognition in Noise

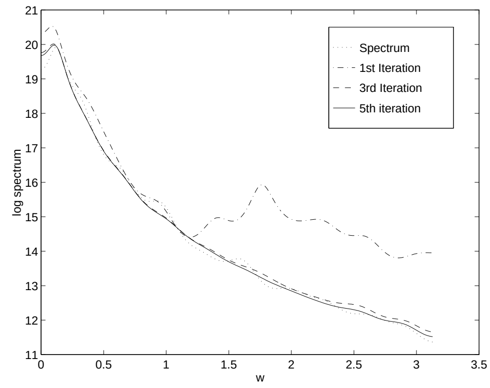

Figure 2. Estimated and Real Power Spectrum of 6dB Coloured Noise A

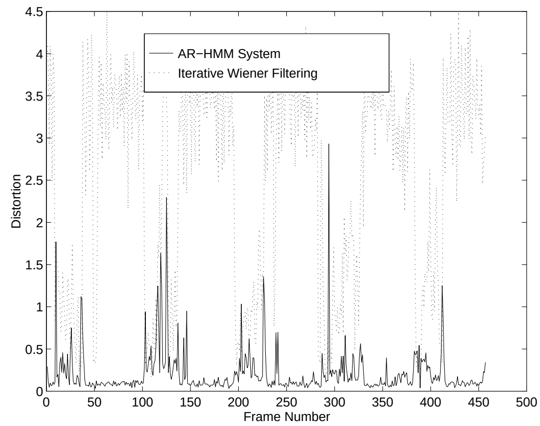

Figure 3. Itakura Distortion Measure for "1 6 3 5 2" 6dB Coloured Noise A

filter system formulated by Lim and Oppenheim [7]. The results for a typical utterance are shown in Figure 3. It is seen that the use of trained clean speech models yields autoregressive parameters that are on better on average than those estimated from the noisy speech.

4. CONCLUSIONS AND FURTHER WORK

A new algorithm that performs enhancement and recognition when only the noisy signal is available has been presented. It uses autoregressive HMMs to model the clean speech and noise. These models are combined and the resulting model used to recognise the speech. The noise model is then reestimated by summing the expected value of the noise statistics given each observation and each HMM state, weighted by the likelihood of being in each state. The process is then repeated until the likelihood converges to a maximum. Enhancement is performed by the application of Wiener filters formed from the speech and noise estimates to each frame. Results presented for additive stationary Gaussian and coloured noise show the algorithm to be effective. The algorithm is potentially extendible to non-stationary noise and this will be the subject of future investigations. The operation of the algorithm on larger databases will also be studied.

5. ACKNOWLEDGEMENTS

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

REFERENCES

[1] Y. Ephraim, "Statistical–model–based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, pp. 1562–1555, 1992.

[2] K. Lee, B. Lee, I. Song, and J. Yoo, "Recursive speech enhancement using the EM algorithm with initial conditions trained by HMM's," in *Proceedings ICASSP*, pp. 621–624, 1996.

[3] H. Sheikhzadeh, H. Sameti, L. Deng, and R. L. Brennan, "Comparative performance of spectral subtraction and HMM-based speech enhancement with application to hearing aid design," in *Proceedings ICASSP*, pp. I13–I16, 1994.

[4] Y. Ephraim, "Gain-adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Transactions on Signal Processing*, vol. 40, pp. 1303–1316, June 1992.

[5] B. Juang, "On the hidden Markov model and dynamic time warping for speech recognition - a unified view," *AT&T Bell Laboratories Technical Journal*, vol. 63, pp. 1213–1243, Sept. 1984.

[6] A. Gray, A. Buzo, R. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-28, pp. 367–376, Aug. 1980.

[7] J. Lim and A. Oppenheim, "All–pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, pp. 197–210, June 1978.

[8] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 40, pp. 725–735, Apr. 1992.

[9] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 245–257, Apr. 1994.

[10] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains," *Ann. Math. Statist.*, vol. 41, pp. 164–171, 1970.

[11] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition," tech. rep., DRA Speech Research Unit, 1992.

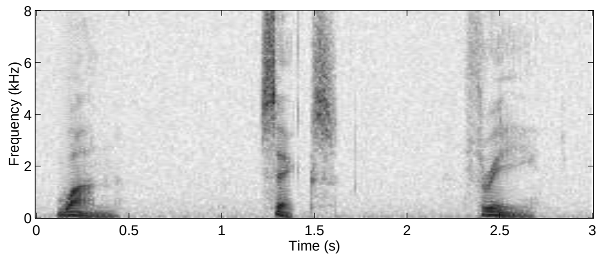

Figure 4. Clean Speech ("1 6 3")

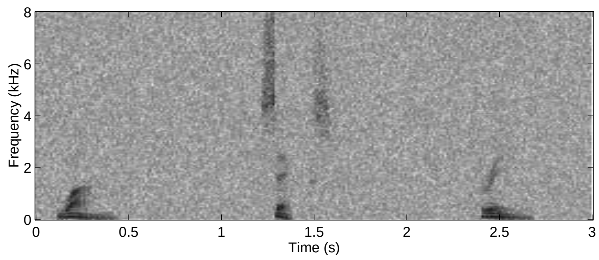

Figure 5. Noisy Speech ("1 6 3") with 6dB Gaussian Noise

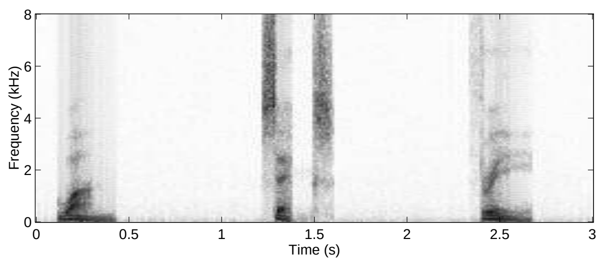

Figure 6. Enhanced Speech ("1 6 3") from 6dB Gaussian Noisy Speech