ELSEVIER SCIENCE B.V. [DTD 4.1.0]
JOURNAL SPECOM ARTICLE No. 1076
PAGES 01-18 DISPATCH 11 July 2000
SPECOM 1076
PROD. TYPE: FROM Disk
A

ELSEVIER
Speech Communication 000 (2000) 000–000
SPEECH COMMUNICATION
www.elsevier.nl/locate/specom

UNCORRECTED PROOF

Adaptive model-based speech enhancement

Beth Logan [a], Tony Robinson [b,*]

[a] *Cambridge Research Laboratory, Compaq Computer Corporation, One Kendall Square, Building 700, Cambridge MA 02139, USA*
[b] *Department of Engineering, Cambridge University, Trumpington Street, Cambridge, CB2 1PZ, UK*

Received 5 October 1999; received in revised form 9 February 2000; accepted 14 February 2000

Abstract

We investigate the enhancement of speech corrupted by unknown independent additive noise when only a single microphone is available. We present adaptive enhancement systems based on an existing non-adaptive technique [Ephraim, Y., 19992a. IEEE Transactions on Signal Processing 40 (4), 725–735]. This approach models the speech and noise statistics using autoregressive hidden Markov models (AR-HMMs). We develop two main extensions. The first estimates the noise statistics from detected pauses. The second forms maximum likelihood (ML) estimates of the unknown noise parameters using the whole utterance. Both techniques operate within the AR-HMM framework.

We have previously shown that the ability of AR-HMMs to model speech can be improved by the incorporation of perceptual frequency using the bilinear transform. We incorporate this improvement into our enhancement systems.

We evaluate our techniques on the NOISEX-92 and Resource Management (RM) databases, giving indications of performance on simple and more complex tasks, respectively. Both enhancement schemes proposed are able to improve substantially on baseline results. The technique of forming ML estimates of the noise parameters is found to be the most effective. Its performance is evaluated over a wide range of noise conditions ranging from −6 to 18 dB and on various types of stationary real-world noises. © 2000 Elsevier Science B.V. All rights reserved.

Résumé

Nous explorons des méthodes d'amélioration de la parole altérée par un bruit additif indépendant avec une source unique. Nous présentons des systèmes d'amélioration adaptative basés sur une technique non-adaptative [Ephraim, Y., 19992a. IEEE Transactions on Signal Processing 40 (4), 725–735]. Cette approche permet de construire des modèles statistiques de la parole et du bruit en utilisant des modèles de Markov cachés auto-regressifs. Nous développons deux méthodes principales. La première méthode estime les modèles statistiques du bruit à partir des silences détectés. La deuxième crée des estimations du maximum de vraisemblance des paramètres du bruit inconnu en utilisant l'ensemble de la phrase. Les deux techniques opèrent sur un schéma auto-regressive hidden Markov models (AR-HMM).

Nous avons montré précédemment que les possibilités des AR-HMMs pour modéliser la parole pouvaientêtre améliorées en incorporant une fréquence de perception utilisant une transformée bilinéaire. Nous introduisons cette correction dans nos systèmes d'amélioration de la parole.

Nos approches sont évaluées sur les bases de données NOISEX-92 et Resource Management, en donnant des indications de la performance respectivement sur des taches simples et complexes. Les deux schémas permettent d'améliorer les résultats de base. La technique qui crée des estimations ML des paramètres du bruit apparaît être la plus efficace. Son efficacité est évaluée sur une large variété de bruits allant de −6 dB à 18 dB et sur divers types de bruit stationnaires réels. © 2000 Elsevier Science B.V. All rights reserved.

[*] Corresponding author. Tel.: +44-1223-332815; fax: +44-1223-332662.
E-mail addresses: beth.logan@compaq.com (B. Logan), ajr@eng.cam.ac.uk (T. Robinson).

0167-6393/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(00)00038-8

Zusammenfassung

Adaptive Model-Basierende Sprachverbesserung

Wir untersuchen die Verbesserung von Sprache die durch eine unbekanntes unabhaengiges additives Geraeusch gestoert ist fuer den Fall, dass nur ein einzelnes Mikrophone verfuegbar ist. Wir presentieren ein adaptives Verbesserungssystem basierend auf existierenden nicht-adaptiven Verfahren [Ephraim, Y., 19992a. IEEE Transactions on Signal Processing 40 (4), 725–735]. Dieser Ansatz modelliert the Sprach und Geraeusch Verteilung durch Benutzung von auto-regressiven "hidden Markov" Modellen (AR-HMMs). Wir haben zwei Haupterweiterungen entwickelt. Die erste bestimmt die Geraeuschstatistik von erkannten Pausen. Die zweite Erweiterung bestimmt "maximum likelihood (ML)" Bewertungen der unbekannten Geraeuschparameter basierenden auf der gesamten Aeuzerung.

Wir haben bereits gezeigt, dass die Faehigkeit von AR-MMs Sprache zu modellieren, verbessert werden kann durch Beruecksichtigung von wahrnehmbaren Frequenzen unter Verwendung der bilinearen Transformation. Wir haben diese Verbessung in unser Erweiterungssystem einbezogen.

Wir bewerten den Erfolg unserer Methoden fuer die NOSEX-92 und "Resource Management (RM)" Datenbanken, resultierend in Bewertungen fuer einfache (im Falle von NOSEX-92) und schwerere Faelle (im Falle von RM). Beide Erweiterungsmethoden erzielen wesentliche Verbesserung wenn verglichen mit Resultaten von Basisverfahren. Es stellt sich heraus, dass die ML Bewertung der Geraeuschparameter die erfolgreichste Methode ist. Ihre Leistung is bestimmt fuer eine weite Spanne von Geraeuschbedingungen (von −6 dB bis 18 dB) und fuer verschiedene Arten von stationaeren "real-world" Geraeuschen. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Speech enhancement; Autoregressive hidden Markov models; Robust speech recognition

1. Introduction

As speech systems have evolved from laboratory demonstrations to real-world applications, the need to maintain performance in a wide variety of situations has emerged. Speech enhancement provides one way of compensating for different environments. It has therefore been investigated by many researchers in recent years (e.g., Gong, 1995).

In this paper, we study enhancement of speech corrupted by additive noise. Our focus is on the situation where the noise statistics are unknown and only one microphone is used. This case is important if speech enhancement is to be used in many real-world systems.

Approaches to speech enhancement can be classified according to the models used to describe the speech and corrupting noise, and the amount of prior information incorporated into these models. The trend in recent years has been to use models of increasing sophistication. One such approach is proposed in (Ephraim, 1992a). This technique models the speech and noise using autoregressive hidden Markov models (AR-HMMs).

We believe the ability to incorporate prior speech information to be one of the main advantages of this enhancement technique. An HMM-based algorithm (Ephraim et al., 1989) has been shown to be superior to spectral subtraction according to Mean Opinion Score evaluations over several noise types (Sheikhzeheh et al., 1994). The approach also has other advantages which make it especially suited to adaptive algorithms. These will become clear in the subsequent discussion.

We present here two extensions to the technique in (Ephraim, 1992a) to make it adaptive. These are based on two main classes of environmental compensation: detecting speech-free portions of the signal and using these to estimate the interfering noise; and making a maximum likelihood (ML) estimate of the noise parameters within a statistical framework.

Pause detection is often cited as means to obtain unknown noise statistics (e.g., Sheikhzeheh et al., 1995). While methods based on energy levels and zero-crossing measures are well established, these techniques tend to rely on thresholds which

UNCORRECTED PROOF

UNCORRECTED PROOF

limit their usefulness in unknown environments (e.g., Deller et al., 1993). Recent work shows that HMMs can be used to more effectively detect pauses in noise (McKinley and Whipple, 1997). In (Logan and Robinson, 1996) we develop an adaptive enhancement system using HMM-based pause detection. In this current paper, we present a more thorough description of our preliminary work with extended results.

ML parameter estimation has been successfully applied to the task of adaptation for speech and speaker recognition (e.g., Lee, 1997). For these applications, schemes based on cepstral features have received the most attention. For example, stochastic matching compensates for unknown convolutional noise (Sankar and Lee, 1996). Here the ML equations are easily solved since speech and noise are additive in this domain.

For the case of additive noise however, working with cepstral features is less trivial. This is because the non-linearity introduced by the logarithm when forming cepstral features makes estimation of unknown parameters using ML mathematically unattractive. In this case, the logarithm must either be approximated by functions, or numerical techniques used to solve for the unknown parameters (Afify et al., 1997; Mokbel, 1997; Moreno et al., 1995).

The problem of adapting to additive noise is easier in the linear spectral domain where again stochastic matching can also be applied (Lee, 1997). However, the distance measure used in this domain is the spectral difference measure which is inferior to the log spectral difference used in the cepstral domain (Rabiner and Juang, 1993).

Enhancement schemes traditionally work in domains which are mathematically suited to additive noise. For example, (Ephraim, 1992a) uses AR-HMMs. These models feature vectors which are additive. Additionally, they use the Itakura–Saito distance measure which is related to the log spectral distance measure. We therefore construct a ML adaptive enhancement scheme based on these models. We have published preliminary results in (Logan and Robinson, 1997a).

Several other adaptive enhancement algorithms which make ML estimates of the unknown parameters have been proposed (Lee et al., 1995; Gannot, 1998; Lee et al., 1996). These approaches estimate the enhanced speech within a Kalman filter framework. Kalman filters, like the Wiener filters used in (Ephraim, 1992a) give the MMSE estimate of the clean speech. However, there is scope in the framework of Ephraim (1992a) to form other estimators. For example, if the enhancement system is used as a front end to a clean speech recogniser, then spectral-based estimators are more applicable (Logan and Robinson, 1998).

In related work, we have presented an extension to AR-HMMs to improve their ability to model speech (Logan and Robinson, 1997b). In this previous work, we incorporate perceptual frequency into the AR-HMM framework and show that this improves recognition performance of a clean speech AR-HMM system. In this current paper, we describe the way in which perceptual frequency can be incorporated into our enhancement systems.

The organisation of our paper is as follows. We first briefly describe the non-adaptive enhancement scheme presented in (Ephraim, 1992a). We then describe our two extensions to make this technique adaptive. In the following section, we describe the incorporation of perceptual frequency into these enhancement schemes.

Sections 5 and 6 detail experimental results. Specifically, we investigate the performance of both the enhancement scheme based on detected pauses and that based on ML noise model estimates. Further, we contrast the effect of using word-based speech models to that of more general speech models. We show results for both small and medium vocabulary systems. Finally, we present conclusions and suggestions for future work.

2. Foundations

We begin with the framework presented in (Ephraim, 1992a). Here, clean speech and noise are modelled using AR-HMMs (Juang, 1984; Juang and Rabiner, 1985). The pdfs for these processes are:

UNCORRECTED PROOF

$$p(\boldsymbol{S}) = \sum_{\boldsymbol{X}} a_{x_0 x_1} \prod_{t=1}^{T} a_{x_t x_{t+1}} b_{x_t}(\boldsymbol{s}_t), \tag{1}$$

$$p(\boldsymbol{D}) = \sum_{\tilde{\boldsymbol{X}}} a_{\tilde{x}_0 \tilde{x}_1} \prod_{t=1}^{T} a_{\tilde{x}_t \tilde{x}_{t+1}} b_{\tilde{x}_t}(\boldsymbol{d}_t), \tag{2}$$

where $\boldsymbol{S}$ is a sequence of K-dimensional clean speech observations, $\boldsymbol{X}$ a sequence of clean speech states, $a_{x_t x_{t+1}}$ the transition probability from state x_t to state x_{t+1} and $b_{x_t}(\boldsymbol{s}_t)$ is the pdf of the output vector $\boldsymbol{s}_t$ from the state x_t. Similarly $\boldsymbol{D}$ is a sequence of noise observations and $\tilde{\boldsymbol{X}}$ is a sequence of noise states.

The pdfs $b_{x_t}(\boldsymbol{s}_t)$ and $b_{\tilde{x}_t}(\boldsymbol{d}_t)$ are assumed Gaussian with zero mean and covariance matrices $\sum_{x_t}$ and $\sum_{\tilde{x}_t}$, respectively. Since the processes are assumed autoregressive, these covariance matrices are dependent on only $P+1$ parameters where P is the order of the autoregressive process (Juang, 1984).

We can combine these speech and noise models to produce a model for noisy speech. The pdf for this model is given by

$$p(\boldsymbol{Y}) = \sum_{\bar{\boldsymbol{X}}} a_{\bar{x}_0 \bar{x}_1} \prod_{t=1}^{T} a_{\bar{x}_t \bar{x}_{t+1}} b_{\bar{x}_t}(\boldsymbol{y}_t), \tag{3}$$

where $\boldsymbol{Y}$ is a sequence of noisy observations and $\bar{\boldsymbol{X}}$ is a sequence of composite states with $\bar{x} \equiv (x_t, \tilde{x}_t)$. For additive, statistically independent noise we have:

$$\boldsymbol{Y} = \boldsymbol{S} + \boldsymbol{D}, \tag{4}$$

$$a_{\bar{x}_t \bar{x}_{t+1}} = a_{x_t x_{t+1}} a_{\tilde{x}_t \tilde{x}_{t+1}}, \tag{5}$$

$$b_{\bar{x}_t}(\boldsymbol{y}_t) = \int b_{\tilde{x}_t}(\boldsymbol{y}_t - \boldsymbol{s}_t) b_{x_t}(\boldsymbol{s}_t) \mathrm{d}\boldsymbol{s}_t, \tag{6}$$

where the pdf $b_{\bar{x}_t}(\boldsymbol{y}_t)$ is Gaussian with zero mean and covariance matrix $\sum_{\bar{x}}$ given by

$$\sum_{\bar{x}} = \sum_{x} + \sum_{\tilde{x}.} \tag{7}$$

We can write the conditional pdf of $\boldsymbol{s}_t$ given $\boldsymbol{Y}$ as

$$p(\boldsymbol{s}_t|\boldsymbol{Y}) = \sum_{\bar{x}_t} p(\bar{x}_t|\boldsymbol{Y}) b_{\bar{x}_t}(\boldsymbol{s}_t|\boldsymbol{y}_t) \tag{8}$$

and thus the MMSE estimate of $\boldsymbol{s}_t$ given $\boldsymbol{Y}$ is given by

$$\hat{\boldsymbol{s}}_t = \sum_{\bar{x}_t} p(\bar{x}_t|\boldsymbol{Y}) H_{\bar{x}_t} \boldsymbol{y}_t, \tag{9}$$

where $H_{\bar{x}_t}\boldsymbol{y}_t$ is the MMSE estimate of the signal $\boldsymbol{s}_t$ given $\boldsymbol{y}_t$. $H_{\bar{x}_t}$ is the Wiener filter formed using the statistics of the composite state $\bar{x}_t$. Thus the MMSE estimator is given by the weighted sum of Wiener filters for each combination of speech and noise states, weighted by the posterior probability of that combination. Other state-based estimators, such as the MMSE spectral amplitude estimator or MMSE log spectral amplitude estimator can also be used instead of Wiener filters (Ephraim, 1992a).

3. Adaptive speech enhancement schemes

We now present two extensions to the work in (Ephraim, 1992c) which enable it to adapt to unknown noise. The first extension estimates the noise statistics using portions of the corrupted signal which have been identified as pauses or 'silence'. The autoregressive probability framework is used for the pause detection. The second scheme makes a ML estimate of the noise parameters within this framework. We assume that the noise statistics are stationary for the duration of the signal to be enhanced.

3.1. Adaptive speech enhancement using detected pauses

If the clean speech model incorporates a model for silence then the compensated model (Eq. (3)) can be used to make a decision about whether a given frame is speech or noise. For example, Viterbi alignment can be performed on the noisy utterance given the compensated model and the labels 'speech' and 'pause' assigned to each frame. If this model was perfect, then the frames labelled as pauses would give good estimates of the noise statistics. This idea forms the basis of a simple adaptive enhancement system.

Fig. 1 shows the algorithm for the system. Frames labelled as pauses are used to reestimate the noise statistics. These statistics are then used to form a new compensated model, and the process repeated. When the likelihood of the utterance

UNCORRECTED PROOF

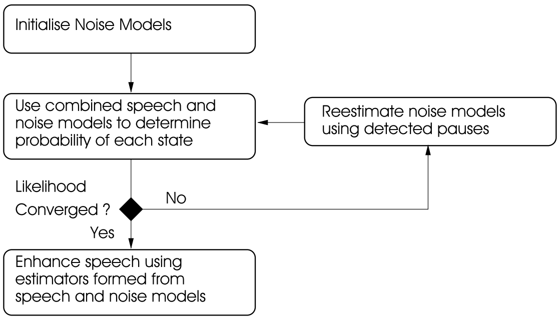
Initialise Noise Models
Use combined speech and noise models to determine probability of each state
Reestimate noise models using detected pauses
Likelihood Converged ?
No
Yes
Enhance speech using estimators formed from speech and noise models

Fig. 1. Basic adaptive enhancement algorithm.

given the compensated model has converged, the speech is enhanced using state dependent estimators.

In this work, we consider two of the estimators proposed in (Ephraim, 1992a): the Wiener filter and a MMSE power spectral density (PSD) estimator. The Wiener filter for composite state $(x_t, \tilde{x}_t)$ is given by

$$H_{x_t,\tilde{x}_t} = \frac{|F_{x_t}|^2}{|F_{x_t}|^2 + |F_{\tilde{x}_t}|^2}, \tag{10}$$

where $|F_{x_t}|^2$ and $|F_{\tilde{x}_t}|^2$, the power spectral densities of the speech and noise states, are estimated using the autoregressive parameters of states x_t and $\tilde{x}_t$, respectively.

The expected PSD of the clean speech at time t given the noisy observations is given by (Ephraim, 1992a)

$$E\{|S_t|^2|\mathbf{y}_t, x_t, \tilde{x}_t, \lambda\} = H_{x_t,\tilde{x}_t}|F_{x_t}|^2 + |H_{x_t,\tilde{x}_t}Y_t|^2, \tag{11}$$

where Y_t is the short-term Fourier transform of the noisy observation at time t.

The required noise statistic is the noise autocorrelation function for each noise state $\mathbf{r}_{\tilde{x}}$ since this is used to generate autoregressive parameters for the noise model. We only consider stationary noise modelled by single state HMMs. We therefore obtain this statistic from the average of the frames labelled as pauses as follows:

$$\mathbf{r}'_{\tilde{x}} = \frac{\sum_{\text{`silence' frames}} \mathbf{r}_{\text{frame}}}{\text{number `silence' frames}}.s \tag{12}$$

3.2. Maximum likelihood noise estimation

The algorithm described in the previous section does not guarantee convergence of the likelihood of the speech utterance given the models. It is also sensitive to modelling errors. We thus turn to a more formal noise estimation scheme.

As described in Section 1, the AR-HMM framework is a good starting point for ML parameter estimation scheme because it models features which are additive, yet it compares features using the Itakura–Saito distortion measure. In this section, we show that the technique of Rose et al. (1994) can be combined with the work of Ephraim to develop an enhancement scheme which can adapt to unknown noise. This is possible because the required likelihood is a linear function of the autocorrelation coefficients.

The procedure closely follows the work in (Rose et al., 1994) but applies it for the first time to AR-HMMs. A full description of the technique is given in (Logan, 1998). To simplify notation, a single mixture component per HMM state is assumed. The extension to multiple mixture systems is straightforward.

Following the method of (Rose et al., 1994), a model for a sequence of noisy observations $\mathbf{Y}$ is derived as:

$$P(\mathbf{Y}|\lambda) = \sum_{\mathbf{X}} \sum_{\tilde{\mathbf{X}}} \int \int_{\mathbf{C}} P(\mathbf{S}, \mathbf{D}, \mathbf{X}, \tilde{\mathbf{X}}|\lambda)\, \mathrm{d}\mathbf{S}\, \mathrm{d}\mathbf{D}. \tag{13}$$

The definition of symbols is as before. λ refers generally to the model parameters: $\{a_{x_t x_{t+1}}\}$; $\{a_{\tilde{x}_t \tilde{x}_{t+1}}\}$; $\{\sum_x\}$ and $\{\sum_{\tilde{x}}\}$. The contour of integration $\mathbf{C}$ is taken over all possible combinations of speech and noise which can form the noisy observation. In this case, additive combinations are considered so $C : S + D = Y$. Given this model, the noise parameters are chosen to maximise the likelihood of the observed data. That is, we find a new estimate of λ, λ', which maximises $P(\mathbf{Y}|\lambda)$.

No closed form solution exists for this maximisation problem. We thus follow the method of Baum et al. (1970) and iteratively maximise an auxiliary function $Q(\lambda, \lambda')$ with respect to λ'. Here, $Q(\cdot)$ is given by

$$Q(\lambda, \lambda') = E\{\log P(\boldsymbol{S}, \boldsymbol{D}, \boldsymbol{X}, \tilde{\boldsymbol{X}}|\lambda')|\boldsymbol{Y}, \lambda\}. \tag{14}$$

To estimate the noise parameters, we need only maximise $Q(\cdot)$ with respect to $\{a_{\tilde{x}_\tau \tilde{x}_{\tau+1}}\}$ and $\{\sum_{\tilde{x}}\}$.

Consider first the maximisation of $Q(\cdot)$ with respect to $\{a_{\tilde{x}_\tau \tilde{x}_{\tau+1}}\}$. As described in (Logan, 1998), the new estimate of $a_{\tilde{x}_\tau \tilde{x}_{\tau+1}}$ is obtained using

$$a'_{\tilde{x}_\tau \tilde{x}_{\tau+1}} = \frac{\sum_{t=1}^{T} p(\tilde{x}_t = \tilde{x}_\tau, \tilde{x}_{t+1} = \tilde{x}_{\tau+1}, \boldsymbol{Y}|\lambda)}{\sum_{t=1}^{T} p(\tilde{x}_t = \tilde{x}_\tau, \boldsymbol{Y}|\lambda)}. \tag{15}$$

Thus the transition probability $a_{\tilde{x}_\tau \tilde{x}_{\tau+1}}$ is reestimated as the sum over all observations of the joint likelihood of state x_τ at time t and state $x_{\tau+1}$ at time $t+1$ and the observation sequence $\boldsymbol{Y}$, scaled by the sum over all observations of the joint likelihood of state x_τ at time t and the observation sequence $\boldsymbol{Y}$.

Now consider the estimation of $\{\sum_{\tilde{x}}\}$. Because the noise is assumed to be modelled by an autoregressive process, each $\sum_{\tilde{x}}$ can be calculated using the autocorrelation function for state $\tilde{x}$, $\boldsymbol{r}_{\tilde{x}}$. We reestimate this statistic as described in (Logan, 1998) using the following equation for each noise state $\tilde{x}$:

$$\boldsymbol{r}'_{\tilde{x}} = \frac{\sum_{t=1}^{T} \sum_{\forall x} p(x_t = x, \tilde{x}_t = \tilde{x}, \boldsymbol{Y}|\lambda) E\{\boldsymbol{r}_{\tilde{x}}|\boldsymbol{y}_t, x_t = x, \tilde{x}_t = \tilde{x}, \lambda\}}{\sum_{t=1}^{T} \sum_{\forall x} p(x_t = x, \tilde{x}_t = \tilde{x}, \boldsymbol{Y}|\lambda)}. \tag{16}$$

Thus the reestimated autocorrelation function is given by the sum over all observations and all clean speech states of the expected value of the autocorrelation function given the particular speech state and noise state, weighted by the joint likelihood of being in that speech and noise state at time t and the observation $\boldsymbol{Y}$.

Note that the forms of Eqs. (15) and (16) are similar to the usual parameter reestimation formulae for AR-HMMs (Juang, 1984) and to the reestimation equations for Gaussian mixture models presented in (Rose et al., 1994).

For stationary noise, only maximisation with respect to $\sum_{\tilde{x}}$ is required. In this case, $p(x_t = x, \tilde{x}_t = \tilde{x}, \boldsymbol{Y}|\lambda)$ can be calculated using the usual forward–backward equations (e.g., Rabiner and Juang, 1993).We can also approximate $p(x_t = x, \tilde{x}_t = \tilde{x}, \boldsymbol{Y}|\lambda)$ by noting that often one state sequence dominates $P(\boldsymbol{Y}, \boldsymbol{X}, \tilde{\boldsymbol{X}}|\lambda)$ (Merhav and Ephraim, 1991). $p(x_t = x, \tilde{x}_t = \tilde{x}, \boldsymbol{Y}|\lambda)$ can thus be replaced by either one or zero depending on whether x_t is part of the dominant state sequence. Therefore Eq. (16) becomes

$$\boldsymbol{r}'_{\tilde{x}} = \frac{\sum_{t=1}^{T} E\{\boldsymbol{r}_{\tilde{x}}|\boldsymbol{y}_t, x_t = x_t^*, \tilde{x}_t = \tilde{x}, \lambda\}}{T}, \tag{17}$$

where $x^* = \{x_t^*, t = 1, \ldots, T\}$ is the most likely clean speech state sequence. This can be found by performing Viterbi alignment using the compensated model on the noisy observations.

The expected value of the autocorrelation function given the composite state $(x_t, \tilde{x}_t)$ and $\boldsymbol{y}_t$ is most easily obtained from the expected value of the noise PSD function $|D|^2$. $|D|^2$ forms a Fourier transform pair with the autocorrelation function which is convenient since it is simpler to work in the frequency domain. Thus the term $E\{\boldsymbol{r}_{\tilde{x}}|z_t, x_t = x_t^*, \tilde{x}_t = \tilde{x}, \lambda\}$ in Eq. (17) can be evaluated as the inverse Fourier transform of $E\{|D|^2|\boldsymbol{y}_\tau, x_\tau = x_t^*, \tilde{x}_\tau = \tilde{x}_t, \lambda\}$.

This can be estimated as shown in (Ephraim, 1992a). Here, the expected value of the kth component of $|D|^2$ is given by

$$E\left\{|D_k|^2|\boldsymbol{y}_\tau, x_\tau = x_t^*, \tilde{x}_\tau = \tilde{x}_t, \lambda\right\} = H_{x_t^*,\tilde{x}_t,k}|F_{x_t^*,k}|^2 + |H_{x_t^*,\tilde{x}_t,k} Y_{t,k}|^2, \tag{18}$$

where $H_{x_t,\tilde{x}_t,k}$ is the kth component of the Wiener filter for the composite state $(x_t, \tilde{x}_t)$, $|F_{x_t,k}|^2$ the kth component of the Fourier transform of the autoregressive coefficients for clean speech state x_t and $Y_{t,k}$ is the kth component of the Fourier transform of the noisy observation at time t. This Wiener filter is designed to return the MMSE estimator of the noise, so its transfer function is given by

$$H_{x_t^*,\tilde{x}_t,k} = \frac{|F_{\tilde{x}_t,k}|^2}{|F_{x_t^*,k}|^2 + |F_{\tilde{x}_t,k}|^2}. \tag{19}$$

Eq. (18) is derived assuming that the covariance matrices $\sum_x$ and $\sum_{\tilde{x}}$ of the speech and noise processes are circulant. This assumption holds assuming sufficiently large K (Ephraim, 1992b).

The new adaptive enhancement algorithm operates as shown in Fig. 2. Comparison of Fig. 1

UNCORRECTED PROOF

UNCORRECTED PROOF

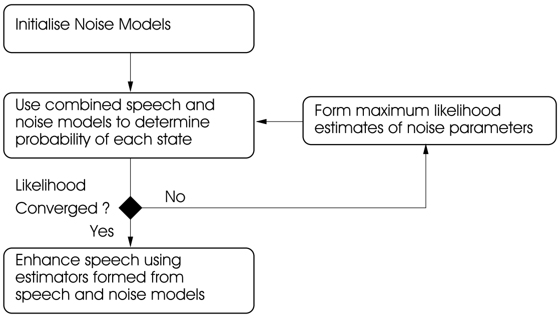
Initialise Noise Models
Use combined speech and noise models to determine probability of each state
Form maximum likelihood estimates of noise parameters
Likelihood Converged ?
No
Yes
Enhance speech using estimators formed from speech and noise models

Fig. 2. Improved adaptive enhancement algorithm.

with Fig. 2 shows that the only difference between the systems is the technique of noise reestimation.

4. Incorporation of perceptual frequency

In (Logan and Robinson, 1997b) we improve the modelling power of AR-HMMs by incorporating perceptual frequency. That is, we use a perceptual frequency scale when calculating the distortion between an observation and a trained model. We show that this improves the performance of clean speech recognition systems substantially.

We incorporate perceptual frequency into AR-HMM systems using the bilinear transform to perform the frequency warping (Oppenheim and Johnson, 1972). This transform has previously been used to improve the performance of linear prediction coding systems (Strube, 1980) and LPC-cepstral recognition systems (Shikano, 1985), (Mokbel and Chollet, 1995).

The bilinear transform converts a time sequence to a new sequence with a warped spectrum. By adjusting the so-called warping factor, the degree of warping can be made to be a very good approximation to the perceptually meaningful Bark scale.

We apply the transform to autocorrelation coefficients as described in (Strube, 1980). For our given sampling rate of 16 kHz, we use a warping factor of 0.57. The 'warped' autocorrelation coefficients are then used to determine LPC coefficients in the usual way and thus train 'warped' AR-HMMs. These warped models can be used in perceptual frequency recognition or enhancement systems. In (Logan and Robinson, 1997b), the focus is mainly on clean speech recognition performance. In this paper, we discuss the construction of enhancement systems.

4.1. Perceptual frequency AR-HMM enhancement systems

Our enhancement systems employ a weighted sum of estimators, where the weights are calculated using the pdf of a compensated AR-HMM given the noisy observations. Until now this compensated AR-HMM has used a linear frequency scale. Using a perceptual frequency scale however leads to a more accurate model, hence a better choice of filters for enhancement (Logan, 1998). We therefore construct a perceptual frequency enhancement system as shown in Fig. 3.

Note here that we use the warped models in conjunction with warped observations for the probability calculations, while non-warped (i.e., unprocessed) AR-HMMs are used to construct the state-based estimators. We perform estimation in the non-warped domain because it is computationally expensive to warp and unwarp time domain observations.

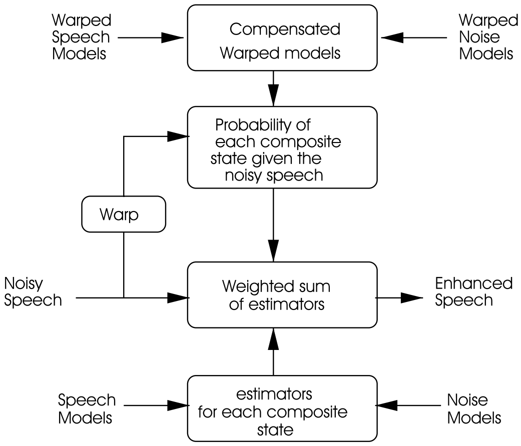
Warped Speech Models
Compensated Warped models
Warped Noise Models
Probability of each composite state given the noisy speech
Warp
Noisy Speech
Weighted sum of estimators
Enhanced Speech
Speech Models
estimators for each composite state
Noise Models

Fig. 3. A perceptual frequency enhancement system. Here the weights for each estimator are determined using perceptual frequency AR-HMMs. The estimators themselves are constructed using non-warped AR-HMMs.

We therefore need to obtain unwarped versions which correspond exactly to each warped speech and noise model for use with the non-warped observations. However, it is never possible to exactly unwarp a given model because the warping process transforms a finite sequence to an infinite sequence (Oppenheim and Johnson, 1972). We could unwarp the filters directly by sampling the warped spectrum but again mismatch is introduced.

We wish to minimise any mismatch introduced by the unwarping process because one of our evaluation procedures is the performance of the enhanced speech with a standard clean speech recogniser. We therefore use parallel warped and non-warped models and train the non-warped versions using single pass retraining.

Given a set of models, single pass retraining (e.g., Young et al., 1996) generates a parallel set of models using different training data. This is achieved by computing the state probabilities using the original models and the original training data, but then switching to a new set of training data to compute parameter estimates for the new model. Thus given parallel warped and non-warped observations and warped models, non-warped models which correspond exactly to the warped models can be trained.

The enhancement process now requires warped and non-warped observations. We use the warped observations in conjunction with warped models to calculate the probability of each composite state. The statistics of each state in the corresponding non-warped model are then used to construct estimators which operate on non-warped observations.

In order to implement the adaptive enhancement schemes described in the previous section, we estimate the noise models in parallel in both the warped and non-warped domains. We again use the compensated warped models in conjunction with warped observations for all probability calculations.

5. Small vocabulary experiments

To evaluate our systems we first conduct small vocabulary, speaker dependent enhancement experiments. We use the male isolated digits additive noise task from the NOISEX-92 database (Varga et al., 1992). This database contains clean speech from two speakers, one male and one female, and the same speech corrupted by various noise sources at SNRs ranging from −6 to 18 dB. The SNR values are defined by NOISEX and based on auditory weighting. The database is generated by adding the noise to the clean speech hence speech artifacts due to speaker stress in noise are not present. We consider the following four stationary additive noise sources: Lynx helicopter noise; speech noise; car noise and F16 aircraft noise.

Our primary evaluation metric is the clean speech recognition performance of the enhanced speech. This quantitatively describes how effective each enhancement scheme would be if used as a front end to a clean speech recognition system. We use a standard clean speech MFCC recognition system for recognition tests, deriving the enhanced MFCC features directly from the enhanced spectra without resynthesising the speech in the time domain.

We additionally quantitatively measure perceptual improvements using the Itakura distortion measure (e.g., Gray et al., 1980). This measures the frame-by-frame distortion between each enhanced signal and the corresponding clean speech. We report the distortion averaged over the entire utterance and additionally over only the speech portions. We refer to these measures as ‘All’ and ‘Speech’ distortion, respectively.

Although we did not conduct formal listening tests, we informally noted a great deal of correlation between the quantitative results and perceptual improvements. As discussed in Section 3.1 we experimented with two of the estimators described in (Ephraim, 1992a): the Wiener filter and a MMSE PSD estimator. In all our experiments, we observed that the former estimator gives perceptually more pleasing speech as it tends to suppress noise in non-speech regions more thoroughly. Conversely, the MMSE PSD estimator produces enhanced speech which gives superior performance with a clean speech recogniser. This is because the clean speech recognition system is strongly influenced by errors in the spectral domain. We there-

UNCORRECTED PROOF

UNCORRECTED PROOF

fore use the MMSE PSD estimator to produce all the quantitative results described here.

The following sections discuss our experiments in detail. We investigate our two noise estimation schemes and examine two different ways of modelling the clean speech: using word-based models and using general mixture models.

To enhance legibility, we show summarised results. These are the distortion measures and recognition error rates averaged over all noise types for each SNR. In these tables, D, S and I are the percentage of deletion, insertion and substitution errors, respectively. The error rate is given by: $\%\mathrm{Error} = D + S + I$. We perform statistical analyses on selected results using the Matched-Pairs test (Gillick and Cox, 1989) at a confidence level of 95%.

5.1. Clean speech word-based AR models

Both warped and non-warped clean speech AR-HMMs are required for the enhancement systems. We initially construct word-based models. We use an 8-emitting state left-to-right HMM model for each digit and a 1-emitting state model for silence. We use autoregressive models of order 20 and 2 mixture components per state.

The speech is parameterised using frames of 32 ms with overlap of 16 ms. These parameters are chosen to be convenient for construction of enhanced time domain waveforms. The models are trained on 100 clean digit utterances. The standard Baum–Welsh algorithm is used for training the warped models and single pass retraining is used to train the non-warped models as described in Section 4.1.

5.2. Noise AR models

We model each type of noise using a single state AR-HMM with autoregressive order 20. Each model is initialised by assuming that the whole utterance is noise.

5.3. MFCC recognition system

We use a standard MFCC HMM recognition system to evaluate the clean speech recognition performance of the enhanced speech. We construct this using the HMM Toolkit V1.5 (Young et al., 1993). Again we use an 8-emitting state left-to-right HMM model for each digit and a 1-emitting state model for silence. We model 15 cepstral coefficients including the zeroth coefficient using one mixture component per state and diagonal covariance matrices. Again the speech is parameterised using frames of 32 ms with overlap of 16 ms.

Again we use the standard Baum–Welsh algorithm for training. Connected word Viterbi decoding is used for recognition (i.e., not isolated word recognition). The syntax for the recognition network is constrained to be a string of digits each separated by silence.

The models are trained using 100 clean digit utterances. Recognition experiments are conducted on 100 enhanced digits for each noise type at each SNR ranging from −6 to 18 dB, giving 400 test digits for each of these SNRs. For processing the digits are grouped into 5 files of 20 digits each.

5.4. Baseline performance

We first investigate the distortion measures and performance of the MFCC recognition system

Table 1
Itakura distortions and word error rates for clean speech and speech corrupted by the four noises recognised using clean and matched MFCC models

SNR (dB)	Distortion		% Error (D, S, I)			
	Overall	Speech	Clean MFCC models		Matched MFCC models	
∞	0.00	0.00	0.00	(0, 0, 0)	0.00	(0, 0, 0)
18	0.66	0.34	54.50	(14.25, 24.25, 16)	0.00	(0, 0, 0)
12	0.84	0.55	77.00	(22, 40, 15)	0.00	(0, 0, 0)
6	0.99	0.78	92.00	(92, 0, 0)	0.25	(0, 0.25, 0)
0	1.10	1.00	95.00	(95, 0, 0)	2.50	(0, 2.5, 0)
−6	1.18	1.18	95.00	(95, 0, 0)	32.50	(9.25, 20.25, 3)

UNCORRECTED PROOF

Table 2
Itakura distortions and word error rates for corrupted speech enhanced adaptively using detected pauses to estimate the noise; MMSE PSD estimation and word-based HMMs

SNR (dB)	Distortion		% Error (D, S, I)	
	All	Speech	MFCC models	
18	0.22	0.17	1.75	(0, 1.25, 0.5)
12	0.34	0.23	6.00	(0, 5.25, 0.75)
6	0.61	0.40	17.75	(0.5, 15, 2.25)
0	0.97	0.77	51.25	(33, 17.75, 0.5)
−6	1.20	1.00	73.50	(25, 27.25, 21.25)

when no enhancement is applied. Table 1 shows the summary distortion and recognition error rates for clean and noisy speech. Results for both clean and matched MFCC models are shown. Here ‘matched’ refers to the situation when the training and testing conditions are identical. We obtain the matched models using single pass retraining for each noise type at each SNR.

These recognition results are obtained by optimising the insertion penalty and another parameter, the silence probability increment, for each test condition. This latter parameter, described in (Seymour, 1996), weights the log observation probability of the silence model by a fixed value to improve the chance of low energy frames at word boundaries being recognised correctly as silence.

We see from the results that the performance of the clean speech recognition system degrades rapidly in the presence of noise. The matched model performance gives an indication of the improvement possible using enhanced speech.

5.5. Word-based models

We now investigate the performance of enhancement systems containing word-based clean speech models. For these experiments, we use the recognition network combined with Viterbi alignment to obtain the most likely speech and noise state for each frame given the noisy observation. The most likely mixture component given this state is then determined. The speech and noise statistics for this mixture component are then used to construct estimators for enhancement. These statistics are also used for the ML noise estimation scheme.

We found that the optimal insertion penalty used during Viterbi alignment varied according to the noise type and SNR. However, the SNR is the dominating factor with deletions more prominent at low SNRs and insertions more prominent at high SNRs. In order to automatically select the best penalty for testing, we first determine the optimal mapping between SNR and insertion penalty for each enhancement system using the training utterances. When testing, we approximate the SNR for each test utterance using the NIST tool *wavmd*. [1] This is then mapped to an insertion penalty.

5.5.1. Noise estimation from detected pauses

We first investigate the enhancement scheme described in Section 3.1 which estimates the noise from detected pauses. The performance of this system is summarised in Table 2. The results demonstrate considerable improvement over the clean speech model results in Table 1. It should be noted though that the performance falls short of the matched model results in this table.

5.5.2. Maximum likelihood noise parameter estimation

We next investigate the ML parameter estimation scheme described in Section 3.2. Table 3 summarises the performance of this system. These results are significantly better than those from the technique of estimating the noise from the detected pauses. They are also comparable to the matched model baseline results in Table 1. We thus con-

[1] *wavmd* is available as part of the Sphere 2.5 software package from NIST at http://www.itl.nist.gov.

UNCORRECTED PROOF

Table 3
Itakura distortions and word error rates for corrupted speech enhanced adaptively using ML noise parameter estimation; MMSE PSD estimation and word-based HMMs

SNR (dB)	Distortion		% Error (D, S, I)	
	All	Speech	MFCC models	
18	0.13	0.12	0.00	(0, 0, 0)
12	0.18	0.18	0.00	(0, 0, 0)
6	0.26	0.27	0.50	(0, 0.5, 0)
0	0.39	0.42	6.25	(0.25, 3.5, 2.5)
−6	0.63	0.69	26.50	(7.5, 13.25, 5.75)

clude that the extra complexity of the ML technique is justified.

5.6. General speech models

We now investigate the effect of using more general speech models in the enhancement system. Here, ‘general’ refers to a model which is trained on clean speech without prior knowledge of the words spoken. In this case, we use an ergodic two-state HMM. The first state models speech using 128 mixture components and the second state models silence using a single mixture component. Transitions between the states are freely allowed. We concentrate on the ML noise parameter estimation scheme since this has been shown to have superior performance.

We construct warped and non-warped clean speech AR-HMM models. The warped AR-HMM models are initialised using single pass retraining from a MFCC system with identical topology. The speech state of this MFCC system is trained using K-means clustering on 100 clean digits. After initialisation, the warped AR HMMs are reestimated using Baum–Welch reestimation. Single pass retraining is then used to train the non-warped AR models from the warped AR models as before.

In the previous experiments, Viterbi alignment is used to obtain the most likely speech and noise states corresponding to each frame. This was logical as we had word-based models and so could perform recognition using the compensated AR-HMMs to determine the best state-frame alignment to use for enhancement.

With the more general speech models used here, we instead determine the posterior probability of each mixture component using the forward–backward equations (e.g., Rabiner and Juang, 1993). This probability is then used to construct a weighted sum of estimators for noise reestimation and enhancement.

Table 4 summarises the results for a 128-mixture enhancement system using ML noise parameter estimates. These results are inferior to the word-based system despite having a comparable number of parameters. Thus we conclude that some performance is sacrificed by the use of simpler speech models in the enhancement system.

We experimented with increasing the number of mixture components in the speech model and with

Table 4
Itakura distortions and word error rates for corrupted speech enhanced adaptively using ML noise parameter estimation; MMSE PSD estimation and general HMMs; 128 mixture components

SNR (dB)	Distortion		% Error (D, S, I)	
	All	Speech	MFCC Models	
18	0.15	0.13	2.50	(0.75, 1.5, 0.25)
12	0.25	0.22	15.75	(4,11, 0.75)
6	0.43	0.37	51.50	(23.25, 25.75, 2.5)
0	0.67	0.56	78.75	(68.5, 10.25, 0)
−6	0.90	0.79	85.75	(83, 2.75, 0)

UNCORRECTED PROOF

Table 5
Itakura distortions and word error rates for corrupted speech enhanced adaptively using ML noise parameter estimation; MMSE PSD estimation and general HMMs; 32 × 4 mixture components

SNR (dB)	Distortion		% Error (D, S, I)	
	All	Speech	MFCC models	
18	0.15	0.14	2.00	(0.25, 1.5, 0)
12	0.22	0.22	10.00	(1, 6, 3)
6	0.34	0.35	30.75	(6.5, 17.5, 6.75)
0	0.57	0.52	63.00	(23.75, 29.75, 9.5)
−6	0.85	0.75	81.00	(45.5, 24, 11.5)

introducing more temporal information by modelling the speech using several states. This latter variation was the most effective.

Table 5 shows the performance for a 33-state model. The first 32 states contain 4 mixture components each and model the speech. The final state contains a single mixture component and models silence as before. The training procedure for this model is different to the 128 mixture component case. Here, the warped AR-HMM is trained by continually splitting mixture components from an initial single mixture system. The non-warped AR-HMM is trained using single pass training as before. The results for this system are significantly better than the 128 mixture component system. Thus there appears to be some advantage in adding temporal information. However, the performance of this system is still substantially worse than the word-based scheme.

6. Medium vocabulary experiments

We now investigate the performance of our algorithm on the more challenging Resource Management (RM) task (Price et al., 1988). This database is suitable for medium vocabulary (around 1000 words) continuous speech experiments and contains speaker dependent and speaker independent data. We use the speaker independent data for our enhancement experiments.

The training data for the speaker independent task consists of 100 speakers with 40 utterances per speaker. The testing data is divided into four sets each consisting of 300 utterances comprising 30 sentences spoken by 10 speakers. These are labelled: “Feb89”, “Oct89”, “Feb91” and “Sep92”. Only clean speech is supplied so noise is added to the test sets as described below.

We conduct experiments using enhancement systems based on general speech models. Given the results of the previous section, we would have liked to use word or phone based models. However, we were unable to construct a clean speech AR-HMM system with sufficient accuracy on the RM task. We believe this is because our AR-HMM system does not incorporate frame-to-frame spectral changes (delta features). Also, only one variance parameter is determined for each mixture component unlike MFCC HMM systems which calculate the variance for each cepstral component (and indeed delta and acceleration components). These modelling deficiencies play a greater role in this medium vocabulary speaker independent task than the simple task in the previous section.

Therefore, we study enhancement schemes based on general speech models. Only the ML noise parameter estimation scheme is investigated since this gives the best performance. We again use enhanced MFCC parameter recognition performance to evaluate our system. We additionally conduct some informal perceptual tests. All recognition results shown are for enhancement system employing MMSE PSD estimators. Conversely, all perceptual results are for systems employing Wiener filters.

6.1. Formation of noise corrupted data

We add noise to the test sets to conduct the enhancement experiments since the RM database

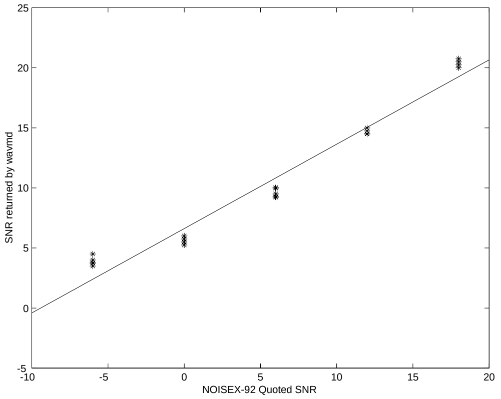

Fig. 4. Performance of the NIST utility *wavmd* on Lynx Noise in NOISEX-92 with a straight line fitted to the data. Note that *wavmd* tends to overestimate the SNR compared to the quoted values.

contains clean speech only. For each test utterance, we extract a random segment of Lynx noise from the NOISEX-92 database, scale it and add it to the utterance.

We consider two noise conditions. These correspond to the attenuation of the Lynx noise by 20 and 12 dB, respectively. The NIST utility *wavmd* is used to estimate the SNR of the corrupted utterances. This gives average SNRs for the two noise conditions of 18 and 12 dB.

wavmd estimates the SNR as a log ratio of speech to noise power where the noise power is estimated without prior knowledge of the noise statistics. Therefore, the SNRs determined are necessarily different to those calculated by NOISEX-92. Fig. 4 shows SNRs calculated using *wavmd* on the NOISEX-92 database verses the quoted SNRs for Lynx noise. We see that for this noise, *wavmd* tends to overestimate the SNR compared to the quoted values, particularly at low SNRs. This should be borne in mind when comparing results from this section with those in Section 5.

6.2. Enhancement system

We model clean speech using general mixture models. Similar to the models described in Section 5.6, we model speech by one or more states with many mixture components and silence by an additional state. The training procedure used is identical to that described in Section 5.6. We use the first three utterances for each training speaker as training data in order to avoid over-training of the models.

We again use one state noise models. For these experiments, we initialise these models from the frame of the test utterance with the minimum power. This was found to give enhanced speech which was perceptually superior to that from a system which initialised the noise using all of the test frames. A small amount of recognition performance is sacrificed by this initialisation technique (about 1% absolute on the 512 mixture system at 18 dB).

We use the approximation of the state-dependent pdf $(b_{\bar{x}_t m_t}(\boldsymbol{y}_t))$ described in (Sheikhzeheh et al., 1995) to calculate the compensated models. This assumes that the autocorrelation function of each state of the compensated model can be approximated by the sum of the corresponding speech and noise autocorrelation functions. This approximation is necessary to make the system computationally tractable.

6.3. MFCC recognition system

The clean MFCC recognition system used for performance evaluation is trained using the RM Toolkit as a template (Young et al., 1993). A 5 mixture component per state, triphone-clustered system is built. Each triphone is modelled by a 3-state left-to-right HMM with diagonal covariance matrices. The feature vectors contain 13 cepstral coefficients including the zeroth coefficient augmented with delta and delta–delta coefficients. These are the first 13 coefficients returned from a MFCC analysis of order 24. The data is pre-emphasised by the filter $H(z) = 1 - 0.97z^{-1}$.

The frame rate and frame size are 16 and 32 ms, respectively as used as in the previous enhancement experiments. These differ from the standard values for these parameters used by the RM Toolkit. The non-standard frame rate affects the modelling of short phones by increasing the minimum duration. This problem was alleviated by the introduction of a skip state into each triphone

UNCORRECTED PROOF

UNCORRECTED PROOF

Table 6
Baseline results for the RM database speaker independent test sets ("Feb89", "Oct89", "Feb91" and "Sep92") for clean speech and speech corrupted by Lynx noise[a]

Model	SNR (dB)	% Error				
		Feb89	Oct89	Feb91	Sep92	Average
Clean	∞	6.3	7.3	5.9	11.0	7.6
	18	38.9	30.4	35.8	43.1	37.0
	12	80.4	81.0	77.7	85.2	81.1
Matched	18	16.7	14.8	14.1	21.0	16.7
	12	40.8	31.3	34.0	40.4	36.6

[a] Performance using clean and matched models is shown.

model. The frame rate also affects the period of time used to calculate the delta and delta–delta coefficients. This effect was not considered.

6.4. Baseline performance

Table 6 shows the word error rates for the clean and noisy speech on the four test sets. The clean baseline is worse than the published performance on this database because of the decreased frame rate as discussed. We see that the addition of noise has a substantial effect on the error rate.

Also in this table are the word error rates achievable when the training and testing conditions are matched. The matched MFCC models are obtained by adding Lynx noise to the training set and then using this data to train models using single pass retraining. These results give an indication of the best performance achievable by any enhancement system.

6.5. Enhancement performance

We first investigate the 18 dB noise condition. Table 7 shows the word error rates for various numbers of mixture components in the models. It can be seen that a substantial improvement has been made on the baseline performance. The performance improves as the number of mixture components increases, although the difference between the 512-mixture and 256-mixture systems is not significant.

The last row of this table shows the error rates for the 512 mixture component system at 12 dB. Again substantial improvements have been made over the baseline performance.

From these two test conditions, it appears that the improvement gained by the enhancement technique halves the error rate. However this performance is significantly worse than the matched model results given in Table 6 suggesting that there is a modelling deficiency.

In Section 5.6 and Seymour (1996), we saw that including temporal information improves the performance of general speech models. Therefore we implement and test a 32-state, 16 mixture component per state model similar to Seymour (1996).

The results of this experiment for the two noise conditions are shown in Table 8. The results at 18 dB are not significantly different from the 512 mixture component system. The 12 dB results are significantly worse than the 12 dB 512 mixture component system.

Table 7
Enhancement results for the RM speaker independent test sets for Lynx noise[a]

SNR (dB)	No. mixes	% Error				
		Feb89	Oct89	Feb91	Sep92	Average
18	128	23.6	20.1	21.7	27.6	23.2
	256	18.7	16.5	18.9	23.9	19.5
	512	18.1	15.5	18.1	24.2	19.0
12	512	42.8	35.7	37.9	46.7	40.8

[a] The speech is enhanced using general speech models with varying numbers of mixture components.

Table 8
Enhancement results for the RM speaker independent test sets for Lynx noise at various SNRs[a]

SNR (dB)	% Error				
	Feb89	Oct89	Feb91	Sep92	Average
18	18.0	16.4	18.0	24.5	19.2
12	46.0	38.9	42.3	48.5	43.9

[a] Speech enhanced using general speech models with 32-state, 16 mixture component/state models.

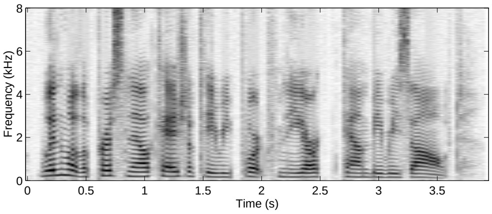

Fig. 5. Clean speech spectrogram for the first sentence for speaker alk0_3.

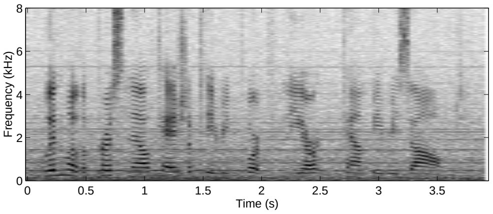

Fig. 6. Speech corrupted by Lynx noise at 12 dB; first sentence for speaker alk0_3.

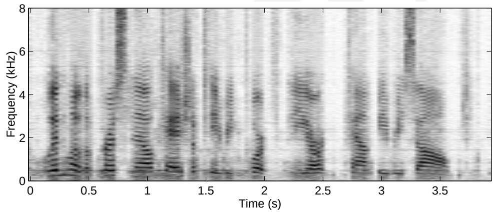

Fig. 7. Speech corrupted by Lynx noise at 12 dB enhanced using Wiener filters formed from 512 mixture component models; first sentence for speaker alk0_3.

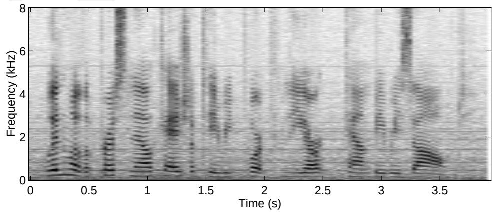

Fig. 8. Speech corrupted by Lynx noise at 18 dB; first sentence for speaker alk0_3.

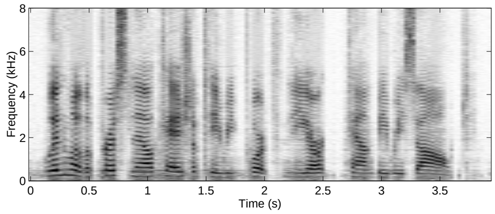

Fig. 9. Speech corrupted by Lynx noise at 18 dB enhanced using Wiener filters formed from 512 mixture component models; first sentence for speaker alk0_3.

Thus we did not observe any improvement in performance for a multi-state system. One reason for the inferior results may be the different training procedures for the models since the 32 × 16 system was formed by continually splitting a AR-HMM system whereas the 512 system was initialised from single pass retraining on a MFCC system. It seems that the superior distortion measure of the MFCC system provides some advantage for initialisation.

Informal listening tests indicate that the enhanced speech contains some residual noise. This is quite considerable and annoying for the speech at 12 dB. Figs. 5–9 show spectrograms of the clean, noisy and enhanced speech for the first sentence for speaker ‘alk0_3’. The enhancement is performed using the 512-mixture system. The text of this sentence is ‘WHEN WILL THE PERSONNEL CASUALTY REPORT FROM THE YORKTOWN BE RESOLVED’. Residual noise is evident in the enhanced spectrograms.

Seymour reported much less residual noise for experiments with a non-adaptive enhancement

UNCORRECTED PROOF

UNCORRECTED PROOF

system on this task at 18 dB (Seymour, 1996). Since in this previous work, the estimators are chosen according to MFCC probabilities, it seems likely that the inferior modelling ability of AR-HMMs is causing the increased distortion. Future work should thus focus on improving the modelling ability of AR-HMMs. We have identified three main areas to be addressed. The first is the incorporation of delta features. This would hopefully allow a word- or phoneme-based enhancement system to be built. The second area for improvement is the incorporation of more variance information. As discussed in (Logan and Robinson, 1997b), AR-HMMs currently have one variance parameter per mixture component whereas MFCC systems train a variance parameter for each cepstral parameter.

The third main area to be addressed is that of gain normalisation. We did not incorporate any gain normalisation into our system since this is a non-trivial problem for compensated AR-HMMs. Ephraim has proposed a technique to iteratively determine a gain contour in this case, but this scheme is too computationally expensive for the medium vocabulary system examined here (Ephraim, 1992b).

7. Conclusions and future work

We have investigated the problem of enhancing speech corrupted by additive noise in an unknown environment when only one microphone is available. We have developed techniques based on a non-adaptive enhancement system by Ephraim (Ephraim, 1992a). This technique models speech and noise statistics using AR-HMMs.

Working in the AR-HMM domain has three advantages for enhancement of additive noise. The first is that the feature vectors used are linearly combinable. This is important when forming a compensated system to model the corrupted speech and also when forming ML estimates of unknown parameters. The second advantage is that the distortion measure used to compare features to templates is the Itakura–Saito distortion measure. This is more effective than a linear spectral distortion measure which would be the metric used if a linear spectral HMM system was built. Finally, the HMM framework allows MMSE spectral estimators as well as time domain estimators to be formed. In our work, we found the former to be better suited to the task of using an enhancement system as a front-end to a clean speech recogniser.

We extended the work in (Ephraim, 1992a) by estimating the noise statistics directly from the signal to be enhanced. Two main approaches were developed. The first considers estimating the noise from detected pauses. The AR-HMM framework is used for the pause detection. The second approach uses ML parameter estimation to estimate the noise statistics given a compensated AR-HMM model of the noisy speech.

Our enhancement schemes use perceptual frequency AR-HMMs. Here we use the bilinear transform to warp the frequency spectrum of our models to an approximation of the Bark scale. We have previously shown that this substantially improves clean speech recognition performance (Logan and Robinson, 1997b).

The schemes were evaluated using the NOISEX-92 and RM databases providing information about performance on small vocabulary speaker dependent and medium vocabulary speaker independent tasks, respectively. We evaluated the enhancement performance using distortion measures and clean speech recognition results. Limited informal listening tests were also conducted.

Both enhancement schemes were able to substantially improve on baseline results. On the small vocabulary task, we found that the technique of making ML estimates of the noise statistics was significantly better than the technique of estimating the noise statistics from pauses. The former technique had performance which was comparable to a matched system. We also found that while speech enhanced using Wiener filters is perceptually more pleasing, if the system is to be used as a front end to a clean speech recognition system, then a MMSE PSD estimator is more appropriate. Finally, we found that word-based models were superior to general speech models, but that temporal information could improve the latter.

The medium vocabulary experiments focused on the ML adaptation technique. Due to model-

UNCORRECTED PROOF

ling deficiencies in AR-HMMs, we were unable to use a word- or phoneme-based enhancement system. We therefore studied an enhancement system based on general speech models. Although the performance of this system was significantly worse than a matched system, substantial improvements over baseline results were seen.

Future work should focus on improving the modelling ability of AR-HMMs in three main areas: the incorporation of delta parameters; the improvement of the distortion measure used and appropriate gain normalisation. These appear to be the main factors limiting the performance of our system on more complex tasks.

Acknowledgements

For the duration of this work, B.T. Logan was funded by a Packer Scholarship, administered by the Cambridge Commonwealth Trust, and an ORS award. Additional funds were received from the Charles Hesterman Merz Fund. Thanks are also due to Carl Seymour for the use of code and scripts, and for helpful discussions. Finally, the authors wish to acknowledge the helpful comments of two anonymous reviewers.

References

Afify, M., Gong, Y., Haton, J., 1997. A unified maximum likelihsood approach to acoustic mismatch compensation: application to noisy Lombard speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.

Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Statist. 41, 164–171.

Deller, J.R., Proakis, J.G., Hansen, J., H.L, 1993. Discrete-Time Processing of Speech Signals. Macmillan, New York.

Ephraim, Y., 1992a. A Bayesian estimation approach for speech enhancement using hidden Markov models. IEEE Trans. Signal Process. 40 (4), 725–735.

Ephraim, Y., 1992b. Gain-adapted hidden Markov models for recognition of clean and noisy speech. IEEE Trans. Signal Process. 40 (6), 1303–1316.

Ephraim, Y., 1992c. Statistical-model-based speech enhancement systems. Proc. IEEE 80, 1526–1555.

Ephraim, Y., Malah, D., Juang, B.H., 1989. On the application of hidden markov models for enhancing noisy speech. IEEE Trans. Acoust. Speech Signal Process. 37 (12), 1846–1856.

Gannot, S., Burshtein, D., Weinstein, E., 1998. Iterative and sequential Kalman filter-based speech enhancement algorithms. IEEE Trans. Speech Audio Process. 6 (4), 373–385.

Gillick, L., Cox, S.J., 1989. Some statistical issues in the comparison of speech recognition algorithms. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 532–535.

Gong, Y., 1995. Speech recognition in noisy environments: a survey. Speech Communication 16, 261–291.

Gray, A.H., Buzo, A., Gray, R.M., Matsuyama, Y., 1980. Distortion measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. ASSP-28 (4), 367–376.

Juang, B.H., 1984. On the hidden Markov model and dynamic time warping for speech recognition – a unified view. AT&T Bell Lab. Tech. J. 63 (7), 1213–1243.

Juang, B.H., Rabiner, L.R., 1985. Mixture autoregresive hidden Markov models for speech signals. IEEE Trans. Acoust. Speech Signal Process ASSP-33 (6), 1404–1413.

Lee, B.G., Lee, K.Y., Ann, S., 1995. An EM-based approach for parameter enhancement with an application to speech signals. Signal Process. 46, 1–14.

Lee, C.H., 1997. On feature and model compensation approach to robust speech recognition. In: Robust Speech Recognition for Unknown Communication Channels, pp. 45–54.

Lee, K.Y., Lee, B., Song, I., Yoo, J., 1996. Recursive speech enhancement using the EM algorithm with initial conditions trained by HMMs. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal, Processing, pp. 621–624.

Logan, B.T., 1998. Adaptive model-based speech enhancement, Ph.D. thesis, University of Cambridge, Available at ftp://svr-ftp.eng.cam.ac.uk/pub/reports/logan_thesis.ps.Z.

Logan, B.T., Robinson, A.J., 1996. Noise estimation for enhancement and recognition within an autoregressive hidden Markov model framework. In: Proceedings of the Sixth Australian International Conference on Speech Science and Technology, pp. 85–90.

Logan, B.T., Robinson, A.J., 1997a. Enhancement and recognition of noisy speech within an autoregressive hidden Markov model framework using noise estimates from the noisy signal. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 843–846.

Logan, B.T., Robinson, A.J., 1997b. Improving autoregressive hidden Markov model recognition accuracy using a non-linear frequency scale with application to speech enhancement. In: Proceedings of the Fifth European Conference on Speech Communication and Technology, pp. 2103–2106.

Logan, B.T., Robinson, A.J., 1998. A practical perceptual frequency autoregressive HMM enhancement scheme. In: Proceedings of the International Conference on Spoken Language Processing.

McKinley, B., Whipple, G., 1997. Model based speech pause detection. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1179–1182.

Merhav, N., Ephraim, Y., 1991. Maximum likelihood hidden markov modelling using a dominant sequence of states. IEEE Trans. Signal Process. 39, 2111–2115.

Mokbel, C., 1997. MUSE: Multipath stochastic equalization a theoretical framework to combine equalization and stochastic modeling. In: Proceedings of the ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels.

Mokbel, C.E., Chollet, G., F.A, 1995. Automatic word recognition in cars. IEEE Trans. Speech Audio Process., 346–356.

Moreno, P.J., Raj, B., Stern, R.M., 1995. A vector Taylor series approach for environment-independent speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 733–736.

Oppenheim, A.V., Johnson, D.H., 1972. Discrete representation of signals. Proc. IEEE 60 (6), 681–691.

Price, P., Fisher, W.M., Bernstein, J., Pallett, D.S., 1988. The DARPA 1000-word Resource Management database for continuous speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 651–654.

Rabiner, L.R., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, New York.

Rose, R.C., Hofstetter, E.M., Reynolds, D.A., 1994. Integrated models of signal and background with application to speaker identification. IEEE Trans. Speech Audio Process. 2 (2), 245–257.

Sankar, A., Lee, C.H., 1996. A maximum-likelihood approach to stochastic matching for robust speech recognition. IEEE Trans. Speech Audio Process. 4 (3), 190–202.

Seymour, C.W., 1996. Model-based speech enhancement, Ph.D. thesis, University of Cambridge.

Sheikhzadeh, H., Sameti, H., Deng, L., Brennan, R.L., 1994. Comparative performance of spectral subtraction and HMM-based speech enhancement with application to hearing aid design. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I13–I16.

Sheikhzeheh, H., Brennan, R., Sameti, H., 1995. Real-time implementation of HMM-based MMSE algorithm for speech enhancement in hearing aid applications. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 808–811.

Shikano, K., 1985. Evaluation of LPC spectral matching measures for phonetic unit recognition, Technical Report CMU-CS-86-108, Carnegie-Mellon University.

Strube, H.W., 1980. Linear prediction on a warped frequency scale. J. Acoust. Soc. Amer. 68 (4), 1071–1076.

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, Technical report, DRA Speech Research Unit.

Young, S.J., Woodland, P.C., Byrne, W.J., 1993. HTK: Hidden Markov Model Toolkit V1.5, Cambridge University Engineering Department Speech Group and Entropic Research Laboratories Inc.

Young, S., Jansen, J., Ollason, D., Woodland, P., 1996. The HTK book for HTK V2.0, Cambridge University Technical Services Ltd. and Entropic Cambridge Research Laboratory Ltd.

UNCORRECTED PROOF