PREDICTIVE MODEL-BASED COMPENSATION SCHEMES FOR ROBUST SPEECH RECOGNITION

M.J.F. Gales[1]

September 1997 (revised February 1998)

Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
England

Email:mjfg@watson.ibm.com

Abstract

For practical applications speech recognition systems need to be insensitive to differences between training and test acoustic conditions. Differences in the acoustic environment may result from various sources, such as ambient background noise, channel variations and speaker stress. These differences can dramatically degrade the performance of a speech recognition system. A wide range of techniques have been proposed for achieving noise robustness. This paper considers one particular approach to model-based compensation, *predictive* model-based compensation, which has been shown to achieve good noise robustness in a wide range of acoustic environments. The characteristic of these schemes is that they combine a speech model with an additive noise model, a channel model and, in the general case, a speaker stress model, to generate a corrupted-speech model. The general theory of these predictive techniques is discussed. Various approximations for rapidly performing the model combination stage have been proposed and are reviewed in this paper. The advantages and the limitations of such a predictive approach to noise robustness are also discussed. In addition, methods for combining predictive schemes with schemes which make use of speech data in the new environment, *adaptive* schemes, are detailed. This combined approach overcomes some of the limitations of the predictive schemes.

[1]The author is now at the IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA.

1 Introduction

The requirement for robustness in speech recognition systems is becoming increasingly important as speech technology is applied to real world applications. Dramatic degradations in system performance can occur as a result of differences between the training and application conditions. Differences may arise from many sources, such as ambient background noise, channel differences and speaker stress. A wide variety of schemes have been proposed to deal with this problem: multiple microphones; inherently robust frontends; enhancement schemes; and model-based compensation schemes. Many of these approaches are reviewed in [Gong, 1995]. This paper considers in detail one particular approach to model-based compensation that has been found to give useful noise-robustness, *predictive model-based compensation*[2]. In this approach the speech models are compensated without the need for any speech data in the new acoustic environment. The schemes, such as parallel model combination (PMC) and vector Taylor series (VTS), only require models of the interfering noise sources to compensate the model parameters. The aim of this paper is to examine some of the predictive compensation schemes described in the literature, the strengths and weaknesses of such an approach, and finally methods which may be used to address some of these weaknesses.

The basic philosophy behind predictive techniques is that optimal performance is obtained when the speech models are trained on data recorded in the same acoustic environment as found in recognition. For many scenarios the original recognition system is trained in a clean acoustic environment. In other words, the training data were collected at a high signal-to-noise ratio using a good quality microphone. It is impossible to ensure that the recogniser is used in the same clean environment, or even that the same microphone or transmission channel is used. Thus the aim of predictive techniques is to convert the *clean-speech model set* to one that approximates the ideal matched train/recognise model set.

The simplest way to create the desired *corrupted-speech model set* is to modify the training data to be representative of the test environment using samples of the interfering noise in the new environment. Then the models can be retrained on this modified data. For large-vocabulary speech-recognition tasks, where there may be many hours of training data and highly complex acoustic models, this is usually impractical. To avoid this, predictive compensation schemes make the assumption that the clean-speech models are a good representation of the clean-speech training data. Instead of modifying the training data and retraining the models, they directly compensate the clean-speech model parameters for the presence of noise. Compensation is performed by combining the clean-speech models with noise models using a *mismatch function* that describes the effect of the train/recognise environment differences on the clean-speech features. A problem with this approach is that for the majority of scenarios in speech recognition there are no simple solutions to finding the corrupted speech model distributions. This has led to the development of a variety of approximate solutions to enable rapid compensation.

The predictive techniques can be contrasted with other forms of noise compensation techniques. One currently popular set of such techniques may be referred to as *adaptive* model-based compensation. These schemes take a model set (again generally the clean-speech model set) and use speech data observed in the new acoustic environment to modify it to be representative of this new environment. Such compensation schemes include maximum *a posteriori* (MAP) adaptation [Gauvain and Lee, 1994] and maximum likelihood linear regression (MLLR) [Leggetter and Woodland, 1995]. These schemes typically require thousands of frames of adaptation data to obtain good estimates of the new model sets. Furthermore, they require a transcription of the speech data. When they are used in an unsupervised adaptation mode, that is the transcription is not given so a recognition hypothesis must be used, the hypothesised transcription may be poor. This limits the possible performance gains.

Alternatively inherently noise robust parameterisations may be used. These include schemes such as RASTA [Hermansky and Morgan, 1994, Morgan and Hermansky, 1992] and the commonly used cepstral mean normalisation (CMN). Although usually effective at relatively high signal-to-noise ratios, the performance of these schemes degrades rapidly as the noise level increases, partic-

[2] This is the terminology used in [Bellegarda, 1997].

ularly for large-vocabulary systems. A third class of noise compensation techniques is speech enhancement. These schemes range from simple spectral subtraction [Boll, 1979] to complex model-based enhancement [Ephraim, 1992]. Enhancement schemes are popular since they are often simple to implement and are computationally cheap. However, the gains of these enhancement schemes are usually limited, due to the simple nature of the speech models used. These speech models are important when noise completely masks the clean speech. The enhancement schemes must still estimate the clean-speech data with only noise observed. Complex speech models have been used for enhancement but predictive compensation techniques are normally required as part of the enhancement process (see section 4.2).

This paper first describes the general acoustic environment and the modelling assumptions behind it. Section 3 presents a general predictive framework. A series of *mismatch functions* for a few standard acoustic environments are given. Using these functions the effects of additive noise on the speech distribution are investigated. Various predictive schemes and their underlying approximations are then described in sections 4 to 6. How *complicated noise* environments are handled is described in section 7, followed by a brief discussion of the advantages and limitations of predictive schemes. Methods to overcome some of the latter are then presented, such as single-pass retraining (section 9) or the combination of predictive schemes with adaptive model-based schemes (section 10). Finally, some experimental results are given in section 11 to illustrate the strengths and weaknesses of predictive compensation schemes.

2 Acoustic Environment

The first step in any environment compensation scheme is to determine the effects of the acoustic environment on the clean-speech signal. One model of the acoustic environment is shown in figure 1 [Hansen and Arslan, 1995].

Figure 1 goes here

Expressing the sources in figure 1 separately yields the following equation

$$y(t) = \left[\left\{\left(\left[s(t)|^{Stress}_{Lombard}\right]_{n_1(t)} + n_1(t)\right) * h_{mike}(t) + n_2(t)\right\} * h_{chan}(t)\right] + n_3(t) \quad (1)$$

where $y(t)$ is the corrupted-speech signal, $s(t)$ is the clean-speech signal, $n_1(t)$ is the ambient (background) noise, $h_{mike}(t)$ is the impulse response of the microphone, $n_2(t)$ and $h_{chan}(t)$ are the additive noise and impulse response, respectively, of the transmission channel, and $n_3(t)$ is the noise present at the receiver. It is hard to generate a specific stress model dependent on the current noise so the effects of stress are usually ignored. All the additive noise sources can then be combined into one variable, $n(t)$, as can all the convolutional noise sources, $h(t)$. This yields the standard, often stated, expression

$$y(t) = s(t) * h(t) + n(t) \quad (2)$$

Normally the data are not modelled in the time domain, with the obvious exception of auto-regressive HMMs. It is therefore necessary to determine the effects of the noise on the speech in the domain in which the speech is to be modelled. Separating the signal into frames, taking the discrete Fourier transform, and then the power, $y_f(\tau)y_f^*(\tau)$, (the subscript f is used to indicate the variable after the Fourier transform) yields

$$y_f(\tau)y_f^*(\tau) \quad = \quad |h_f(\tau)s_f(\tau)|^2 + |n_f(\tau)|^2 + 2n_f(\tau)h_f(\tau)s_f(\tau)\cos(\theta) \quad (3)$$

where θ is the angle between the speech vector and the noise vector. The following assumptions are commonly made.

1. The signal is smoothed over sufficient bins that the last term may be ignored.
2. $h_f(\tau)$ is constant over time, independent of τ.
3. $h_f(\tau)$ is independent of the incoming signal level.

The last two assumptions are generally poor, but greatly simplify the resulting expression. With these assumptions the following relationship is obtained in the linear-spectral domain[3]

$$\mathbf{o}(\tau) = \mathbf{h}\mathbf{s}(\tau) + \mathbf{n}(\tau) \tag{4}$$

where $\mathbf{h}$ is the convolutional noise term in the linear-spectral domain, $\mathbf{o}(\tau)$ is the corrupted speech, $\mathbf{n}(\tau)$ the additive noise and $\mathbf{s}(\tau)$ the speech at frame τ in the linear-spectral domain. Other domains will be considered in this paper based on this expression. Throughout this paper a consistent notation will be used to denote the domain of the model parameters or observation. Superscripts will be used to denote the domain, for example for the mean parameter, μ is the clean-speech linear-spectral mean, μ^l is the clean-speech log-spectral mean, and μ^c is the clean-speech cepstral mean. For this paper, unless otherwise stated "cepstral" refers to cepstra obtained from log-spectral filterbank analysis, not linear prediction. The noise parameters will be denoted as, for example, $\tilde{\mu}^c$, and the corrupted-speech parameters as $\hat{\mu}^c$. The clean speech, noise, and corrupted speech in the cepstral domain will thus be indicated as $\mathbf{s}^c(\tau)$, $\mathbf{n}^c(\tau)$ and $\mathbf{o}^c(\tau)$ respectively. Where distributions are used to approximate the observations the indexing on τ is dropped.

Thus, by making some simplifying assumptions about the acoustic environment and its effects on the speaker, it is possible to derive a simple equation in the linear-spectral domain describing the effects of the acoustic environment on the clean-speech data.

3 A Predictive Framework

Figure 2 goes here

The general framework for predictive compensation schemes is very simple. A predictive model-based compensation scheme has to generate estimates of corrupted-speech models in the new acoustic environment without having observed any speech data from the new environment. This is achieved using models of the various noise sources and a function that describes how each of these noise sources alters the speech signal. Figure 2 shows the general framework for this process. Thus the corrupted-speech model is a combination of the clean-speech model with an additive-noise model, channel model, and stress model. The latter is to model the effect of stress and noise on the speech itself, the Lombard effect [Junqua and Anglade, 1990]. This effect is highly complex and will not be explicitly considered in this paper (although section 10 shows how simple estimates may be made of this model).

To perform the compensation requires a function describing the effects of channel differences and noise on the speech for all elements of the particular speech feature vector used. These functions are known as the *mismatch functions*. In addition to their use in compensation, the functions can also be used to examine the nature of the corrupted-speech distribution. Knowledge of the form of this distribution is important as it influences the choice of the appropriate distribution to use in the corrupted-speech model. The mismatch functions, effect of noise on the corrupted-speech distribution and choice of corrupted-speech model are discussed in detail below.

[3]In this paper, the linear-spectral domain refers to the feature domain that exists after applying the mel-spaced bins to the values in equation 3. The log-spectral domain refers to the feature domain that exists after taking the log() of the linear-spectral domain.

3.1 Mismatch Functions

To combine the various models in a sensible way the effect that the noise has on the elements on the feature vector used to represent the speech data must be described. This is done by defining a mismatch function. Since their effects vary, each noise condition requires a different mismatch function. Four noise conditions are considered here. All the functions described assume that the clean-speech and corrupted-speech data are modelled using features derived from mel-frequency cepstral coefficients (MFCCs) [Davis and Mermelstein, 1980]. The functions come directly, or are derived, from the expressions for the effect of the acoustic environment on the clean-speech signal presented in the previous section.

1. **Additive Noise**: The cepstral domain is a linear transformation of the log-spectral domain.

   $$\mathbf{o}^c(\tau) = \mathbf{C}\mathbf{o}^l(\tau) \tag{5}$$

   where for the static features (from equation 4),

   $$o_i^l(\tau) = \log\left(\exp(s_i^l(\tau)) + \exp(n_i^l(\tau))\right) \tag{6}$$

   and the elements of the cosine transform matrix $\mathbf{C}$ are defined as

   $$c_{ij} = \cos(i(j-0.5)\pi/N) \tag{7}$$

   where N is the number of log-spectral filterbanks. In addition to static features, dynamic features are commonly appended to the feature vector [Furui, 1986]. Mismatch functions for these dynamic features are described in appendix A.

2. **Convolutional Noise**: If there is just convolutional noise, then

   $$\mathbf{o}^c(\tau) = \mathbf{s}^c(\tau) + \mathbf{h}^c \tag{8}$$

   If the speech and the noise are both modelled by Gaussian components, the resultant corrupted-speech distribution will also be Gaussian distributed.

3. **Additive and Convolutional Noise**: In many situations both additive and convolutional noise are present. For static features

   $$o_i^l(\tau) = \log\left(\exp(s_i^l(\tau) + h_i^l) + \exp(n_i^l(\tau))\right) \tag{9}$$

   models the effects of additive and convolutional noise.

4. **Bandwidth Limited Channel**: The situation where data are bandwidth limited compared to the training data cannot easily be handled in the simple convolutional noise framework. An alternative mismatch function is therefore used.

   $$\mathbf{o}^{bc}(\tau) = \tilde{\mathbf{W}}\mathbf{o}^c(\tau) \tag{10}$$

   where $\tilde{\mathbf{W}}$ is defined in appendix B and $\mathbf{o}^{bc}(\tau)$ is the corrupted-speech observation after the bandwidth-limited channel. This function may be used in conjunction with any of the other mismatch functions.

The rest of this paper concentrates on the cases where there is additive noise or both additive and convolutional noise present.

3.2 Effects of Noise on Speech Distributions

There are two options for investigating the effects of noise on the speech distributions. The first is to use actual corrupted-speech data from a stereo database and examine the corrupted-speech distributions (such as in [Openshaw and Mason, 1994]). Alternatively, and the approach adopted here, the effects of idealised noise on idealised speech can be examined. The idealised speech is modelled by a Gaussian distributed in the log-spectral domain. The noise is also assumed to be Gaussian distributed in this domain. This allows a very controlled investigation of the effects of noise. It is also a good (although simplified) representation of combining a speech model and noise model using a predictive model-based compensation scheme. For plots in this paper the speech has a mean of 10 and a standard deviation of 6. The noise has a mean of 4 and a standard deviation of 1.

Figure 3 goes here

For the simplified idealised scenario with a single dimension in the log-spectral domain, the corrupted-speech distribution when additive noise is present has the form shown in figure 3. The following general, often stated, observations may be made about the corrupted-speech distribution.

1. The distribution is non-Gaussian.
2. The mean increases.
3. The variance decreases[4].

The effects of the convolutional noise on the distribution are not plotted here since, using the simplified acoustic environment, convolutional noise just shifts the position of the clean-speech mean prior to adding the noise signal.

Of course the actual corrupted-speech distribution will be far more complex than this. The elements of the speech feature vector are correlated with one another and not Gaussian distributed.

3.3 Corrupted-speech Distribution Modelling

In predictive model-based compensation, each clean-speech model state(-component) is combined with each noise state(-component). Two basic forms may be used to model the corrupted-speech distributions that result from each speech and noise state(-component) pairing.

1. **Single Gaussian distribution**: Each speech and noise state(-component) pairing is modelled by a single Gaussian component. Depending on the nature of the noise, and the domain in which the speech is being modelled, this may be a crude representation of the corrupted-speech distribution.
2. **Multiple and non-Gaussian distributions**: To improve the accuracy of the corrupted-speech models, distributions other than single Gaussian distributions have been considered. The nature of the distributions used varies from the use of multiple Gaussian distributions to thresholding schemes which use different models of the corrupted speech depending on the observed signal level.

Various forms of these two approaches are described in sections 4 and 5.

[4]This assumes that the noise variance is smaller than the speech variance, which is commonly true.

4 Single Gaussian Schemes

The simplest approach to modelling the corrupted-speech distribution is to use a single Gaussian component to model each speech and noise state(-component) pairing. Although this is only a crude approximation of the corrupted-speech distribution shown in figure 3, it results in corrupted-speech models with distributions of the same form as the clean-speech models. This allows standard recognisers to be used with no modifications.

Using a single Gaussian component, the likelihood of the corrupted-speech observation being produced by the corrupted-speech models is

$$\mathcal{L}(\mathbf{o}^c(\tau)|q_j(\tau), q_v^n(\tau), \mathcal{M}_s, \mathcal{M}_n) \approx \mathcal{N}(\mathbf{o}^c(\tau); \hat{\mu}^c, \hat{\mathbf{\Sigma}}^c) \tag{11}$$

where $\mathcal{N}()$ is the Gaussian probability density function, $q_j(\tau)$ and $q_v^n(\tau)$ denote occupation of speech state j and noise state v at time τ, $\mathcal{M}_s$ is the clean-speech model and $\mathcal{M}_n$ the noise model. Compensation schemes are required to rapidly estimate the new means, $\hat{\mu}^c$, and, preferably, the new variances, $\hat{\mathbf{\Sigma}}^c$. The aim of the schemes is to find

$$\hat{\mu}^c = \mathcal{E}\left\{\mathbf{o}^c\right\} \tag{12}$$

and

$$\hat{\mathbf{\Sigma}}^c = \mathcal{E}\left\{(\mathbf{o}^c - \hat{\mu}^c)(\mathbf{o}^c - \hat{\mu}^c)^T\right\} \tag{13}$$

where the expectations are taken over the clean-speech and noise distributions.

4.1 Parallel Model Combination

In parallel model combination (PMC)[5] many different approximations have been used to obtain the model parameters. In its basic form PMC just deals with additive noise. Thus for the additive noise case

$$\hat{\mu}_i^l = \mathcal{E}\left\{\log\left(\exp(s_i^l) + \exp(n_i^l)\right)\right\} \tag{14}$$

The statistics for s_i^l and n_i^l are obtained from the appropriate speech and noise state(-component) pair. A similar form may be obtained for the variance (or covariance matrix).

Unfortunately equation 14 has no simple closed-form solution when Gaussian distributions are used to model the clean speech and the noise in the log-spectral domain (or cepstral domain). Various approximations have therefore been used. All the expressions in this section are given in the log-spectral domain, whereas the model parameters are assumed to be in the cepstral domain. A linear transform maps model parameters from the log-spectral to the cepstral domain.

$$\begin{aligned} \hat{\mu}^c &= \mathbf{C}\hat{\mu}^l & (15) \\ \hat{\mathbf{\Sigma}}^c &= \mathbf{C}\hat{\mathbf{\Sigma}}^l\mathbf{C}^T & (16) \end{aligned}$$

$\mathbf{C}$ is defined in equation 7. Mapping from the cepstral to the log-spectral domain is simply the reverse of the above[6].

1. **Log-Normal Approximation**: This was the original approximation used [Gales and Young, 1992]. It is based on the assumption that the sum of two log-normally distributed variables is itself approximately log-normally distributed. Using this approximation it is only necessary to estimate the mean and variance of the corrupted-speech data in the linear-spectral domain

[5] Some authors refer to all predictive compensation schemes as PMC.

[6] In many systems only the lower-order cepstral parameters are used. Hence it is not possible to truly map the model parameters from the cepstral to log-spectral domain. However this has been found to have little effect on the performance [Gales, 1995].

to determine the corrupted-speech distribution in the cepstral domain. In the linear-spectral domain the speech and noise are additive so

$$\hat{\mu} = \mu + \tilde{\mu} \tag{17}$$
$$\hat{\mathbf{\Sigma}} = \mathbf{\Sigma} + \tilde{\mathbf{\Sigma}} \tag{18}$$

where the clean-speech linear-spectral domain parameters are related to the log-spectral parameters by

$$\mu_i = \exp(\mu_i^l + \Sigma_{ii}^l/2) \tag{19}$$
$$\mathbf{\Sigma}_{ij} = \mu_i\mu_j \left[\exp(\Sigma_{ij}^l) - 1\right] \tag{20}$$

The same expressions can be used to obtain the noise statistics in the linear-spectral domain. Using the inverses of the above equations give the corrupted-speech mean and variance in the log-spectral domain.

$$\hat{\mu}_i^l \approx \log(\hat{\mu}_i) - \frac{1}{2}\log\left(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i^2} + 1\right) \tag{21}$$
$$\hat{\Sigma}_{ij}^l \approx \log\left(\frac{\hat{\Sigma}_{ij}}{\hat{\mu}_i\hat{\mu}_j} + 1\right) \tag{22}$$

This approximation has also been used to compensate models based on linear prediction cepstral coefficients [Martin et al., 1993]. This approximation is usually only used to compensate the static model parameters, not the dynamic parameters (although a similar approach has been used to crudely compensate dynamic parameters).

Although algebraic expressions for the mean and variance are given, evaluating these expressions is computationally expensive, in particular transforming the variances from the cepstral to the log-spectral domain. This has a computational cost of $\mathcal{O}(n^3)$ operations to map the variances and $\mathcal{O}(n^2)$ to map the means (where n is the dimensionality of the static feature vector). The calculation of the mismatch function is only performed $\mathcal{O}(n^2)$ times to compensate the variances and $\mathcal{O}(n)$ times to compensate the means per speech and noise component pairing. For speech model sets with large numbers of Gaussian components this may be prohibitively expensive.

2. **Log-Add Approximation**: By ignoring the effects of the variances on the estimates, the corrupted-speech mean is approximated by[7]

$$\hat{\mu}_i^l \approx \log\left(\exp(\mu_i^l) + \exp(\tilde{\mu}_i^l)\right) \tag{23}$$

This scheme is only applicable to compensating the means, not the variances, of the models. However, it is highly computationally efficient. As the effects of the variance are ignored there is no need to transform the variances between the various domains, a major overhead in the log-normal approximation. Furthermore the mean vector may be stored in the log-spectral domain, saving the cost of mapping the means parameters from the cepstral to the log-spectral domain, without dramatically increasing the memory requirements.

3. **Numerical Integration**: Direct numerical integration of the expressions may be used to obtain the model parameters. One simple and computationally efficient method (in the sense that few points are required) is to use Gaussian integration with Gauss-Hermite abscissa and weights [Gales, 1995]. If implemented directly solving equation 14 involves integrating over the speech distribution, then over the noise distribution . The process may be simplified using the equalities described in [Gales, 1995]. The static mismatch function may be rewritten as

$$\log\left(\exp(s_i^l) + \exp(n_i^l)\right) = s_i^l + \log\left(1 + \exp(n_i^l - s_i^l)\right) \tag{24}$$

[7] The log-add approximation for the static parameters in additive noise is very closely linked to hypothesised Wiener filtering (HWF) [Berstein and Shallom, 1991, Beattie and Young, 1992].

A new variable x_i^l is introduced, where $x_i^l = n_i^l - s_i^l$. If the speech and noise are modelled by Gaussian distributions x_i^l is also Gaussian distributed. Performing the numerical integration over x_i^l, allows the mean to be estimated using a single dimension integration.

Numerical integration is able to yield very accurate estimates of the model parameters. By increasing the number of points used in the numerical integration, these estimates become increasingly accurate. However, as the number of points increases, the computational cost increases dramatically, particularly if variances are compensated, or delta and delta-delta parameters are used. Furthermore, when compensating the variances, numerical integration schemes have the same cost of converting parameters between the cepstral and log-spectral domains as the log-normal approximation.

When the variances are updated full covariance matrices are obtained in the cepstral domain. This is even true when diagonal covariance matrices are used to model the speech and noise. Only the elements on the leading diagonal are commonly used in the corrupted-speech model, since using full covariance matrices increases both the memory requirements to store the models and the computational requirements to calculate likelihoods during recognition.

4.2 Auto-Regressive HMMs

An alternative scheme, that does not have the problem of the highly complex mismatch function when the data are modelled in the cepstral domain, is to use auto-regressive (AR) HMMs [Juang and Rabiner, 1985]. As the data are usually modelled in the time domain and the speech and noise are assumed to be additive in this domain, the mismatch function is the standard expression[8] given in equation 2. This is a far simpler mismatch function than that required for standard speech models in the cepstral domain. In [Ephraim, 1992] it is shown how to obtain a Gaussian (not AR) HMM from AR noise and speech models. AR HMMs are attractive because they provide good models for corrupted speech. However to date, they are commonly not used for speech recognition, as the recognition performance of these models is generally poor compared to that of standard HMMs.

AR HMMs have been successfully used to enhance the corrupted speech as a pre-cursor to recognition [Ephraim et al., 1989]. This type of enhancement scheme will be referred to as a predictive model-based enhancement scheme. The basic structure of a predictive model-based enhancement scheme is shown in figure 4. Its concept is that, by using the corrupted-speech model to recognise the corrupted-speech data, an hypothesised clean-speech model state sequence may be obtained from the corrupted speech[9]. Using this hypothesised state sequence, the clean-speech statistics and noise statistics for a particular frame are hypothesised. The clean-speech data can then be estimated by applying standard estimation schemes, such as minimum mean square error estimators and simple linear estimators.

Figure 4 goes here

Predictive model-based enhancement schemes have also been examined in domains other than the time domain using standard HMMs, for example the cepstral domain [Seymour and Niranjan, 1994] and the linear-spectral domain [Gagnon, 1992]. A major advantage of predictive model-based enhancement schemes is that the model set set used for enhancement need not be the same as that used for recognition. Thus, simpler ergodic models of speech may be used in the model compensation scheme, greatly reducing the computational cost of compensation.

[8] To date systems handling convolutional noise have not been designed.

[9] The corrupted-speech state sequence has a one-to-one mapping to the clean-speech state sequence.

4.3 Vector Taylor Series

Vector Taylor series approximations (VTS) [Moreno et al., 1996] are another alternative single Gaussian scheme. The corrupted-speech distribution is calculated in the log-spectral domain in a similar fashion to PMC. However, now the mismatch function is approximated by a finite length Taylor series and the expectation over this Taylor series is taken to find the corrupted-speech parameters. For example, the static mismatch function for additive and convolutional noise may be written as[10]

$$\begin{aligned}\mathbf{o}^l = \mathcal{F}(\mathbf{s}^l,\mathbf{h}^l,\mathbf{n}^l) &\approx \mathcal{F}(\mathbf{s}^l_0,\mathbf{h}^l_0,\mathbf{n}^l_0) + \frac{\partial}{\partial \mathbf{s}^l}\mathcal{F}(\mathbf{s}^l_0,\mathbf{h}^l_0,\mathbf{n}^l_0)(\mathbf{s}^l-\mathbf{s}^l_0) + \frac{\partial}{\partial \mathbf{h}^l}\mathcal{F}(\mathbf{s}^l_0,\mathbf{h}^l_0,\mathbf{n}^l_0)(\mathbf{h}^l-\mathbf{h}^l_0)\\ &\quad + \frac{\partial}{\partial \mathbf{n}^l}\mathcal{F}(\mathbf{s}^l_0,\mathbf{h}^l_0,\mathbf{n}^l_0)(\mathbf{n}^l-\mathbf{n}^l_0) + \ldots \end{aligned} \tag{25}$$

It is possible to simplify this expansion by using the same equalities as the PMC numerical integration case. Since all the variables are assumed to be Gaussian distributed and independent, it is possible to consider expansions based on $\mathbf{x}^l$, where $\mathbf{x}^l = \mathbf{n}^l - \mathbf{s}^l - \mathbf{h}^l$. The corrupted-speech mean is set to the expected value of the Taylor series which now becomes (considering only one dimension)

$$\begin{aligned}\hat{\mu}^l &= \mathcal{E}\left\{\mathcal{F}(s^l,h^l,n^l)\right\}\\ &= \mu^l + h^l + \mathcal{F}'(x^l_0) + \mathcal{E}\left\{\frac{\partial}{\partial x^l}\mathcal{F}'(x^l_0)(x^l-x^l_0)\right\} + \mathcal{E}\left\{\frac{\partial^2}{\partial {x^l}^2}\mathcal{F}'(x^l_0)(x^l-x^l_0)^2\right\} + \ldots \end{aligned} \tag{26}$$

where

$$\mathcal{F}'(x^l) = \log\left(1+\exp\left(x^l\right)\right) \tag{27}$$

Since x^l is Gaussian distributed the expected value of all the odd terms in the Taylor series will be zero (provided the expansion is based around the mean of x^l). The mean can then be approximated by (assuming that h^l is a fixed value)

$$\begin{aligned}\hat{\mu}^l = \mathcal{E}\left\{\mathcal{F}(s^l,h^l,n^l)\right\} &\approx \mu^l + h^l + \mathcal{F}'(x^l_0) + \mathcal{E}\left\{\frac{\partial^2}{\partial {x^l}^2}\mathcal{F}'(x^l_0)(x^l-x^l_0)^2\right\}\\ &= \mu^l + \mathcal{F}'(\tilde{\mu}^l-\mu^l-h^l) + (\Sigma^l+\tilde{\Sigma}^l)\frac{\partial^2}{\partial {x^l}^2}\mathcal{F}'(\tilde{\mu}^l-\mu^l-h^l)\end{aligned} \tag{28}$$

By increasing the number of terms in the expansion, the approximation may be made very accurate. The expression in equation 28 may be viewed as an extension of the PMC log-add approximation. VTS may be applied to update the variances as well. Unfortunately, unless the elements of the covariance matrix are assumed to be independent in the log-spectral domain, the Taylor series is more complex so it is more expensive to compute. Again similar equalities as those used in numerical integration [Gales, 1995] may be used to simplify the VTS estimate of the variance. However, the variances must also be mapped from the cepstral domain to the log-spectral domain and vice-versa.

The computational cost of using VTS to compensate the models depends heavily on the number of terms taken in the Taylor series and the model parameters being compensated. If only static mean parameters are compensated the cost may be made relatively low[11]. However if variances, or dynamic parameters, are compensated the computational cost rises rapidly.

[10] The infinite length Taylor series yields an exact solution. The notation $\frac{\partial}{\partial \mathbf{s}^l}\mathcal{F}(\mathbf{s}^l_0,\mathbf{h}^l_0,\mathbf{n}^l_0)$ indicates the value of the derivative with respect to $\mathbf{s}^l$ at the point $(\mathbf{s}^l_0,\mathbf{h}^l_0,\mathbf{n}^l_0)$.

[11] To remove the need to transform the speech and noise model variances into the log-spectral domain, the diagonal elements of the speech and noise covariance matrices in the log-spectral domain can be stored.

5 Multiple and Non-Gaussian Schemes

The techniques in the previous section assume a single Gaussian component can be used to model each corrupted-speech distribution. In many situations this is a poor assumption because the corrupted-speech distribution obtained from combining the speech and noise models is not Gaussian distributed. The actual corrupted-speech distribution (for the additive noise case) may be expressed as

$$\mathcal{L}(\mathbf{o}^c(\tau)|q_j(\tau), q_v^n(\tau), \mathcal{M}_s, \mathcal{M}_n) = \int \mathcal{L}(\mathbf{s}^c(\tau), \mathbf{n}^c(\tau)|q_j(\tau), q_v^n(\tau), \mathcal{M}_s, \mathcal{M}_n) \tag{29}$$

where the integral is over all pairs of observations such that $\mathbf{o}^c(\tau) = \mathbf{s}^c(\tau) \otimes \mathbf{n}^c(\tau)$. Unfortunately it is only possible to obtain closed-form solutions for a few situations [Rose et al., 1994]. By relaxing the single Gaussian component restriction, alternative approximate solutions to this problem can be found. This section presents a number of schemes which use either multiple Gaussian or non-Gaussian distributions to model each speech and noise state(-component) pairing. The schemes presented are; noise masking, speech and noise decomposition, and iterative PMC.

5.1 Noise Masking

A variety of noise masking schemes have been implemented for robust speech recognition [Varga et al., 1988]. The basic assumption behind noise masking is that either the speech or noise dominates in the log-spectral domain. Hence, according to the relationship between the speech and noise a different approximation for the distribution is assumed.

Table 1 goes here

A typical noise masking scheme [Klatt, 1979] calculates the probability distribution according to one of the entries in table 1. In this scheme, if either the speech model mean or the observation is below the noise mean then they are assumed to be masked by the noise. The noise mean is then used in place of the masked speech mean and/or observation in the likelihood calculation. Other masking schemes are possible [Varga et al., 1988].

These noise masking schemes are simple to implement and computationally efficient. The only additional cost is during recognition. Thus, the computational cost of a noise masking scheme is dependent on the number of observations in the acoustic environment, not the number of Gaussian distributions in the speech model. It is therefore efficient for very large speech recognition systems. Two comparisons per dimension per Gaussian component per observation are required in addition to a standard Gaussian likelihood calculation. However, these schemes are limited in where they can be applied. As each channel of the filterbank must be dealt with separately, this scheme is only applicable in the log-spectral domain and to components in this domain with diagonal covariance matrices. This limits the form of speech models that may be used. Furthermore, masking schemes are not easily extended to handle dynamic features.

5.2 Speech and Noise Decomposition

Speech and noise decomposition (SND) is a natural extension of the noise masking scheme described above. Rather than modelling the noise as a simple mean level, it is modelled as a Gaussian distribution. In order to obtain the likelihood of the corrupted-speech observation a max() approximation between the speech and noise is used [Varga and Moore, 1990]. Thus it is assumed that

$$o_i^l(\tau) \approx \max(s_i^l(\tau), n_i^l(\tau)) \tag{30}$$

Under this assumption a closed-form expression for the likelihood is possible. Two situations may occur for each channel of each frame. Either the speech dominates, in which case the noise is known to be less than the observed signal, or the noise dominates and the speech is less than the observed signal. Mathematically this is expressed as

$$\begin{aligned}\mathcal{L}(o_i^l(\tau)|q_j(\tau), q_v^n(\tau), \mathcal{M}_s, \mathcal{M}_n) \;\approx\; & \mathcal{N}(o_i^l(\tau); \mu_i^l, \Sigma_{ii}^l)\mathcal{C}(o_i^l(\tau); \tilde{\mu}_i^l, \tilde{\Sigma}_{ii}^l) \\ & + \mathcal{C}(o_i^l(\tau); \mu_i^l, \Sigma_{ii}^l)\mathcal{N}(o_i^l(\tau); \tilde{\mu}_i^l, \tilde{\Sigma}_{ii}^l) \qquad (31)\end{aligned}$$

where $\mathcal{C}()$ is the cumulative Gaussian distribution. The likelihood calculation is now a weighted sum of the likelihood of the observation given the speech model and the likelihood given the noise model.

Figure 5 goes here

Figure 5 shows the distribution obtained when this approach is applied to the idealised scenario shown in figure 3. The distribution resulting from SND is relatively close to the idealised corrupted-speech distribution. Furthermore, SND is computationally efficient. As the speech models are not altered the computational load during recognition is a function of the number of frames decoded, not the number of speech model distributions. This makes SND efficient for very large model sets. The additional computational overhead for each likelihood calculation is one Gaussian distribution likelihood and a cumulative distribution lookup. This is compared to the standard single Gaussian likelihood normally required. However many of the limitations of the noise masking scheme also apply to SND. Again, the speech models are restricted to be diagonal covariance in the log-spectral domain. Also the scheme cannot be easily applied to systems with dynamic features.

5.3 Iterative PMC

A standard technique used in speech recognition (and more generally in pattern recognition) is to use multiple Gaussian distributions to model non-Gaussian distributions. Using a sufficiently large number of Gaussian distributions, any distribution may be modelled. This is the concept behind iterative PMC (IPMC). Many Gaussians are used to model each speech and noise state(-component) pairing. The likelihood of an observation is now given by

$$\mathcal{L}(\mathbf{o}^c(\tau)|q_j(\tau), q_v^n(\tau), \mathcal{M}_s, \mathcal{M}_n) \approx \sum_i \hat{w}_i \mathcal{N}(\mathbf{o}^c(\tau); \hat{\mu}^{(i)c}, \hat{\mathbf{\Sigma}}^{(i)c}) \qquad (32)$$

where i is summed over all Gaussian components associated with a combined speech and noise component and $\hat{w}_i$ is the component weight associated with component i of the corrupted-speech model.

Figure 6 goes here

Figure 6 shows IPMC applied to the idealised scenario shown in figure 3. Here the corrupted-speech distribution is modelled using Gaussian components. Although the corrupted-speech distribution is not perfectly modelled it is significantly better than the single Gaussian approximation. The problem is how to obtain the estimates of the parameters of the corrupted-speech model. An approach called data-driven PMC (DPMC) [Gales, 1995] may be used. Here the speech and noise models are used to generate corrupted-speech "observations". These observations are generated by using the speech and noise distributions to generate values, and then combining these values using the appropriate mismatch function. This is now a standard multiple component training problem[12].

Although gains in performance have been obtained for both standard and complicated noise environments [Gales, 1995] (see section 7), these models are computationally very expensive to train. In order to generate an accurate representation of the corrupted-speech data many observations are required (anywhere in the range 25-1000 [Gales, 1995]). The generation of each observation has the same computational cost as the PMC log-add approximation. In addition, iterative schemes must be used to estimate the multiple components once the corrupted-speech observations have been generated.

6 Summary of Techniques

Table 2 goes here

The techniques presented in the previous sections are summarised in table 2. The table gives; the technique, the final form of the corrupted-speech model for each speech and noise pairing, the domain it is normally applied in, the features that may be compensated, the model parameters that may be compensated, and the computational cost of the technique[13]. AR HMM compensation is not included in the table since it is not usually used for speech recognition.

Table 3 goes here

Table 3 lists the abbreviations used in table 2. The ordering of techniques in table 2 is not meant to indicate any measure of how good the techniques are. The best choice of technique for a particular task will be dependent on the signal-to-noise ratio, the complexity of the task and the compute power available.

7 "Complicated" Noise Environments

So far single state noise models have been assumed. This section addresses how to handle the situation where there are multiple states in the additive noise model.

Some noise environments have distinct regions associated with them. An extreme example is that of machine gun noise where there are distinct "bang" and "silence" regions of the acoustic signal. Modelling both regions with a single-state single-Gaussian-component noise model is a poor representation of the signal. Multiple state noise models allow a far better representation of the signal. Noise environments which require the use of these multiple-state noise models will be referred to as *complicated noise* environments. All the predictive techniques described may be used in these complicated noise environments. However, either the decoding stage must now be modified or the topology of the corrupted-speech model changed to accommodate the multiple-state noise model. These modifications are discussed in this section.

[12] This scheme may, of course, also be used to estimate the simpler single Gaussian models.

[13] For many of the schemes this is highly dependent on the model parameters being compensated and the number of terms, or points, used in estimating the parameters. For example if VTS is used to compensate only the static mean parameters it has a low computational cost. However when used to compensate all the parameters the cost is high. The computational cost indicated in table 2 is to compensate all possible parameters for all possible features.

7.1 3-Dimensional Viterbi Decoder

Figure 7 goes here

As it is not possible to know *a priori* which state of the noise model and which state of the speech model is most likely at recognition time, the concept of a 3-dimensional Viterbi decoder [Varga and Moore, 1990] must be used. Now all possible paths through both the speech model and the noise model are considered. These may be found in an efficient fashion very similar to standard Viterbi decoding. A new variable $\phi_{j,v}(\tau)$ is introduced which is the maximum joint probability of being in state q_j of the speech model and state q_v^n of the noise model at time τ and observing the sequence $\mathbf{o}^c(1)$ to $\mathbf{o}^c(\tau)$. $\phi_{j,v}(\tau)$ is defined as

$$\phi_{j,v}(\tau) = \max_{\theta \in \mathbf{\Theta}_{\tau-1}} [\mathcal{L}(\mathbf{o}^c(\tau), q_j(\tau), q_v^n(\tau), \theta | \mathcal{M}_s, \mathcal{M}_n)] \tag{33}$$

where $\mathbf{\Theta}_{\tau-1}$ is the set of all partial paths of length $\tau-1$, and θ is the path through both the speech and the noise model. The following recursion is used in a time synchronous fashion to extend all the currently active paths.

$$\phi_{j,v}(\tau) = \max_{i,u} \phi_{i,u}(\tau-1) a_{ij} \tilde{a}_{uv} \mathcal{L}(\mathbf{o}^c(\tau) | q_j(\tau), q_v^n(\tau), \mathcal{M}_s, \mathcal{M}_n) \tag{34}$$

This allows a multiple state noise model to be efficiently traversed.

7.2 Expanded Models

Figure 8 goes here

When the noise model is ergodic (i.e. fully connected), the 3-D decoding scheme described above can be synthesised using a standard 2-D decoder operating on an expanded model. Here each of the original speech states, q_i, is replaced by N compensated states (N is the number of noise states), $q_{i1}, q_{i2}, \ldots, q_{iN}$, with transition probabilities given by [Gales and Young, 1993, Martin et al., 1993]

$$\hat{a}_{iu,jv} = a_{ij} \tilde{a}_{uv} \tag{35}$$

This is shown diagrammatically in figure 8. There is a small loss of information at the model boundaries since the effective self-loop transition probabilities for the noise state cannot be preserved exactly. In practice this seems to make little difference.

Both the 3-D Viterbi decoder and the expanded model approach increase the number of possible paths through the corrupted-speech models. This increases the computational load during decoding, although the increase may be limited using standard pruning techniques.

8 Advantages and Limitations

In the previous sections various predictive model-based compensation schemes have been described. For each scheme some of its limitations were given. This section details some general advantages and disadvantages of predictive compensation techniques that apply to most of the schemes described.

8.1 Advantages

1. **Parameter compensation**. All the model parameters are individually compensated. Most other compensation schemes have to make assumptions about model parameters being affected in a similar fashion to enable rapid adaptation to a new acoustic environment. For example in MLLR many components share the same linear transform.

2. **Meaningful models**. The models generated by predictive compensation schemes may be directly related to the corrupted-speech models. In many instances, where single Gaussian component schemes are used, the "target" corrupted-speech model is the single-pass retrained model (see section 9). By comparing the two model sets the compensation algorithms may be examined at a far finer level than simple word error rate.

3. **Complicated noise environments**. These are handled in a simple, meaningful way (section 7). If more standard schemes, such as MLLR or MAP adaptation, are used to compensate the models there is no simple approach to handling these conditions.

4. **No speech data required**. No speech data in the new acoustic environment is required, only the various noise models. This allows models to be adapted to an acoustic environment as soon as these noise models are available, as opposed to, for example, having to wait for sufficient speech data to adapt the models as in an adaptive scheme.

5. **Non-Gaussian distributions**. Some schemes are able to explicitly model non-Gaussian corrupted-speech distributions. Most standard model-based compensation schemes implicitly assume that the corrupted-speech distributions are still Gaussian.

8.2 Limitations

1. **Mismatch function**. The performance is highly dependent on how appropriate the mismatch function is for describing the acoustic environment. If the approximations used to obtain the mismatch function are poor the compensated models will be poor.

2. **Noise models**. Explicit models of both the channel and additive noise are required[14]. For ambient additive noise these are relatively easy to obtain[15], however channel models, particularly convolutional noise, may be hard to determine *a priori*. Techniques for estimating these convolutional noise models are given in section 10. An important question is the amount of speech data required to generate a good noise model. For the predictive schemes described, the mean, and sometimes the variance, are required to describe the noise. The actual number of noise samples to obtain good estimates of these parameters is highly dependent on the acoustic environment. In particular, if a complicated noise environment is to be modelled, where the noise has some systematic temporal variation, significantly more noise data is required to ensure that all possible variations have been observed.

3. **Parameterisations**. It is not simple to define mismatch functions for all speech parameterisations, for example RASTA-PLP [Hermansky and Morgan, 1994] or cepstral mean normalisation (CMN). Without a mismatch function it is not possible to compensate the models.

4. **Speaker characteristics**. Predictive schemes only adapt the models to the new acoustic environment. Stress and speaker characteristics cannot be simply compensated for within this framework. In contrast adaptive schemes that make use of the corrupted-speech data can simultaneously compensate for both stress and speaker characteristics in addition to the acoustic environment.

[14] This is unfortunate - as a quotation attributed to Bishnu Atal says: "We call it noise because we know nothing about it!" [Furui and Sondhi, 1992].

[15] One further important assumption will be made. That is that the original speech signal was recorded in an acoustic environment with very high signal-to-noise ratio. Generating corrupted-speech models from speech models produced in noise corrupted environments is not considered in this paper.

5. **Accurate models**. All the models used in the compensation process are required to be accurate. Inaccurate models, or those which don't model all the required features of the speech and noise, will result in poor performance. This can become particularly important when compensating dynamic parameters (see appendix A).

6. **Speed**. The speed of any compensation scheme is very important. In non-stationary acoustic environments, the ability to adapt rapidly is particularly important as the model may have to be adapted many times. The computational cost of many model based compensation schemes is a function of the number of distinct components. (The obvious exceptions to this are the noise masking schemes and speech and noise decomposition.) For large vocabulary speech models the number of Gaussian components may be very large (> 100000), hence the compensation is computationally expensive. In contrast, for enhancement style schemes the cost is a function of the number of frames observed and for inherently robust schemes there is no cost! One particularly expensive aspect of these model-based compensation schemes, when cepstral parameters are used, is the mapping of the variances from the cepstral to and from the log-spectral domains[16]. For any compensation scheme making use of the variances in the log-spectral domain this is an unavoidable cost.

9 Single-Pass Retraining

Single-Pass Retraining (SPR) [Gales, 1995] addresses some of the problems described in the previous section. The clean-speech training data is modified to be representative of the data in the new acoustic environment. Thus, an artificial stereo database with pairs of clean-speech and corrupted-speech training data is generated. Given this stereo database, new acoustic models are trained. If standard HMM training is used, for example where K-means clustering is followed by repeated iterations of Baum-Welch re-estimation, the computational load associated with training the models is very high. SPR allows the models to be trained in a single pass of Baum-Welch re-estimation, thus reducing the computational load.

In SPR, the posterior Gaussian component probabilities (or frame/state-component alignment) obtained from aligning the clean-speech models with the clean-speech data are assumed to be approximately the same as the posterior probabilities that would be obtained from aligning the corrupted-speech data with corrupted-speech models. This allows all the components to be updated in a single pass. Thus the means and variances of the models are trained using

$$\hat{\mu}^c = \frac{\sum_{\tau=1}^{T} \gamma_m(\tau)\mathbf{o}^c(\tau)}{\sum_{\tau=1}^{T} \gamma_m(\tau)} \tag{36}$$

$$\hat{\mathbf{\Sigma}}^c = \frac{\sum_{\tau=1}^{T} \gamma_m(\tau)(\mathbf{o}^c(\tau) - \hat{\mu}^c)(\mathbf{o}^c(\tau) - \hat{\mu}^c)^T}{\sum_{\tau=1}^{T} \gamma_m(\tau)} \tag{37}$$

where the component occupation probability $\gamma_m \tau$ is defined as

$$\gamma_m(\tau) = p(q_m(\tau)|\mathbf{s}^c(1), \ldots, \mathbf{s}^c(T), \mathcal{M}_s) \tag{38}$$

and $q_m(\tau)$ indicates occupation of the component being re-estimated at time instance τ. The component weights and transitional probabilities are unaltered. If Viterbi or hard state alignments are used, instead of probabilistic ones, and are stored, then even for large amounts of training

[16] In contrast to the mean parameters, storage of covariance matrices in the log-spectral domain is normally too memory intensive unless diagonal covariance matrices are assumed in this domain. This assumption is usually poor, hence the use of cepstral parameters. Diagonal covariance matrices in the log-spectral domain are required for many of the mean compensation schemes. If only means are to be compensated using one of the schemes requiring variances, it is worth storing diagonal covariance matrices in the log-spectral domain, hence negating the requirement to map between domains.

data the computational cost may be quite small [Gauvain et al., 1996]. In [Gales, 1995] the performance of a SPR trained system is compared to one that is completely retrained by performing multiple iterations of Baum-Welch. Surprisingly, the SPR trained system performed as well as the completely retrained system.

There are two advantages to using SPR over standard predictive techniques. First, it is no longer necessary to assume that the speech models are a good representation of the speech data. Second, it is possible to use any speech parameterisation, since the speech and noise data can be combined in the time domain[17]. The disadvantages of SPR are that the complete speech database must be on-line and the computational cost of adding noise to the whole database is high.

If semi-continuous HMMs (SCHMMs) are used then it is not necessary to use all the training data due to the very large amount of tying [Moreno, 1996]. The relatively small number of distributions can be rapidly and accurately updated. However the large number of unaltered parameters, the component weights, tends to limit the performance of these schemes for SCHMMs. In a slightly modified version of SPR, MLLR [Leggetter and Woodland, 1995] is used in a supervised fashion on training data corrupted by the noise sources [Woodland et al., 1996]; this requires a much smaller amount of training data. Good performance was achieved using less than 1% (about 200 sentences) of the available training data.

In addition to being used for model compensation, SPR has been used to assess the various single Gaussian approximations [Gales, 1995]. For single Gaussian schemes, such as PMC, SPR yields the correct corrupted-speech model, ie the one generated if the speech and noise models were perfect representations of the acoustic noise and the speech data. Thus, SPR allows a far finer analysis of how the compensation schemes are performing. For example, it can show whether the delta-delta parameters are being compensated correctly using a particular approximation, without being affected by approximations in estimating the static or delta parameters.

10 Combined Schemes

In many ways predictive model-based compensation schemes are complementary to adaptive compensation schemes. Predictive schemes may be used to adapt all the parameters to approximately match the new acoustic environment. Then, for example, linear adaptation schemes may be used to fine-tune model parameters for particular speakers or to compensate for approximations in the mismatch function. Alternatively, slightly modified versions of adaptive schemes may be used to obtain channel characteristics to feed into the predictive model-compensation scheme. A variety of ways of combining adaptive and predictive schemes are possible. One fairly general combined approach to estimating the corrupted-speech model $\hat{\mathcal{M}}$ is

$$\hat{\mathcal{M}} = \mathcal{G}(\mathcal{K}(\mathcal{M}_s, \mathbf{o}_T) \otimes \mathcal{M}_n, \mathbf{o}_T) \tag{39}$$

where $\mathcal{G}()$ and $\mathcal{K}()$ are adaptive transforms and $\otimes$ indicates combining the models according to the predictive scheme used. The two transforms in equation 39 perform different tasks.

1. The function $\mathcal{K}()$ is used to model (to a certain extent) the convolutional noise, speaker stress and speaker characteristics. The estimation of the parameters of $\mathcal{K}()$ is non-trivial and is examined in the next section.

2. The function $\mathcal{G}()$ is used to correct inaccuracies in the mismatch functions and any errors in the estimation of the parameters of $\mathcal{K}()$. The parameters of the transform, normally linear, may be estimated using standard techniques.

The rest of this section only considers the case of linear transformations of the models. Other adaptive schemes, such as MAP schemes [Gauvain and Lee, 1994], may also be used. In this

[17] This assumes only additive noise is present.

section the ambient additive noise model, $\mathcal{M}_n$, is assumed to be known[18] and only schemes using Gaussian components to model the corrupted-speech distribution are considered.

10.1 Parameters of $\mathcal{K}(\mathcal{M}, \mathbf{o}_T)$

As previously mentioned, the function $\mathcal{K}(\mathcal{M}, \mathbf{o}_T)$ may be used to model channel differences, speaker stress and speaker characteristics. The effects of these are complex to model, so it would be preferable to have the most flexible transform possible. Unfortunately it is awkward to estimate the transform parameters in the presence of additive noise. For the case when there is just convolutional noise, the stochastic additive transform (SAT) [Sankar and Lee, 1995] and MLLR [Leggetter and Woodland, 1995] have been used. However, to date only simple transformations have been used when there is both additive and convolutional noise[19]. These simple transforms have the form

$$\hat{\mu}^c = \mu^c + \mathbf{h}^c \tag{40}$$

($\mathbf{h}^c$ is zero for the delta and delta-delta parameters). This assumes that the effects of the channel difference (and any stress related distortions of the original speech) may be modelled by a simple offset to the static parameters. Most systems assume a single bias for all components. It would be useful to have more general transforms such as MLLR acting on all the parameters including the delta (and delta-delta) parameters[20]. However to date, such general transforms have not been implemented. The rest of this section details various approaches to estimating the bias term in equation 40.

Some techniques for estimating $\mathbf{h}^c$ in equation 40 when there is additive and convolutional noise have been described in the literature. All the schemes require some speech data in the new acoustic environment to be available. The bias is usually estimated in a ML sense, so

$$\mathbf{h}^c = \arg\max_{\mathbf{h}^c} \left[\mathcal{L}(\mathbf{o}_T^c | \mathbf{h}^c, \mathcal{M}_s, \mathcal{M}_n)\right] \tag{41}$$

This may be rewritten as an expectation-maximisation (EM) optimisation problem [Minami and Furui, 1995]. This expression has been directly optimised using gradient descent, combining models using the PMC log-normal approximation [Minami and Furui, 1995]. By making a series of assumptions detailed in [Gales and Young, 1995], the estimation may be approximated by

$$\mathbf{h}^l = \arg\min_{\mathbf{h}^l} \left[\left|\hat{\mu}^l(\mathbf{h}^l) - \frac{1}{T}\sum_{\tau=1}^{T} \mathbf{o}^l(\tau)\right|\right] \tag{42}$$

where $\hat{\mu}^l(\mathbf{h}^l)$ indicates the dependence of the corrupted-speech mean on the estimate of the bias. This is a simple one-dimensional line search and is significantly faster than direct optimisation, particularly if the PMC log-add approximation is used.

A couple of estimation schemes have been based on an alternative approximation. In [Gauvain et al., 1996] a scheme using the observed data and the training data was proposed.

1. Make initial estimate of $\mathbf{h}^l$, $\mathbf{h}^{l(0)}$, $k = 0$.

2. Calculate

$$\delta h_i^{l(k+1)} = \frac{1}{T_{test}} \sum_{\tau=1}^{T_{test}} o_i^l(\tau) - \frac{1}{T_{train}} \sum_{\tau=1}^{T_{train}} o_i^{l(k)}(\tau) \tag{43}$$

[18]This paper does not look at approaches to estimating the additive noise model. If an ambient noise model is not available, then an iterative scheme where an initial estimate of the noise is used in recognition can be run. Any data assigned to "silence" may then be used to estimate the additive noise model [Gauvain et al., 1996].

[19]There is nothing to prevent standard optimisation schemes for more complex transforms being used, other than possibly efficiency.

[20]This is not a constraint of many of the schemes described here, but does complicate the estimation process.

where

$$\mathbf{o}^{l(k)}(\tau) = \log\left(\exp(\mathbf{s}^{l}(\tau) + \mathbf{h}^{l(k)}) + \exp(\mathbf{n}^{l}(\tau))\right) \tag{44}$$

3. Set $\mathbf{h}^{l(k+1)} = \mathbf{h}^{l(k)} + \delta\mathbf{h}^{l(k+1)}$ and repeat step 2 with $k = k+1$ until convergence criterion met.

An alternative scheme, not using the actual observed data in the estimation process just statistics derived from it, is described in [Minami and Furui, 1996]. Here step 2 is replaced by an estimate using the SAT approach [Rose et al., 1994, Sankar and Lee, 1995]

$$\delta h_i^{c(k+1)} = \frac{\sum_{\tau=1}^{T}\sum_{m=1}^{M}\frac{\gamma_m(\tau)(o_i^c(\tau)-\hat{\mu}_i^{c(k)})}{\hat{\sigma}_i^{c(k)2}}}{\sum_{\tau=1}^{T}\sum_{m=1}^{M}\frac{\gamma_m(\tau)}{\hat{\sigma}_i^{c(k)2}}} \tag{45}$$

where

$$\hat{\mathcal{M}}^{(k)} = \mathcal{K}(\mathcal{M}_s, \mathbf{h}^{c(k)}) \otimes \mathcal{M}_n \tag{46}$$

In this implementation SAT is used only to generate an estimate of the average offset; no variance is estimated[21]. These schemes are not guaranteed to increase the likelihood of the data at each iteration but in practice have been found to be stable. Similar schemes based on the VTS approximation have also been proposed [Moreno, 1996].

An interesting extension is possible when using this simple form of compensation. Previously it was mentioned that it is not possible to perform PMC when CMN is applied to the training data. This problem may be overcome using these simple bias estimation schemes. Rather than considering CMN, it is possible to consider a ML estimate of the bias on a per-sentence level in training. This is very similar to CMN, but is more reliable for shorter sentences. Models are trained on data with this ML bias subtracted, instead of using CMN. During testing the bias, equivalent to that applied to the training data, is estimated using one of the schemes described above. This bias is only applied to the clean-speech models, hence it is consistent with the training. It is important to stress that during testing no bias removal is applied to the data, any bias effects are handled by the model compensation process.

10.2 Parameters of $\mathcal{G}(\mathcal{M}, \mathbf{o}_T)$

Estimating the parameters of $\mathcal{G}(\mathcal{M}, \mathbf{o}_T)$ is a standard problem, since now the transform is directly acting upon a model set. The task is the same as the standard adaptive schemes, however the model to be adapted has previously been modified using predictive schemes. To date only linear transformations based on MLLR [Leggetter and Woodland, 1995] have been used [Gauvain et al., 1996, Gales and Woodland, 1996]. The transformations typically have the form

$$\hat{\mu}^c = \mathbf{A}\mu^c + \mathbf{b} \tag{47}$$

$$\hat{\mathbf{\Sigma}}^c = \mathbf{L}\mathbf{\Sigma}\mathbf{L}^T \tag{48}$$

where $\mathbf{A}$ and $\mathbf{L}$ are transformation matrices associated with the mean and variance respectively, and $\mathbf{b}$ is the bias vector applied to the mean. The transformation parameters are estimated in a ML scheme. Of course, other transformations may be used, many of which have been applied in speaker adaptation e.g. [Neumeyer et al., 1995]. Although it is not necessary to use simple linear transforms (for example MAP estimation could be used instead), linear transforms are popular since they are able to robustly adapt model parameters using a small amount of adaptation data.

[21] An interesting question is whether the optimisation is stable when this variance term is incorporated, as there is no strict proof of convergence for either of these schemes.

11 Results

This paper has deliberately avoided quoting performance figures as it is notoriously hard to obtain comparable results for a variety of well-implemented schemes. Rather than trying to compare a series of schemes, some results are given that illustrate both the strengths and weaknesses of the predictive compensation schemes. The results are taken from [Gales and Woodland, 1996] and [Woodland et al., 1996]. The test sets were from the 1994 ARPA CSRNAB evaluations. For details of the acoustic models used see [Gales and Woodland, 1996]. The two tasks considered were:

1. **Spoke 5: Microphone Independence**. This task uses data recorded with unknown microphones. These microphones had a variety of acoustic attributes, plus in many cases in addition to convolutional noise there is a low level background noise. It is a closed 5k word recognition task with about 10 sentences from each of 20 speakers.
2. **Spoke 10: Noisy Channel**. The S10 spoke is a closed 5k word task with car noise artificially added onto clean speech. Three noise levels are given, however only the noisiest *Level 3* condition, approximately 10dB SNR, is presented in this paper. There were approximately 10 sentences from each of 10 speakers.

Table 4 goes here

The baseline performance for these tasks used cepstral mean normalisation (CMN) in both training and testing. This should yield a good degree of immunity to convolutional noise when there is very little additive noise and the simple mismatch function described for convolutional noise is a good representation of the acoustic environment. On the S5 task the word error rate was 17.4% and on the S10 task 54.3%. These performance figures are poor compared to the clean system performance of around 6%. The performance on the S5 task illustrates that for practical systems, where there is invariably some low level additive noise, CMN does not achieve microphone robustness. For the S5 task incremental MLLR, an adaptive compensation scheme, was also performed. In this case, and for all other MLLR transforms, a block-diagonal transform was estimated with separate blocks for the static delta and delta-delta parameters. It was assumed that the speaker boundaries were known and the transforms reset for each speaker. Using MLLR the word error rate was reduced to 12.1%. Although yielding a good reduction in word error rate (about 30%), the performance is still poor compared to that of the speaker-independent clean system despite the advantage of some speaker adaptation. The reason for the relatively poor performance is that the initial models were poorly matched to the acoustic environment and there is little adaptation data, particularly for the first few sentences. This has two consequences. First the initial alignments for the adaptation training, $(\gamma_m(\tau))$, are poor, due to not knowing the correct transcription. Furthermore the ability of a few linear transform to compensate for all the effects of the acoustic environment is limited. The obvious solution to both these problems is to improve this initial estimate, using predictive compensation schemes.

The predictive scheme chosen was PMC with the log-add approximation, updating the means for the static, delta and delta-delta parameters[22]. The basic additive noise PMC was only performed on the S10 task as the S5 task is known to have primarily convolutional noise present. PMC reduced the error rate on the S10 task to 11.2%. Using a combined scheme, where a simple bias is estimated per speaker for the transform $\mathcal{K}()$ in equation 39 (this bias was estimated using the first sentence for each speaker), the error rate on S10 was further reduced to 10.7%. This

[22] Although a simple scheme, its performance in compensating the means is close to that of SPR [Gales, 1995]. The variances are not adapted.

shows that a simple bias can achieve a very crude form of speaker adaptation. Using PMC and the same bias estimation procedure on the S5 data gave a word error rate of 10.3%, lower than the incremental MLLR result. Despite these reductions in word error rate the performance is still significantly worse than that achieved in clean environments.

The performance can be further enhanced using an adaptive scheme in conjunction with a predictive scheme. Incremental MLLR, in the form of the transform $\mathcal{G}()$ in equation 39, was used in addition to PMC log-add scheme. On the S5 data this gave a word error rate of 8.6% when the means were adapted with MLLR. This was further reduced to 8.0% when a diagonal transform of the variance was used. For the S10 data the results were 9.3% with a mean transform and 8.9% with both a mean and variance transform. On the S5 task, the error rate was 34% lower than that achieved using MLLR alone. It is interesting to note the gains that are obtained using variance adaptation. To compensate the variances using PMC is far more computationally expensive than compensating the means (the variances were not adapted in these experiments since log-add PMC was used). Using simple linear transforms in an adaptive scheme to compensate the variances is a computationally very efficient alternative.

These results illustrate that although good noise robustness may be achieved using predictive schemes (in this case PMC), it is possible to further improve the performance by combining predictive and adaptive schemes. As an additional contrast it is worth examining the performance of a multi-environment trained system. This is really only applicable to the S5 task as in the vast majority of situations it is not possible to predict the level or nature of the additive noise. On this task using models trained on the secondary channel data (i.e. using data from many different microphones), with no noise compensation other than CMN, the word error rate was 9.0%. Using incremental MLLR adaptation on these models an error rate of 7.2% was achieved. This illustrates one of the limitations of predictive compensation schemes, as the additive and convolutional noise mismatch function does not fully model the effects of the environment.

12 Conclusions

This paper has reviewed a class of noise robustness schemes known as predictive model-based compensation. The aim of these schemes is to compensate clean-speech models to approximate training in the same corrupted-speech environment as used for recognition. The model-based compensation is performed by combining a variety of models: the speech model, channel model and additive noise model (plus possibly a stress model), to generate a corrupted-speech model. No speech data is required in the new acoustic environment. The combination is based upon a mismatch function that describes the effect of the noise in the new environment on clean-speech. To efficiently estimate the corrupted-speech model a series of assumptions and approximations to this function are required. A number of approximations have been proposed, such as parallel model combination and vector Taylor series. The choice of approximation determines the computational load, and which model parameters can be compensated. Predictive model-based techniques have a number of advantages. In addition to not requiring any speech data, they can individually compensate all the model parameters and generate meaningful models. Non-Gaussian corrupted-speech distributions can be modelled which also allows complicated noise environments to be handled. Predictive techniques also suffer various limitations. In particular, they require noise models (of both the channel and additive noise) and a good model of how the acoustic environment affects the speech signal to be available. Furthermore, many of the schemes are computationally expensive, especially if dynamic parameters or model variances are to be adapted. Despite their limitations, predictive schemes have been shown to achieve good levels of noise robustness.

Some of the limitations of predictive schemes may be overcome if the schemes are combined with adaptive schemes which make use of speech data in the new acoustic environment. These adaptive schemes may be used to both estimate, for example the channel difference, and to compensate for any slight errors in the model parameters, or mismatch functions. They may also be used to efficiently adapt the variances, which are usually very computationally expensive to estimate using predictive schemes. Other limitations of the predictive framework are harder to overcome.

It is difficult to derive mismatch functions for all the feature vectors commonly used in speech recognition. How to compensate for the effects of stress, other than in very crude ways, has not been extensively investigated. Additionally, the current mismatch functions are only crude representations of the actual environment.

In summary, there are a variety of predictive model-base compensation schemes. These schemes have been shown to achieve good levels of noise robustness in a variety of noise environments. Combining these schemes with adaptive schemes yields improved performance. However, the final choice of noise robustness scheme, whether a predictive scheme or another approach, is highly dependent on the nature of the task and difference in acoustic environment.

Acknowledgments

The author was funded as a Research Fellow at Emmanuel College, Cambridge. Kate Knill made many useful comments during the writing of this paper. The author would also like to thank the anonymous reviewers for their many helpful comments.

A Dynamic Parameters

For speaker-independent and large-vocabulary tasks it is normal to add dynamic features in the form of delta and delta-delta features to the feature vector. In [Gales, 1995] the need to compensate these dynamic parameters for good performance in noise was demonstrated. Compensating these dynamic parameters may be performed in the standard framework by obtaining mismatch functions for the dynamic features. If delta features are calculated using simple differences over a window size of w, then the mismatch function may be shown to be [Gales and Young, 1996]

$$\begin{aligned}
\Delta o_i^l(\tau) &= \mathcal{F}^{\Delta}(s_i^l(\tau-w), n_i^l(\tau-w), \Delta s_i^l(\tau), \Delta n_i^l(\tau)) \\
&= \log\left(\exp(\Delta s_i^l(\tau) + s_i^l(\tau-w)) + \exp(\Delta n_i^l(\tau) + n_i^l(\tau-w))\right) \\
&\quad - \log\left(\exp(s_i^l(\tau-w)) + \exp(n_i^l(\tau-w))\right)
\end{aligned}$$

where $\mathbf{\Delta s}(\tau)$ is the speech delta feature vector at time τ (and similarly for the noise and corrupted speech). Similar expressions may also be derived for the delta-delta features [Gales and Young, 1996] (indicated by $\mathbf{\Delta}^2\mathbf{s}(\tau)$). It is interesting to note that the "noisy" delta parameters at time τ are a function of the static parameters at time $\tau - w$. In [Gales, 1995] the need to explicitly model data at time $\tau - w$ are investigated and found to be important.

An alternative mismatch function is obtained by using matrix-based dynamic parameters [Gales, 1995]

$$\begin{bmatrix} \mathbf{o}^c(\tau) \\ \mathbf{\Delta o}^c(\tau) \\ \mathbf{\Delta}^2\mathbf{o}^c(\tau) \end{bmatrix} = \begin{bmatrix} \mathbf{0}^n & \mathbf{I}^n & \mathbf{0}^n \\ -\frac{1}{2}\mathbf{I}^n & \mathbf{0}^n & \frac{1}{2}\mathbf{I}^n \\ \frac{1}{2}\mathbf{I}^n & -\mathbf{I}^n & \frac{1}{2}\mathbf{I}^n \end{bmatrix} \begin{bmatrix} \mathbf{o}^c(\tau - w) \\ \mathbf{o}^c(\tau) \\ \mathbf{o}^c(\tau + w) \end{bmatrix}$$

Using these matrix-based schemes mismatch functions for the both linear-regression and simple difference dynamic parameters may be generated. Two-dimensional cepstral parameterisation of the speech has also been examined [Milner and Vaseghi, 1994] for use with predictive schemes.

One final approximate scheme is to use a continuous time approximation to the dynamic parameters [Gopinath et al., 1995]. Here $\Delta o_i \approx \partial o_i / \partial t$ and so, for example, assuming that the mean of the noise delta parameters is approximately zero

$$\Delta\hat{\mu}_i^l \approx \left(\frac{\exp(\mu_i^l)}{\exp(\mu_i^l) + \exp(\tilde{\mu}_i^l)}\right) \Delta\mu_i^l \tag{49}$$

It is worth noting that, other than the continuous time approximation, all the mismatch functions are "exact". However, to model the speech and noise data well there may be the need for additional models and inter-element correlation modelling [Gales, 1995]. Care must be taken about the assumptions of the accuracy of the speech and noise models when compensating dynamic parameters. Furthermore, not all the predictive compensation schemes mentioned are applicable to compensating dynamic parameters, for example SND only handles static parameters. Other schemes, such as, numerical integration PMC, may become computationally too expensive.

It is also interesting to note that dynamic model parameters are robust to "ideal" convolutional noise, as defined in the previous section. However, they are not robust when additive noise is present.

B Bandwidth-Limited Data

The standard convolutional noise model cannot easily handle the case of band-limited data, for example, if the clean-speech models are trained on data recorded at a 16KHz bandwidth and the test data is recorded over an 8KHz bandwidth channel. An alternative model is required.

Figure 9 goes here

Figure 9 shows the standard MFCC filterbank over the complete frequency range and the position of the maximum band limited frequency. From the diagram, only the first M frequency bins will be correctly estimated, $M+1$ to N will contain noise. This maps into a simple convolutional noise model. Replace the noise in bands $M+1$ to N by 1 (or 0 in the log-spectral domain) and the parameterisation is now

$$o_i^{bc}(\tau) = \sum_{j=1}^{N} w_{jj} c_{ij} o_j^l(\tau) \tag{50}$$

where $\mathbf{o}^{bc}(\tau)$ is the corrupted-speech data in the cepstral domain after the bandlimited channel,

$$w_{ij} = \begin{cases} 1, & \text{iff } i = j \text{ and } j \leq M \\ 0, & \text{otherwise} \end{cases} \tag{51}$$

The channel model is therefore $\mathbf{W}$. In terms of how this affects the mean

$$\mu^{bc} = \mathbf{C}\mathbf{W}\mathbf{C}^{-1}\mu^c = \tilde{\mathbf{W}}\mu^c \tag{52}$$

and variance

$$\mathbf{\Sigma}^{bc} = \tilde{\mathbf{W}}\mathbf{\Sigma}^c\tilde{\mathbf{W}}^T \tag{53}$$

In practice, if the test data channel bandwidth were known *a priori* to be bandwidth limited, the filterbank analysis would be altered in training to improve the resolution in the pass band. This is not possible using predictive schemes.

References

[Beattie and Young, 1992] Beattie, V. L. and Young, S. J. (1992). Hidden Markov model state-based Cepstral noise compensation. In *Proceedings ICSLP*, pages 519–522.

[Bellegarda, 1997] Bellegarda, J. (1997). Statistical techniques for robust ASR: Review and perspective. In *Proceedings Eurospeech*, pages KN–33 – KN–36.

[Berstein and Shallom, 1991] Berstein, A. D. and Shallom, I. D. (1991). An hypothesized Wiener filtering approach to noisy speech recognition. In *Proceedings ICASSP*, pages 913–916.

[Boll, 1979] Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions ASSP*, 27:113–120.

[Davis and Mermelstein, 1980] Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions ASSP*, 28:357–366.

[Ephraim, 1992] Ephraim, Y. (1992). Gain-adapted hidden Markov models for recognition of clean and noisy speech. *IEEE Transactions Signal Processing*, 40:1303–1316.

[Ephraim et al., 1989] Ephraim, Y., Malah, D., and Juang, B. H. (1989). On the application of hidden Markov models for enhancing noisy speech. *IEEE Transactions ASSP*, 37:1846–1856.

[Furui, 1986] Furui, S. (1986). Speaker independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions ASSP*, 34:52–59.

[Furui and Sondhi, 1992] Furui, S. and Sondhi, M. M. (1992). *Advances in Speech Signal Processing.* Marcel Dekken.

[Gagnon, 1992] Gagnon, L. (1992). A noise reduction approach for non-stationary additive interference. In *Proceedings ESCA Workshop in Speech Processing in Adverse Conditions*, pages 139–142.

[Gales, 1995] Gales, M. J. F. (1995). *Model-Based Techniques for Noise Robust Speech Recognition.* PhD thesis, Cambridge University.

[Gales and Woodland, 1996] Gales, M. J. F. and Woodland, P. C. (1996). Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 10:249–264.

[Gales and Young, 1992] Gales, M. J. F. and Young, S. J. (1992). An improved approach to the hidden Markov model decomposition of speech and noise. In *Proceedings ICASSP*, pages 233–236.

[Gales and Young, 1993] Gales, M. J. F. and Young, S. J. (1993). Cepstral parameter compensation for HMM recognition in noise. *Speech Communication*, 12:231–240.

[Gales and Young, 1995] Gales, M. J. F. and Young, S. J. (1995). Robust speech recognition in additive and convolutional noise using parallel model combination. *Computer Speech and Language*, 9:289–307.

[Gales and Young, 1996] Gales, M. J. F. and Young, S. J. (1996). Robust speech recognition using parallel model combination. *IEEE Transactions Speech and Audio Processing*, 4:352–359.

[Gauvain et al., 1996] Gauvain, J. L., Lamel, L., Adda, D., and Matrouf, D. (1996). Developments in continuous speech dictation using the 1995 ARPA NAB news task. In *Proceedings ICASSP*, pages 73–76.

[Gauvain and Lee, 1994] Gauvain, J. L. and Lee, C. H. (1994). Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions Speech and Audio Processing*, 2:291–298.

[Gong, 1995] Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech Communication*, 16:261–291.

[Gopinath et al., 1995] Gopinath, R. A., Gales, M. J. F., Gopalakrishnan, P. S., Balakrishnan-Aiyer, S., and Picheny, M. A. (1995). Robust speech recognition in noise — performance of the IBM continuous speech recognizer on the ARPA noise spoke task. In *Proceedings ARPA Workshop on Spoken Language Systems Technology*, pages 127–130.

[Hansen and Arslan, 1995] Hansen, J. H. L. and Arslan, L. M. (1995). Robust feature-estimation and objective quality assessment for noisy speech recognition using the Credit Card corpus. *IEEE Transactions Speech and Audio Processing*, 3:169–184.

[Hermansky and Morgan, 1994] Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions Speech and Audio Processing*, 2:578–579.

[Juang and Rabiner, 1985] Juang, B. H. and Rabiner, L. R. (1985). Mixture autoregressive hidden Markov models for speech signals. *IEEE Transactions ASSP*, 33:1404–1413.

[Junqua and Anglade, 1990] Junqua, J. and Anglade, Y. (1990). Acoustic and perceptual studies of Lombard speech: Application to isolated-word automatic speech recognition. In *Proceedings ICASSP*, pages 841–844.

[Klatt, 1979] Klatt, D. H. (1979). A digital filterbank for spectral matching. In *Proceedings ICASSP*, pages 573–576.

[Leggetter and Woodland, 1995] Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech and Language*, 9:171–186.

[Martin et al., 1993] Martin, F., Shikano, K., and Minami, Y. (1993). Recognition of noisy speech by composition of hidden Markov models. In *Proceedings Eurospeech*, pages 1031–1034.

[Milner and Vaseghi, 1994] Milner, B. P. and Vaseghi, S. V. (1994). Comparison of some noise-compensation methods for speech recognition in adverse environments. *IEE Proceedings Vision, Image and Signal Processing*, pages 280–288.

[Minami and Furui, 1995] Minami, Y. and Furui, S. (1995). A maximum likelihood procedure for a universal adaptation method based on HMM composition. In *Proceedings ICASSP*, pages 129–132.

[Minami and Furui, 1996] Minami, Y. and Furui, S. (1996). Adaptation method based on HMM composition and EM algorithm. In *Proceedings ICASSP*, pages 327–330.

[Moreno, 1996] Moreno, P. J. (1996). *Speech Recognition in Noisy Environments*. PhD thesis, Carnegie Mellon University.

[Moreno et al., 1996] Moreno, P. J., Ray, B., and Stern, R. M. (1996). A vector Taylor series approach for environment-independent speech recognition. In *Proceedings ICASSP*, pages 733–736.

[Morgan and Hermansky, 1992] Morgan, N. and Hermansky, H. (1992). RASTA extensions: Robustness to additive and convolutional noise. In *Proceedings ESCA Workshop on Speech Processing in Adverse Conditions*, pages 115–118.

[Neumeyer et al., 1995] Neumeyer, L. R., Sankar, A., and Digalakis, V. V. (1995). A comparative study of speaker adaptation techniques. In *Proceedings Eurospeech*, pages 1127–1130.

[Openshaw and Mason, 1994] Openshaw, J. P. and Mason, J. S. (1994). On the limitations of Cepstral features in noise. In *Proceedings ICASSP*, volume 2, pages 49–52.

[Rose et al., 1994] Rose, R. C., Hofstetter, E. M., and Reynolds, D. A. (1994). Integrated models of signal and background with application to speaker identification in noise. *IEEE Transactions Speech and Audio Processing*, 2:245–257.

[Sankar and Lee, 1995] Sankar, A. and Lee, C. H. (1995). Robust speech recognition based on stochastic matching. In *Proceedings ICASSP*, pages 121–124.

[Seymour and Niranjan, 1994] Seymour, C. W. and Niranjan, M. (1994). An HMM based Cepstral-domain speech enhancement scheme. In *Proceedings ICSLP*, pages 1595–1598.

[Varga et al., 1988] Varga, A. P., Moore, R., Bridle, J., Ponting, K., and Russell, M. (1988). Noise compensation algorithms for use with hidden Markov model based speech recognition. In *Proceedings ICASSP*, pages 481–484.

[Varga and Moore, 1990] Varga, A. P. and Moore, R. K. (1990). Hidden Markov model decomposition of speech and noise. In *Proceedings ICASSP*, pages 845–848.

[Woodland et al., 1996] Woodland, P. C., Gales, M. J. F., and Pye, D. (1996). Improving environmental robustness in large vocabulary speech recognition. In *Proceedings ICASSP*, pages 65–68.

List of Tables

List of Figures

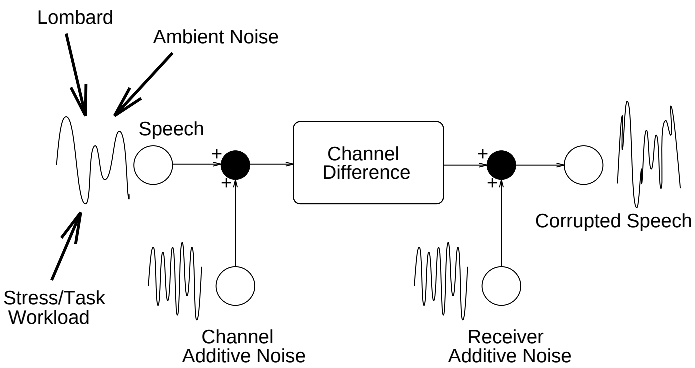

Figure 1: The acoustic environment.

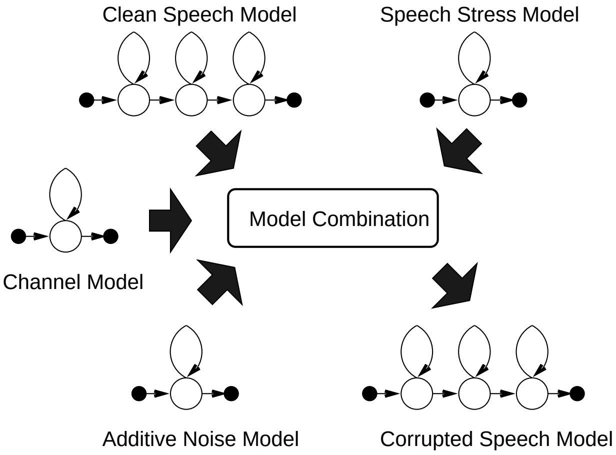

Figure 2: A predictive model compensation framework.

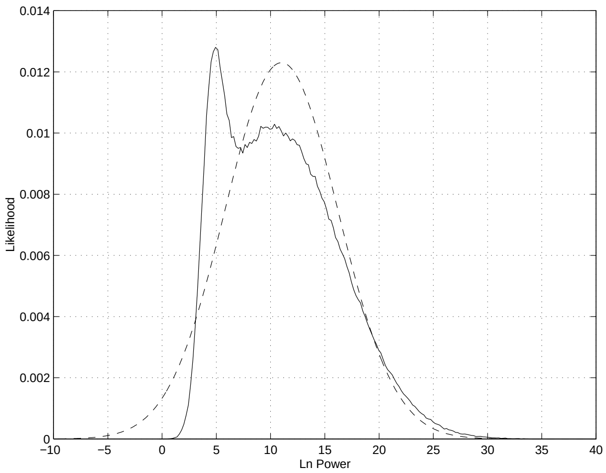

Figure 3: Effects of additive noise, the solid line (-) is the actual distribution , the dashed line (- -) is the single Gaussian approximation.

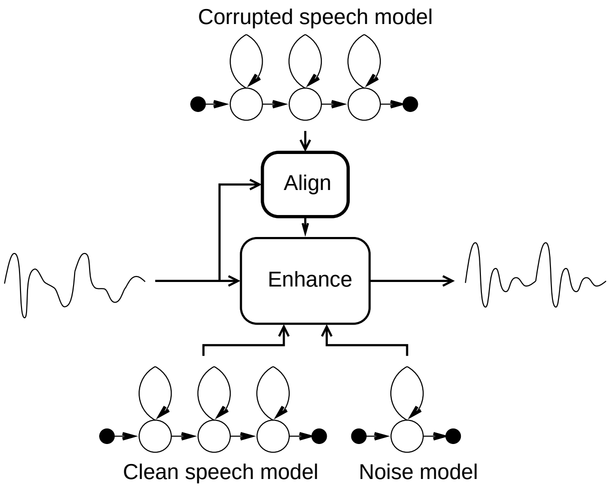

Figure 4: Predictive model-based enhancement.

Condition	Likelihood
$\tilde{\mu}_i^l < o_i^l(\tau) < \mu_i^l$	$\mathcal{N}(o_i^l(\tau); \mu_i^l, \Sigma_{ii}^l)$
$o_i^l(\tau) < \tilde{\mu}_i^l < \mu_i^l$	$\mathcal{N}(\tilde{\mu}_i^l; \mu_i^l, \Sigma_{ii}^l)$
$o_i^l(\tau) < \mu_i^l < \tilde{\mu}_i^l$	$\mathcal{N}(\tilde{\mu}_i^l; \tilde{\mu}_i^l, \Sigma_{ii}^l)$
$\tilde{\mu}_i^l < \mu_i^l < o_i^l(\tau)$	$\mathcal{N}(o_i^l(\tau); \mu_i^l, \Sigma_{ii}^l)$
$\mu_i^l < \tilde{\mu}_i^l < o_i^l(\tau)$	$\mathcal{N}(o_i^l(\tau); \tilde{\mu}_i^l, \Sigma_{ii}^l)$
$\mu_i^l < o_i^l(\tau) < \tilde{\mu}_i^l$	$\mathcal{N}(\tilde{\mu}_i^l; \tilde{\mu}_i^l, \Sigma_{ii}^l)$

Table 1: The Klatt noise masking scheme [Klatt, 1979].

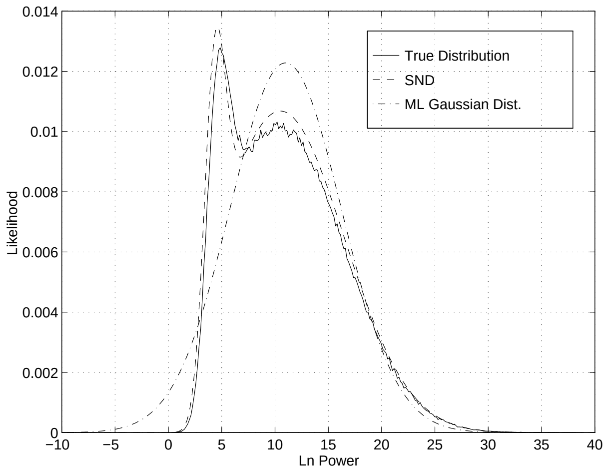

Figure 5: Speech and noise decomposition, the solid line (-) is the actual distribution, the dashed line (- -) is the SND approximation and the dashed-dot line (-.) is the single Gaussian approximation.

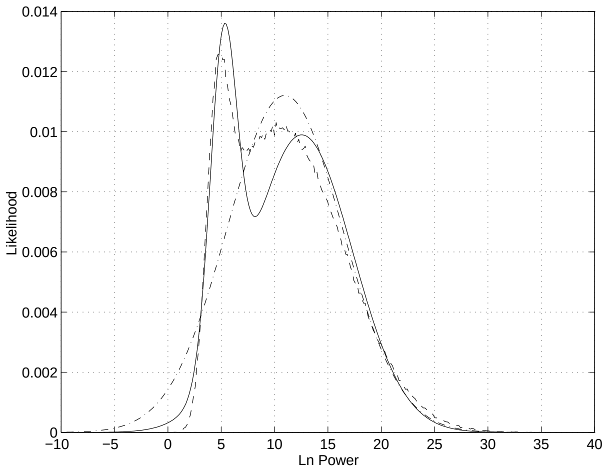

Figure 6: Iterative parallel model combination, the solid line (-) is the data modelled by two Gaussian distributions, the dot-dashed line (-.) is the data modelled by a single Gaussian distribution and the dashed line (- -) is the distribution of the data.

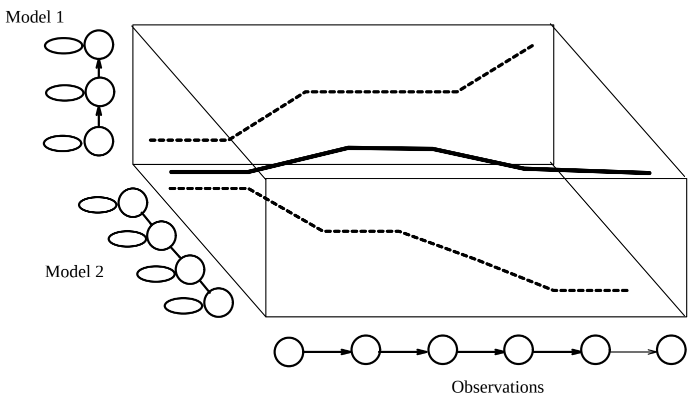

Figure 7: 3-dimensional Viterbi decoder.

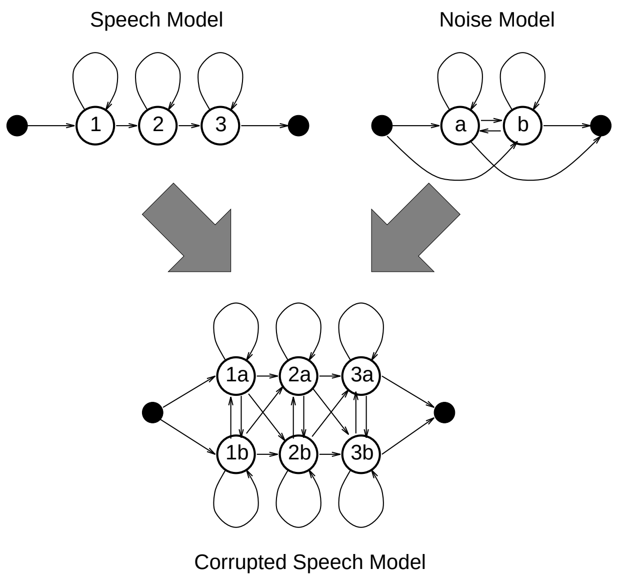

Figure 8: Mapping a 3-D Viterbi decoder to a 2-D Viterbi decoder.

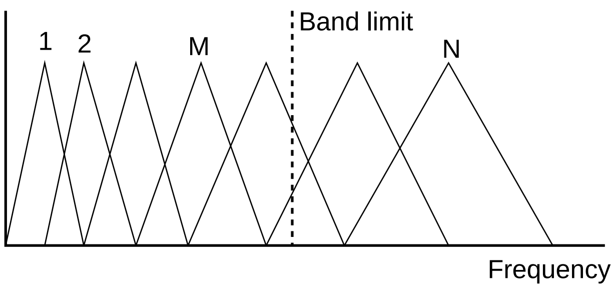

Figure 9: Bandwidth limited data.

Technique	Final Model	Domain	Compensate Features			Compensate Parameters		Cost
			$\mathbf{o}$	$\boldsymbol{\Delta}$	$\boldsymbol{\Delta}^2$	μ	$\boldsymbol{\Sigma}$	
PMC Log-Normal	Gauss.	cep	y	n	n	y	y	High
PMC Log-Add	Gauss.	cep	y	y	y	y	n	Low
PMC Num. Int.	Gauss.	cep	y	y	y	y	y	High
VTS	Gauss.	cep	y	y	y	y	y	High
Noise Masking	Non-Gauss.	log	y	n	n	n/a	n/a	Low
SND	Non-Gauss.	log	y	n	n	n/a	n/a	Low
IPMC	Mult-Gauss.	cep	y	y	y	y	y	High

Table 2: Summary of predictive model-based schemes described.

Gauss	Single Gaussian distribution
Mult-Gauss	Multiple Gaussian distributions
cep	Cepstral domain
log	Log-spectral domain
$\mathbf{o}$	Static features
$\boldsymbol{\Delta}$	Delta dynamic features
$\boldsymbol{\Delta}^2$	Delta-delta dynamic features
y	May be compensated with this technique
n	Cannot be easily compensated with this technique
n/a	Not applicable

Table 3: Glossary of abbreviations used in table 2.

Type Compensation	Compensation Scheme	Task	
		S5	S10
Inherent	CMN	17.4	54.3
Adaptive	+ MLLR	12.1	—
Predictive	PMC	—	11.2
Combined	+ $\mathcal{K}()$	10.3	10.7
Combined	+ $\mathcal{G}()$ (Mean)	8.6	9.3
Combined	+ $\mathcal{G}()$ (Variances)	8.0	8.9
Multi-Environment	CMN	9.0	—
Adaptive	+ MLLR	7.2	—

Table 4: Word error rates on the ARPA 1994 S5 and S10 tasks using various compensation schemes.