Literature Review - Speech Recognition for Noisy Environments

Introduction

The Current State of Speech Recognition Systems

In recent years Automatic Speech Recognition has reached very high levels of performance, with word-error rates dropping by a factor of five in the past five years.

This current state of performance is largely due to improvements in four main areas of the field . The first is the use of common speech corpora allows the easy use of large training sets and a 'level playing field' when comparing the results of new recognition systems. Secondly, new ideas in acoustic modeling have also led to significant improvements in performance. Context specific HMM phonetic models, the modelling of cross-word effects, changes in feature vectors over time, are just a few of these new techniques that have helped reduce word error rates by up to a factor of two. The use of language modelling with statistical n-gram grammars have also played an important part in improving recognition of large vocabulary corpora using probabilities calculated from millions of lines of text. Finally, improvements in search algorithms and workstation speed and memory have allowed for a shorter experimentation cycle. This allows for far more complex algorithms and databases (e.g. tri-gram grammar databases) to be used whilst operation and training times have been reduced.

Current levels of performance can be seen by examining the recognition rates from two common speaker independent continuous speech corpora.

The first, a small corpora designed for use with a closed vocabulary (i.e. all words tested with the recogniser are contained within its lexicon) is the TI connected digits corpus (Leonard,1984). This has a vocabulary of 10 words with 4 hours of training data, the current word error rate for this corpora is 0.3%. The second corpora is far larger, and can therefore be used with either open or closed test vocabulary, this is the ARPA Wall Street Journal Dictation (Paul,1992) corpora. This has a vocabulary of 20,000 words with 12 hours of training data, current state of the art recognisers achieve an error rate of 13% for a closed test vocabulary, 26% for open vocabulary.

Speech Recognition Systems and Noise

Unfortunately, the speech recognition systems reviewed above all have one criteria in common, they are designed to work in controlled environments using clean speech. If these systems are exposed to speech taken from noisy environments then their performance degrades rapidly. For example, an isolated word recogniser trained on real speech gives 100% accuracy with clean speech, this can typically drop to 30% when used in a car travelling at 90kmh (Lockwood and Boudy 1992). As can be seen the differences between using a recogniser in clean and noisy environments are extreme, this causes one of the major obstacles in producing a commercial recognition system to be used in 'normal' environments.

If a recognition system is to be used in these noisy environments it must be 'robust' to many different types and levels of noise, categorised as either additive noise, or changes in the speakers voice *due* to noise. Additive noise (such as car engine noise, background babble, white noise etc.) contaminates the speech signal changing the data vectors which represent the speech. Changes in the speakers voice are caused by modifications of articulation to increase intelligibility where auditory feedback is affected by excess levels of

noise. This is known as the 'Lombard Effect'(Lunqua,1993), causing highly variable intra and inter speaker distortion highly damaging to recognition performances . The main variances can be found in increases in the speakers pitch, amplitude, vowel duration, and spectral tilt, as well as shift in formant frequencies F1 and F2. However these changes are by no means constant, even for the same speaker under similar noise conditions (average 3:1 difference ratio of distortion for the same speaker) making this type of noise highly difficult to either remove or model.

Approaches to Speech Recognition in Noise

It is the effect of these types of noise that produce serious mismatches between the training and recognition data used with the recognition systems seen above. Therefore if the error-rate for recognition in noisy environments is to be reduced then a function must be found which will reduce the differences between these two environments. This can be done in two ways, by changing the speech model parameters to match the speech environment or by transforming the recognition speech data from the noisy environment to the environment where the models were trained.

These methods can be divided into the following three basic approaches:

Use of Noise Resistance Features and Similarity Measurements - This approach assumes that the system is independent of noise, with the same configuration used for both clean and noisy speech. Therefore these techniques focus of the effects of noise on the speech signal rather on the removal of noise, with attempts made to derive features and parameters of speech which are noise resistant. Along with feature and parameter identification suitably robust similarity measures must also be used, since it has been found (Bateman et al, 1992) that these measures are not usually independent of the representations. These techniques, although not the most successful have the significant advantage that they are applicable to wide ranges of noisy environments, since assumptions about the characteristics of the noise are rare.

Speech Enhancement - This involves the transformation of noisy speech into as close an approximation of the training environment as possible. By preprocessing speech enhancement techniques are intended to recover either the waveform or specific parameters of clean speech embedded in noise. However, these techniques are not usually directly related to improving speech recognition performance. Most of these techniques were originally developed for speech quality enhancement, and while the distortion caused by then techniques are tolerable for humans they cause problems for recognition systems.

Speech Model Compensation for Noisy Environments - Model Compensation entails the transformation of speech models created in a reference environment to accommodate a specific noisy environment in order to recognise noisy speech. This technique, rather than attempting to derive a clean estimate of speech allows for noise in the recognition process itself. The usual statistical modelling techniques (Hidden Markov Models, Artificial Neural Networks etc.) are trained using clean speech, with the model parameters then adapted to accommodate noisy speech. This is used to compensate for the discrepancy between training and operational conditions, unfortunately, the usual case is that a model is required for each operating noise environment. Therefore a noise model trained in one environment may only be used in that specific noise environment, making the system too specific for general use.

There are many approaches to the implementation of the methods of noisy speech recognition covered above, although some of the techniques, while starting from different objectives, have led to similar solutions but are differently classified. Until recently most of these were based upon pure signal processing and mathematical methods to produce the most accurate representations and features possible by any means. However, more recently techniques have been developed which take advantage of speech production and perception knowledge. Because the human hearing system is the most efficient recognition system known then it is hoped by using auditory, physiological, and production knowledge in the design of recognition systems that

some of the success of human recognition can be duplicated. A brief explanation of both signal processing and knowledge based approaches will be covered in the next section, this will serve as a background to the summary of the noisy speech recognition techniques shown later in this document.

Background on Speech Processing Tools and Auditory Knowledge

Basic Speech Analysis Tools

One the current problems in speech analysis and representation is in finding a set of parameters that can encapsulate the inherent variability found in speech signals. Unfortunately, due to the limitations of current mathematical frameworks, and our limited knowledge of speech analysis and perception, this robust representation is not available.
Instead, most speech analysis techniques encode speech parameters in the frequency domain, a domain which enables most speech signals to be discriminated accurately. Conversion from time to frequency domain is based on three basic methods: *fourier transforms, digital filter-banks, and linear prediction*. The Fourier transform allows the passage of a signal from the time to frequency domain, and vice versa (Inverse Fourier Transform). This transform has also been extended for use with discrete time signals, sampled at regular intervals, known as the Discrete Fourier Transform (DFT). This uses a set Window length of samples for analysis which is proportional to the frequency resolution. A large window produces greater frequency resolution, but is naturally at the cost of temporal resolution. This is one of the essential tools of speech processing, along with its more efficient counterpart, that of Fast Fourier Transform (FFT), used in most applications where spectrum estimates are required. The second method for estimating the spectral envelope is via a filter-bank which separates the signal frequency bandwidth in a number of frequency bands where the signal energy is measured. This method offers two main advantages over DFT, that of the small number of parameters used to represent the spectrum envelope, and the possibility of having different frequency resolutions for each envelope. This last advantage, along with the characteristics of the filters and the spacing of their central frequency has been found to be important, for instance, in the simulation of cochlea like filtering in auditory periphery modelling.
Finally, Linear Predictive Coding uses a different approach utilising the modelling the human speech production tract. This is based upon the idea that voiced speech is, at least in some sense, periodic, and so predictable. Therefore if past samples from a speech waveform are modelled, it should be possible to predict the forthcoming samples as a weighted combination of those past samples. The number of previous samples used for LPC defines the number of *co-efficients* (weightings) used in minimising the errors between past and present samples, naturally this error is to be kept to a minimum. The number of co-efficients is also equivalent to the number of *poles* in the linear system, a system with *p* poles can model a spectrum with *p/2* frequency resonances. Therefore, LPC will, theoretically allow us to model, say, the first three formants of a speech signal with 6 poles. This offers both smoothing and data reduction of the speech signal with a minimum loss of information. Unfortunately LPC does not always place the poles on the areas of interest, therefore a number of 'spare' poles are used to allow for these. The choice of order of co-efficients for LPC is a compromise between spectral accuracy and computation time/memory, and the target application. In general an 8th to 14th order model is used to model the first three to five formant peaks, however for some purposes it is sometimes useful to make a gross characterisation of the speech spectrum by using a two-pole model.
Another useful method for speech analysis is that of cepstram analysis, also based on speech production modelstaking advantage of the fact that a speech signal is a convolution of the source (source waveform) and filter (vocal tract). In order to separate these signals a fourier transform is used to convert the convolution to multiplication, which is in turn converted to addition by taking logarithms. Therefore cepstral analysis involves taking a DFT and log magnitude of the speech signal, followed by an inverse fourier transform. The provides a *cepstrum* domain of the waveform with a quefrency (cepstral equivalent of frequency) peak corresponding to the pitch (shown on a DFT as a ripple caused by the harmonic fine structure), and a number

of low quefrency peaks representing the formants. These features would normally be extracted by applying a low or high pass 'lifter' (cepstral equivalent of filter). However, these are not features generally used in recognition systems, but instead the low-time part of the spectrum, the *cepstral co-efficients*, are used. Since most of the formant information is represented by the first 12 or so points in the cepstrum, and found to be relatively robust, it is this that many recognisers use as data vectors for recognition. A variant of these co-efficients also widely used in recognition are *mel-frequency cepstral co-efficients*, these are generated in a similar way, however before the inverse DFT the frequency axis is warped to become more like that used in human pitch perception.

As well as using the basic forms of speech parameter's discussed above many of the speech recognition techniques discussed in later sections will use variants of the two most popular classifiers used in speech recognition, *Hidden Markov Models (HMM)* and *Artificial Neural Networks (ANN)*. These approaches are based on stochastic modelling, which provides a framework for modelling patterns statistically and formalising the decision making process such that the average loss per decision is as small as possible.

The first of these, the HMM, is based upon a statistical state-sequence known as a Markov chain, consisting of a set of states, with transitions between the states. Each state corresponds to a symbol, and to each transition is associated a probability. Symbols are produced as the output of the Markov model by the probabilistic transitioning from one state to another. However, this model is too restrictive to study the complex problems associated with speech recognition. For this purpose it is necessary to extend the model to be able to treat the case where the observations are probabilistic functions of the states. This resulting Hidden Markov Model is similar to the Markov Model except that the output symbols are probabilistic, with all symbols possible at each state, each with its own probability. In other words a HMM is composed of a non-observable "hidden" Markov chain, and an observation process which links the acoustic vectors extracted from the speech signal to the states of the hidden chain. HMM's are found to be particularly effective at modelling speech since they are able model quasi-stationary segments of speech onto states, dependent on the speech unit being modelled which could be anything from sentences to phonemes

The second approach, that of ANN, also known as a connectionist model or Parallel Distributed Processing (PDP) model consists of a large number of interconnected simple non-linear cells. The elementary cell, called a node or 'neuron' consists of a number of inputs and a single output which is a non-linear function of the weighted sum of its inputs (the function is usually sigmoid). The design of an ANN involves a number of issues, namely the choice of network topology and characteristics, and the specification of training methods for adjusting the neuron's weights. The network topology is one of the key issues, and it is this that usually sets the networks training. The most popular of these in ASR are *Multi-layer perceptrons (PLP), self organising maps,* and *recurrent networks* . The most common of which are MLP's, feedforward systems in which the outputs of the nodes of later q form the inputs for layer $q+1$.

ANN's have been used in a number of speech recognition projects of the past few years, some concerned with signal processing, but most concerned with the classification phase of speech recognition. However, whilst similar in application to HMMs ANNs offer a number of advantages over their counterparts. For instance ANN's require weaker assumptions about statistical properties of the input data than HMMs, and can perform discriminative training. This linked with ANN's abilities to produce non-linear functions from outputs and to learn by example has made them particularly useful.

The Use of Auditory Knowledge

When simulating a natural process it seems natural to look to that process for inspiration, for instance, in the design of the camera early inspiration came from the physiological studies of the eye. Therefore when hoping to simulate speech recognition it seems natural to study the best speech processor currently available, the human auditory system. Thus, a great deal of effort has been made to duplicate parts of this system, or take inspiration from it, when designing automatic speech recognition systems.

This area is known as 'auditory modelling', referring to a computational model of the peripheral hearing system. Unfortunately knowledge of the auditory system is fragmented, and whilst much is known of the

auditory periphery knowledge of the central and higher auditory processes still sketchy. Therefore at the current time it is the physiological functions of the basilar membrane and other cochlear processes, up to the neural level, are considered as the primary functions for auditory modelling. The auditory periphery system has the capability of integrating events corresponding to complex spectral and temporal information, by modelling this system it is hoped that the advantages offered over more traditional processes will be conferred on the ASR system. Such advantages include a better temporal localisation of important cues, better detectability in degraded environments, and a reduction in the variability of integrated information known as auditory cues.

Current auditory knowledge used in ASR fall in two different categories, that of modelling *physiological* functions, and *psychoacoustic* features. Physiological functions refer to auditory known gained from the examination of the physical operations of the ear and other auditory processes, psychoacoustic concepts from psychological experiments in hearing and perception.

Physiological models can be classified in two different categories, models which attempt to give a spectral representation of speech sounds at the level of auditory nerve fibres, and models which attempt to reproduce dynamic speech mechanisms. Auditory models belonging to the first type are numerous and widespread across many auditory based ASR systems. The basic model consists of linear formulation of basilar membrane, transformation of basilar membrane vibrations into hair cell displacements, and a simplified non-linear cell transduction and compression into electrical potentials. Additions to this model include reproduction of the lateral inhibition phenomenon by use of Lateral Inhibitory Networks, and efferent-induced effects simulated by using a feedback control mechanism to provide a more robust speech representation in noise. From our point of view dynamic speech mechanisms are essentially reproduced by short-term adaption and forward masking phenomena. Short-term adaption has been shown to enhance the fast changes in intensity which occur in the system Both of these types of model have been applied to speech recognition, however it has been found that the optimal model for feature detection can be found to vary as a function of the acoustic environment. Psychoacoustic models such as loudness and critical-band concepts have now become more or less standard in auditory periphery models, along with the concepts of masking, and saturation.

Techniques for Noisy Speech Recognition

Noise Resistant Features and Distance Metrics

Speech Representation and Similarity Measures

One of the most common representations for speech signals used in recognisers has been in the frequency domain, produced via Discrete Fourier Transform or filter-bank analysis. However, it has been found that whilst the combination of speech signal and noise are additive, and relatively easy to process, speech representations in the cepstral domain can increase the performance of recognisers (Erell and Weintraub,1993) in both clean speech and noise.

Additional improvements to this representation have been introduced in order to improve robustness under noise. White noise scale factors have been added to the cepstral Euclidean distance between noisy and clean vectors to compensate for the reduction in the norm of the cepstral vectors. However, whilst this approach holds for white noise it has been found (Openshaw and Mason,1994) that other types of noise not only cause a reduction of the norm but changes in the statistical parameters of other cepstral coefficients. Other distortion measures for reducing this corruption are many and varied, from the simple Weighted Likelihood Ratio's (WLR) (Nocerino et al,1985) which gives more emphasis to spectral peaks less affected to noise, weighted cepstral distance's, Root Power Sums (RPS), and Sine Wave Liftering (SWL). However, all of these methods have basically the same approach, that is to emphasise spectral peaks over valleys, and to de-

emphasise low frequency cepstral terms when contributing to the distance measure. One of the more successful approaches has been the Sine Wave Liftering, this seeks to truncate the higher cepstral coefficients with the trailing edge of a raised sine lifter (basically a smoothed log power spectrum). Because the lower-order terms of a cepstral representation describe the smooth features of the spectrum due to vocal tract response rather than the fine spectral structure (which adversely affect spectral pattern matching) by removing the higher-order terms the effects of spectral tilt are reduced and spectral peaks enhanced. This technique has been found to give improvements in noisy speech recognition over that of any other distortion measure in testing involving car noise (Nakamura et al,1993).

Distortion measures have also been introduced in improve the application of LPC coefficients, frequency weighted Itakura spectral distortion measures try to compensate for bandwidth broadening due to noise, and high order derivatives of LPC used to compute estimates of the clean LPC. However the most effective method is that of Short-Term-Modified Coherence (SMC) (Mansour and Juang,1989). This differs from normal LPC by all-pole modeling of the auto-correlation sequence of a noisy signal instead of from the speech waveform. By taking advantage of the coherence of adjacent speech segments SNR can be improved by roughly 10-12 dB for noisy speech within a SNR range of 0-20dB. When combined with a cepstral liftering technique SMC was found to give equal recognition rates to standard LPC-cepstrum methods when an a SNR of 15dB lower (Mena et al,1990).

Using Perceptron Learning as a Similarity Measure
The multi-layer perceptron (MLP) Neural Network has become one of the most effective methods for classifying different distributions of speech parameters in the recognition of noisy speech. Tests on the classification of cepstral co-efficient frames of noisy vowel data (Paliwal,1990) showed that MLP gave better accuracy than other statistical classifiers at all levels of SNR with slower rates of degradation. This approach has also been extended from discriminative vector similarity to vector-sequence similarity where certain vectors are given more weighting than others in the recognition process. When tested with a vocabulary specific recogniser (Anglade et al.,1993) it was found that under clean, Lombard and Lombard with white noise effects recognition was better than with a continuous density HMM system.

Linear Discriminant Analysis
Another form of statistical pattern classifier is Linear Discriminant Analysis, this performs a linear transformation of a speech representation by minimising within-class differences and maximising those between classes. This has shown to provide significant reduction in the dimensionality of a speech representation whilst yielding as good or better performance than the original (Bocchieri and Wilpon,1992) representaton.

Applications of Auditory Modelling
The computational modelling of various psychoacoustic and neurophysical knowledge into a auditory periphery front-end's for a variety of speech processing applications is becoming increasingly popular. These applications take advantage of a wide variety of techniques covered by the field of 'auditory periphery modelling' including critical band filtering, loudness curve properties, non-linear energy compression, haircell modelling, short-time adaptation and other peripheral and central auditory processing phenomena. By the use of these models improvements in temporal localisation, speech detectability in degraded environments, and the reduction in the variability of various auditory cues have been seen. This will therefore increase the system's insensitivity to noise, and so increase recognition accuracy (Gao et al,1992) .
An improvement on Linear Prediction utilising auditory peripheral knowledge is PLP or perceptually based linear prediction. This differs from standard LP by adding critical band integration, equal loudness pre-emphasis, and intensity to loudness compression before the standard inverse fourier transform. These simulate the properties of the auditory system, which, followed by an all-pole model obtain analysis parameters compatible with standard LP. However PLP has the advantage that because of its increased efficiency only 5 coefficients are usually required instead of LP's 15, this makes PLP as computationally efficient as LP, but provides dramatic savings on storage space. When applied to ASR PLP has been shown to

yield improvements when used with cepstral analysis (Hermanski et al,1985).
Another auditory phenomenon that is being introduced into periphery models is that of 'lateral inhibition', this is the suppression of the activity of nerve fibres on the basilar membrane caused by the activity of adjacent fibres. This accounts for the phenomenon caused when two tones of different amplitude are similar in frequency, in this situation the perception of the weaker tone will be inhibited. This has been used to improve noise resistance by convolving a frequency dependent lateral inhibition function with noisy speech (Cheng and O'Shaughnessy,1991). Because the narrow-band SNR is higher on spectral peaks then by emphasising these areas and attenuating spectral valleys then the SNR of the entire signal will hopefully increase.
On a different level, auditory models based on the temporal characteristics of nerve fibre firing rates have shown to offer robust representations for various recognition applications. One such model, the Ensemble Interval Histogram (EIH) (Ghitza,1986) consists of a set of cochlea filters followed by a zero crossing detector with a calculation of an interval histogram. Cochlea filters are equally spaced on a log-frequency scale, each of which are connected to a set of 7 level crossing detectors set at positive threshold levels uniformly distributed in the log scale. The multi-dimensional point output by the level-crossing detectors simulates auditory nerve firing patterns which is used to create a pseudo interval histogram for each fibre/filter. The relative spectral density of underlying frequency components can be located by finding regions of the fibre-array which fire synchronously. This can be calculated by finding the 'ensemble interval histogram' of the fibre array, the sum of all individual fibre histograms. Synchronisation is coded as corresponding height's in the EIH because every individual histogram in the same synchronisation region will contribute to each other. This representation has been found to have two main properties, fine spectral details are well preserved at low frequencies (but fuzzy at high frequencies), and increased robustness to noise compared with standard fourier analysis.
When used as a front end to a simple DTW recogniser the EIH-based recogniser was found to give comparable recognition rates to an FFT based front end with clean speech. However, as SNR decreases recognition scores with the EIH-based front end drop more slowly than the FFT based system, quantitatively the EIH based system achieves a given recognition score with global SNR values which are between 5dB and 15dB lower. However, when the cochlea filters used in EIH analysis were replaced with standard band-pass filters (Ghitza,1992) it was found that noise robustness was mainly due to the timing synchrony characteristics rather that to the shape of the filters.
A similar temporal method (Dobrin et al,1995) uses a two stage Periphery Auditory System (PAS) and Central Auditory System (CAS). The PAS uses the same array of cochlea filters, however with this technique, a non-linear model of the inner hair cell is used comprising of half-wave rectification, rapid adaptation, low-pass filtering, and fast adaptation. The output at this level simulates the instantaneous firing rate of the auditory nerve predicted for small populations of fibres. This information is then passed to the CAS providing a 'quasi-place-temporal representation' of the speech signal. The model used in this situation is the Coherence Co-operation Measure (CCM) (Wu et al,1990), found to be particularly effective at removing speaker-dependent information. The CCM operates by analysing the firing rate of the PAS using an N neuron neural network (N being the number of channels) with the *i*th neuron receiving input from channel CF*i* and +/- *v* neurons around CF*i*. For each neuron coherence is calculated by summing the cross-correlations between all fibres leading to the neuron. Results for digit recognition tests using both FFT and CCM recognition show that whilst CCM suffers a 5% drop in recognition rates with clean speech, at a noise rate of 0dB this changes to a 20% improvement.

Slow Variation Noise Filtering
One feature of many background noises and distortions found to occur alongside speech is that they vary slowly relative to speech. Therefore various schemes for removing slow variations in recognition feature vectors in an attempt to improve recognition accuracy. Applications of this filtering technique has been applied to a variety of parameters, from the log-power-spectrum to cepstral feature vectors. It is to the latter parameter that one of the simpler and more efficient methods is applied, know as CMN (Atal,1974) or Cepstrum Mean Normalisation. This involves removing the mean of all cepstral vectors and has found to improve recognition significantly, especially that of channel distortion (such as that caused by microphone

changes), without degrading the baseline system.
Another popular technique applied to this area is RASTA (Hermansky et al, 1991), suppressing constant additive offsets in every log spectral component of the short term spectrum. This can also be applied to mel-cepstral (Hirsch et al,1991) and PLP parameters (called RASTA-PLP) by filtering with a sharp spectral zero at zero frequency, making the average of each band also zero. These techniques are also particularly effective at reducing channel distortion, both from microphones and over telephone links. RASTA is particularly effective at reducing noise caused by such channel distortions, since they are additive in the log domain where RASTA operates. However signals which are additive over the time domain cannot be removed by this process, another technique called J-RASTA (Morgan and Hermansky,1992) is required. This consists of filtering the time series of a function of the spectrum, and applying its inverse on the result. If the noise is additive the function will be identity, if convolutional in nature, logarithmic. The shape of this function is dependent on *J*, which in turn is dependent on the additive noise level. This has been found to reduce both additive and convolutional noise increasing recognition rates for whole word models, however since time series are used then there is a time-memory effect which reduces its effectiveness with sub-word models.
In all, these techniques have been found to be most effective at reducing constant distortion noised caused by changes in channel. Unfortunately their efficiency reduces considerably when applied to other types of slow variance noise making these techniques rather limited to very specific noise.

Speech Enhancement

Mapping Transformation
In parametric space a speech signal can be represented as a point which moves as different sounds are produced by a speaker. An utterance can therefore be represented as a trajectory of these points across the chosen domain. Speech restoration can now be viewed as either a transformation of the noisy and clean trajectories to a reference trajectory (normalisation) or a transformation from the noisy/clean trajectory to the clean/noisy trajectory (mapping). In the area of speech enhancement it is the latter of these two transformations which is of interest to us, that is to transform noisy speech parameters to clean speech parameters.
One of the simplest forms of mapping is the linear transformation, here examples of the clean and noisy speech are aligned using the Dynamic Time Warping algorithm and a transformation found which would be useful in the case of both additive and Lombard noise. The simplest method for finding this transformation is that of minimising the mean square error between the parameter vectors. The most effective method is that of linear regression which takes a slightly different approach, that of the iterative mapping of clean speech onto noisy speech until a satisfactory result is obtained. This technique has been found to be highly effective, superior to both spectral mapping and a method for adjusting HMM state mean vectors (Mokbel et al,1992).
Multi-layer perceptrons can also be used for mapping transformations and have been used successfully to transform noisy cepstral parameters to their clean counterparts (Tamura and Nakamura,1990). Even for signals which differ from the training data in both the original speech input as well as the type of environmental noise a neural network has provided improvements in recognition performance. Since neural networks can achieve arbitrarily complex mappings, when compared to linear transformations they have shown to provide superior results (Mokbel et al,1992). Noise reduction using this technique has also been attempted on the time signal (Barbier and Chollet,1992), however whilst SNR was significantly improved recognition scores did not improve due to the distortion and occlusion of the speech signal.
Unfortunately, whilst performance can be impressive using ANN's for spectral mapping since there are no parametric models of speech or noise the performance is highly dependent whether the system is used with similar types and levels of noise. Another disadvantage this approach is that clean versions of all noisy speech are required for mapping which in applications such as telephony is not always available.

Spectral Subtraction
Spectral subtraction is a noise suppression technique used to reduce the effects of added noise in speech. It estimates the power of clean speech by explicitly subtracting the noise power from the noisy speech power.

This of course assumes that the noise and speech are uncorrelated and additive in the time domain. Also, as spectral subtraction techniques necessitate estimation of noise during pauses it is supposed that noise characteristics change slowly. However, because noise is estimated during speech noises this makes the method computationally efficient.
Unfortunately, for these reasons, spectral subtraction is beset by a number of problems. Firstly, because noise is estimated during pauses the performance of a spectral subtraction system relies upon a robust noise/speech classification system. If a misclassification occurs this may result in a mis-estimation of the noise model and thus a degradation of the speech estimate. Spectral subtraction may also result in negative power spectrum values, reset to non-negative values. This results in residual noise known as musical noise. Finally subtraction techniques cannot be used in the logarithmic spectrum domain because noise becomes signal dependent.
Spectral subtraction has been used for various kinds of applications as well as speech enhancement, such as recognition (Van Compernolle,1989), and speech coding. In a speech enhancement application it has been shown (Berouti et al,1979) that at a 5 dB SNR the quality of the speech signal is improved without decreasing intelligibility, however at lower SNR speech this reduces rapidly. When used in ASR the trade-off between SNR improvement and spectral distortion is important, although various attempts have been made to reduce musical noise. One method (Boll,1979) proposed a scheme where the frame-by-frame randomness of the noise is measured. For a given frequency bin, the residual noise is suppressed by replacing its current value with a minimum value chosen from the adjacent analysis frame. Combinations of spectral subtraction and other noise compensation techniques such as noise masking (Van Compernolle,1989), have also been applied and have shown to increase performance. Such a scheme (Lockwood and Boudy,1991) showed an increase in performance from 31% to 98% with such combinations for robust speech recognition in cars. Continuous spectral subtraction techniques have also been applied (Nolazco-Flores and Young, 1994) to avoid the problem of speech boundary detection in noise. With this method a smoothed estimate of the long term spectrum is continuously calculated and subtracted from the system. However, it still requires the detection of occasional periods of non-speech activity to update its noise HMM model.

Noise Masking
Noise masking is a psychological phenomenon of reduction of perceptibility of a signal in the presence of noise. That is, people cannot detect an acoustic stimulus whose level is lower than the masking threshold generated by other stimuli. When listening to speech in a noisy environment, the effect of noise can be decreased by the masking mechanism.
When implemented in ASR, its effect is to reduce the contribution of low energy regions in the discrimination process. Speech spectra are modified to immune the system from variations in background noise. This has been implemented as a noise floor normalisation (Klatt,1976) applied after a filter-bank analysis and log transformation. Only the frequency regions in the spectrum with an energy level higher than the masking level are then used in the recognition process. An improvement to this basic model used a masking level set at the maximum noise level found during the whole training process instead of at each template (Holmes and Sedgewich,1986). This method was also extended to the HMM framework and was shown to provide robust speaker-dependent digit recognition with SNR as low as 3 dB (Varga and Ponting,1989).
Noise masking has also been applied in transformed spectral space such as the cepstral domain. As the masking operation is a non-linear operation, it is carried out in the log-energy spectral domain before the cepstral transformation. When compared with to HMM decomposition, noise masking followed by recognition in the cepstral domain was found to give lower performance at very low SNR (less than 3 dB) (Mellor and Varga,1993). However it has been shown to be a good alternative at higher levels of SNR, while requiring lower computational complexity. An improvement to this approach was gained by adding a constant, independant of the noise level, to each spectral component to stabilise the cepstral space to changes in the presence of noise. At high masking levels important decreases in error rates were observed at low and medium SNR with performance reduced with clean speech.
As can be seen, noise masking is similar to spectral subtraction, since the thresholds correspond

approximately to the background noise level. Unfortunately this means that most of the problems that affect this method arise in noise masking. e.g. that of noise level estimation and residual noise.

Comb Filtering
If the period of noisy voiced speech can be determined, then comb filters can be applied in the frequency domain to increase the SNR. Comb filtering assumes that the noise is additive, and, due to the delay in pitch detection short-term stationary. It is also naturally limited to voiced speech, and will not be applicable to both unvoiced speech and speech segments with fast transitions or voiced fricatives. Comb filtering multiplies in the frequency domain the observation signal by a sequence of Dirac functions whose interval is the period of the speech signal. Naturally this approach depends upon accurate estimations of the period of the noisy speech which must be tracked as it varies over time, this can be difficult under noisy conditions
As speech is quasi-periodic with frequency modulation the spectral peaks of higher harmonics are broader and lower in amplitude compared to lower harmonics. To correct this non-stationarity in the periodicity comb filters can be applied to the time-domain re-warped signal. The clean speech estimate is then obtained by re-warping the comb filtered speech to re-introduce the quasi-periodicity of the original speech (Graf and Hubing,1993).
Comb filtering has been tested on the SNR of speech under a number of different types of noise (Lim and Oppenheim,1978) and has been found to improve SNR but at the cost of intelligibility. To minimise temporal distortions a dynamic time warping comb filter has been proposed (Graf and Hubing,1993) which warps adjacent pitch periods into the current pitch period to produce an enhanced pitch period. Performance was found to improve with this method at the cost of greater computational overhead.

Template Based Estimation
Template-based estimation involves finding the best sequence of clean speech templates to match noisy speech data. By using templates the signals can be restricted to a parameter subspace defined by the templates and the combination coefficients. An advantage of this system is that the output is almost noise free insofar as the noise influence is converted into incorrect estimation of templates.
There are a number of different approaches to template construction, phoneme-based, vector quantisation (VQ), and linear combination of sine-waves have all been used. For the purposes of improving noisy speech intelligibility VQ templates have been used for resynthesis (O'Shaughnessy,1988), with formant distance used as a similarity measure. The output of such processing is noise-free speech, with degradation appearing as a spectrum mismatch. One of the most successful applications of template estimation used linear combinations of clean speech templates for clean speech estimates. These were extracted from a set of phonemes, with combination coefficients based on the similarity between the noisy vector and noisy speech templates. Using phonemes as a recognition units a 206 word vocabulary under 10 dB Gaussian noise achieved a 90% recognition rate (Gong,1993) (95% using clean speech). However, training a base transformation under one type of noise cannot deal with different noise types or levels. An extension of the base transformation technique has been proposed which first recognises a specific noise category, then applies a base transformation relevant to that category (Treurniet and Gong,1994). The resulting system was tested on 10-40 dB SNR giving a flat recognition rate response. However, whilst this method outperformed both model compensation and training data contamination it still suffers from the problem of limited noise modelling, that is it will only work efficiently under a closed set of noise conditions.

Statistical Modelling
A popular method for coping with speech and noise is to use statistical methods for modelling some aspects of noise and variability. The basic idea is to take the presence of speech into account and to separately model an independent noise source as well as a speech source, assuming an additive relationship between the two sources.
To this end Bayesian approaches have been used with a number of distortion measures. The most common of these are the squared error function, leading to Minimum Mean Square Error (MMSE) estimation, and the uniform cost function, leading to Maximum A Posteriori (MAP) estimation. In MMSE estimation the

parameter values minimise the mean square error between the estimate and the actual value, whilst MAP estimation maximises the conditional probability of the parameters given the observation.
A great many applications have applied MMSE estimation methods, one of the most successful was applied to the estimation of filter bank log energies (Van Compernolle,1989). This was found to be superior to spectral subtraction methods, obtaining a significant increase in recognition rates. Improvements to this method were obtained by incorporating the correlation between different channels by means of conditioning the estimator on the total frame energy. Later this was extended by modelling the filter log energy vectors with a mixture of components representing speech classes (corresponding to a HMM state), making the assumption that spectral energies in different frequency channels are uncorrelated within each class (Erell and Weintraub,1993).
Another solution to speech enhancement is to take advantage of the capability of HMM to segment a speech utterance into quasi-stationary segments. This is based on statistical modelling of both clean and speech and noise, and because state-based statistics on noise are used this system can model non-stationary noise. A HMM-based MMSE estimator has been proposed (Ephraim,1990) whose estimation introduces pairs of states of signal models and noise models. The estimator is a weighted sum of Wiener filters of noisy speech which are conditioned on the pairs of states. The weights are the probabilities of speech-noise composite HMM states given the noisy observations. The MMSE determines the filter weights from the noisy speech. Since HMMs are used in the selection of composite states, the estimate exploits information on the neighbourhood of the analysed vector.
A MAP estimated approach implemented with the EM algorithm (Ephraim et al,1989) allowed the enhancement of all frames in the noisy utterance simultaneously. This consists of a maximisation of the log likelihood of the noisy speech over all sequences of states and mixture components. The estimation of the most likely sequence of states is carried out using the Viterbi algorithm, and the estimation of the clean speech vectors by applying Wiener filters on the noisy speech.
A comparison of both MAP and MMSE found that MMSE was superior in informal listening tests, and easier to implement. However, in the case of complex stochastic models for speech and noise the MAP estimator was necessary. Unfortunately one of the drawbacks of both these methods is the assumption that noise is both additive and statistically independent, therefore noise created by the Lombard effect will not be affected.

Computational Auditory Scene Analysis
Auditory Scene Analysis offers a general approach to the problem of extracting information concerning a sound source embedded in a noisy auditory background. The auditory background can be environmental noise, but also other speakers and music. One way of looking upon this problem is to consider the auditory scene as a mixture of several 'auditory objects' and to design a pre-processor where components are separated and grouped object by object before identification. This approach is generally more complex than the other methods described, however perceptual data shows that this is the strategy employed by human listeners (Bregman,1990). It is the area of 'grouping' the low level auditory objects which most attention lies. The acoustic properties and ASA models currently available suggest that onset-offset time, temporal dynamics of amplitude and frequency (e.g. pitch and formant trajectory), and spatial location are the most important features which govern auditory grouping decisions. However at the current time no common definition of the low-level auditory objects, and the rules governing their formation exist.
The grouping method which seems most natural is that of spatial location, where different sounds are can be grouped by their location. This area is split into two different approaches, the first restricts methods to those which have some basis in psychological and physiological evidence, and therefore detect spatial information from just two 'ears'. The second, and currently more successful area, is in the use of microphone arrays for the detection of spatial location. One such approach uses a 16 microphone array fixed to a wall for spatial localisation based on measuring time delays (Brandstein et al,1995). This is a frequency-domain based delay estimator shown to be capable of obtaining precision delay estimates over wide ranges of SNR conditions, whilst being computationally simple enough for real-time systems. The sound delay information can then be used to localise talker positions to within a few centimeters in diameter, and to track moving sources.
Another application of CASA takes advantage of two of the grouping cues described above, that of pitch and

AM (Berthommier and Meyer,1995). This approach is based upon the physiological experiments which suggest that the cells of the central nucleus of the inferior colliculus are sensitive to very selective amplitude modulation rates, with suggestions that neurones maybe organised into 'best stimulus' and 'best amplitude modulation' frequencies (Langner and Schreiner,1988). Therefore by modelling these effects (functionally) it is hoped that by the examination of long-term periodicity an efficient grouping mechanism will result, at least for the limited task of double vowel separation. From a functional point of view this task requires a precise estimate of AM frequencies to produce a 'AM Map' of the spectra, in this case obtained by Discrete Fourier Transform of the output of a band-pass filtered, half-wave rectified gamma-tone filterbank. The resulting `AM Map` produces a representation much like that of an autocorrelogram (frequency/delay/time), but with some degree of biological plausibility. The next stage uses a an harmonic sieve to produce a number of estimations for pitch, each of these estimates is then 'pooled' with the AM map to produce an estimate of the spectral energy related directly of harmonically with each candidate F0. The resulting spectra are then identified by use of a Neural Network, either an LVQ-Kohonen classifier or a Multi-Layer Perceptron. When tested with both white and AM noise the system provided favorably with a success rate of 60% at SNR -6 dB, on vowel/vowel recognition around 60% of the stimuli lead to recognition of both vowels, with at least one vowel detected 100% of the time. Moreover, this system has shown a hierarchy of spectral dominance of vowels where certain vowels were significantly easier to detect than others. This corresponds to the concept of 'dominance' of one source to another, opening the necessity of tracking all sources simultaneously.
Another approach to CASA follows the model for Auditory Scene Analysis described earlier more diligently than the scheme recently discussed. This method segregates the auditory scene into objects (in this case known as 'strands') before going on to a two-stage grouping process. As in the previous example the system uses a gamma-tone filterbank front end, modelling the auditory periphery by a bank of cochlea filters and simulations of neuromechanical transduction of inner hair cells. The output of the auditory periphery is then processed for the formation of the auditory objects, known as 'synchrony strands'. This process consists of three stages, initially dominant frequency estimation computes the most prominent frequency in the output of each filter. Cross-channel grouping takes place next, reducing the redundancy of dominant frequencies across channels to provide a summary of synchronous activity within a specific time frame. This results in a sequence of groups of channels with similar characteristics, known as place groups. Finally place-groups of aggregated over time to produce an explicit time-frequency description of auditory synchrony, resulting in the synchrony strands. The strands are then passed to the first stage of the grouping mechanism which uses cues such as common AM, harmonicity, common frequency modulation and movement, and onset and offset synchrony. This too is a two-stage process, the simultaneous stage examines all strands which overlap a starting (or seed) strand for similarities, those sufficiently similar are recruited into a group. The next stage, or sequential stage, chooses a strand in the recently constructed group which starts before, or finishes after the previous seed strands. This new seed is usually selected by its length of extension to the group, dominance of extension, or similarity to the current seed. This new strand now becomes the new seed for the group, and we return to simultaneous grouping once more. This continues until no more strands can be recruited into a group, whereupon the process starts again with a new unrecruited strand. When all strands have been grouped the second stage of grouping takes place, that of relating the groups found in the first stage. This higher-level grouping is produced from the derived properties of the low-level groups, in this case pitch contour. From this stage it is possible to listen to the results of grouping (or use them in ASR) by resynthesising the strands contained within the group. These have been found to both intelligible and preserve speaker-specific cues, however because high-frequency strands have a tendency to be rejected from groups (due to unreliable AM information) the sound sounds slightly muffled. Tests show that in clean conditions around four-fifths of the auditory scene is grouped, lower than expected due to the difficulties explained above. This drops to 67% to 78% with non-intrusive noise, remaining at this level for 6 out of the 10 tested noise types.

Speech Model Compensation for Noisy Environments

HMM Composition and Decomposition

The basic idea of signal decomposition is to recognise concurrent signals simultaneously. Parallel HMM's are used to model the concurrent signals and the composite model is modelled as the function of their combined outputs. This technique has been found to be particularly useful (Varga and Moore,1990) in the decomposition of speech and noise due to three main effects. Firstly, because the noise is modelled by a separate HMM various types of noise can be handled. Secondly, the technique does not suppose that the composite signal is derived from a particular type of signal combination, therefore additive as well as convolutional noise can be dealt with. Finally, the noise power is not assumed to have zero variance as in standard spectral subtraction schemes. Unfortunately due to its very nature this method is computationally expensive since the recognition is done by searching through a three-dimensional lattice state-space. This requires an extension of the standard Viterbi algorithm to three dimensional encoding. Therefore we move from an *N (states) x T (observations)* state model to an *N x M x T* model, where M is the number of states in the noise model. Experiments (Varga and Moore,1990) show that for both additive and non-stationary noises HMM decomposition significantly improves recognition of noisy speech over a standard HMM recogniser down to a SNR of -3dB. Unfortunately, because these improvements lie in noise modelling this system will only operate under the noise conditions in which it was trained, although a global noise model has been proposed (Gales and Young,1992).
Another technique similar to HMM signal decomposition is that of Parallel Model Combination (PMC) (Gales and Young,1993), however it differs in a few aspects from its predecessor. PMC is a composition method, so the noisy observation is modelled before recognition. It also operates in the cepstral domain, and, depending on noise variability its computational complexity can be significantly less than that on signal decomposition. When noise is stationary PMC can model noise using only a simple HMM model (1 state), however a non-stationary noise signal requires a more complex HMM with the optimal combination of speech and noise done at a decoding stage. When tested against the HMM decomposition technique PMC showed improved performance, with similar performance to non-linear spectral subtraction. Also, on the NOISEX-92 database PMC was found to be robust against additive noise down to 0dB for both stationary and nonstationary noise (Gales and Young,1993). This technique has also been used to compensate for speech distortions caused by speech enhancement techniques. For instance the HMMs of a recogniser were compensated by PMC for the signal distortion caused by a spectral subtraction stage (Kobayashiet al,1994).

State-dependent Wiener filtering
State-dependent Wiener filtering is a method for noise cancellation during recognition similar to the methods of Wiener filtering used for Speech Enhancement (see previous section). Previously Wiener filtering was limited to cancellation of noise in non-stationary speech signals. However by using HMM's to divide speech into quasi-stationary segments a Wiener filter can be used to implement noise cancelling filters within the speech recognition process. Basically this involves using a HMM state for each filter channel with a FIR Weiner filter attached as an additional parameter of the model. During recognition a filtered estimate of clean speech is calculated on the basis of a sequence on noisy input vectors. The estimate is then used to compute the output probability of the state.
This scheme can be employed in both the frequency and cepstral domain, however the most successful applications have used cepstral-time features giving recognition rates very close to matched test conditions down to 0dB for digit recognition (Vaseghi et al,1994).

Adaption of HMM Parameters and Duration models
Because of the performance degradation when HMM-based speech recognisers are introduced to noise a variety of adaption measures have been developed to improve the robustness of the input parameters. These include such methods as re-estimation of noise statistics, Bayesian Adaption, and linear transformation of HMM parameters.
For discrete HMM recognisers instead of using a Euclidean Distance measure a probabilistic mixture model based on a weighted sum of models from clean speech and noise has been introduced (Nadas et al,1989). This method models the energy of the noisy speech in each frequency band as the maximum of clean speech and noise energies. Variations in the noise characteristics are accounted for by continually adapting the noise

model, and by using statistical compensation the confusions which occur in noisy speech (not clean speech) can be accounted for. Unfortunately this method, whilst producing improvements in stationary noise, was found to be limited when dealing with impulsive noise.
Bayesian learning for HMM's (Lee at al,1991) was originally introduced for mismatch reduction between speakers, however it has been found that this method can also be applied to noisy speech or microphone adaption. By using several repetitions of the same word with this speaker adaption technique achieved better results for speaker-independent recognition than that speaker-dependent recognition using the same training data. The strength of this technique comes from the effective use of the set of trained models to derive new models through adaption.
As well as adapting HMM parameters improvements in HMM recognition rates can be gained by effective modelling of speech duration structures. It has been noted that when speech is produced at a low level in a noisy environment that whilst the spectral representation of speech may change significantly the duration structures remain intact. Therefore the use of statistics in HMM state duration modelling (Ferguson,1980) (Russell and Cook,1987) more robust towards robust noise than other parameters. In an experiment (Nicol et al,1992) duration information was modelled by splitting each HMM state into a number of sub-states. On a speaker-independent isolated word recognition task this model managed to reduce 60% of the errors for a variety of different noises.

Minimum Error Training
The usual method for training HMM's is Maximum Likelihood Estimation (MLE) where each model of a given class is trained independently, so no discrimination can be made between models can be made during training. This operates on the assumption that the models constitute a good representation of the data. However when small amounts of data are used during training the models are not correctly estimated. Therefore a mismatch occurs between what has been estimated and the models which are suitable for the training data. By replacing MLE criterion with a minimisation of mis-classifications (Minimum Error Training) or maximise the difference between the correct class the probability of the most probable incorrect class (Error Corrective Training) during training, error minimisation and discrimination can be improved. These are Minimum Error Classification (MEC) techniques, also called corrective or discriminative training. Basically this consists of an interative process to adjust the parameter's values as to make correct words more probable and incorrect words less probable. In the case of HMM based recognition these parameters are usually the means, variances, or weights of mixture densities, although the means and weights of mixture densities are found to be most effective.
The use of minimum error training for noisy speech recognition is thought to be particularly effective because, in theory, maximum likelihood training is optimal only if the probabilistic models fit the speech. This is effective with clean speech because the distribution can be assumed as Gaussian, unfortunately noise speech is usually far from Gaussian. Therefore the use of corrective or discriminative training is hoped to compensate for this deficiency. It has been found that in the case of isolated word recognition discriminative training can be highly effective in improving the recognition of noisy speech (Mizuta and Nakajima,1992). By training a HMM initialised on clean speech and continuing training using noisy speech (20dB) with minimum error training the MEC HMM had better phoneme and word error rates than the MLE algorithm even in a multi-style speaker training case.

Training Data Contamination
In order to reduce the mismatch between test data in noisy environments and speech models trained under clean conditions one solution is to add the noise experienced under test conditions to the training data. The use of such training data contamination has been shown to give good improvements in a number of recognition systems (Dautrich et al,1983)(Furui,1992). This approach is similar to the approach of spectral transformation of clean training data for the approximation of noisy environments. Unfortunately, whilst spectral transformation can take account of acoustic changes in the speakers voice due to the Lombard effect data contamination can only deal with additive noise.
However, an approach to redress this shortcoming has been to automatically convert the training data to

stressed speech by various distortion measures. Recognisers trained using the contaminated stressed speech (Bou-Ghazale and Hansen,1994) have shown that this can reduce error-rates caused by the Lombard effect, albeit only for isolated word recognition.
The use of data contamination can also be used to help learning algorithms to perform better recognition. The use of discriminative HMM training with data at different SNRs (Mitzuta and Nakajima,1992) have shown to produce a robust recognition system to noises with various spectral characteristics.
Improvements in the use of data contamination have also come about by varying the SNR ratio of the speech used in training (Tsuboi et al,1990). By starting the training process using clean speech and gradually increasing the SNR of training data, a process called noise immunity learning, has been tested successfully in both word-spotting and speaker-dependent isolated word recognition experiments (Sankar and Patravali,1994).
Unfortunately all data contamination is by its nature restricted to specific noisy environments. A system trained using one type of noise will degrade significantly under even small changes in the environment characteristics (Kitamura et al,1992). Approaches to generalise noise characteristics have been attempted (Lippman et al,1987) by using several different types of noise, trained using multi-speaker style training procedures. However even these approaches are still extremely limited in their operating environments making data contamination only applicable to situations where noise levels and characteristics are relatively constant.

References

Leonard,R.G. IEEE International Conference on Acoustics, Speech, and Signal Processing, Paper 42.11. 1984

Paul,D. Proceedings of the DARPA Speech and Natural Language Workshop,pp.7-14,1992.

P.Lockwood and J.Boudy,'Experiments with a Non-linear Spectral Subtractor, HMM's and the projection, for robust speech recognition in cars', Vol.11,Nos.2-3,pp.215-228.1992

Jean-Claude Lunqua,'The Lombard Reflex and its role on human listeners and automatic speech recognisers',JASA 93(1) Jan 1993 p510-524.

D.C.Bateman,D.K.Bye, and M.J.Hunt (1992),'Spectral contrast noralisation and other techniques for speech recognition in noise',Proc. IEEE. Internat. Conf. Acoust. Speech Signal Process., Vol. I,pp.241-244,1992.

B.Atal,'Effectiveness of Linear Prediction Characteristics of the speech wave for automatic speaker identification and verification',J.Acost.Soc.Amer.,Vol.55,pp.1304-1312.1974.

H.Hermansky,N.Morgan,A.Bayya and P.Kohn,'Compensation for the effect of communication channel in auditory-like analysis of speech (RASTA-PLP)',Proc.European Conf. on Speech Technology,Genova,Italy.pp. 1367-1371.1991

H.G.Hirsch,P.Meyer,H.W.Ruehl,'Improved Speech Recognition using high-pass filtering of subband envelopes',Proc.European.Conf.Speech Technology, Genova,Italy.pp-413-416,1991.

K.K.Paliwal,'Neural Net classifier for robust speech recognition under noisy Environments',Proc.IEE Internat.Conf.Acoust.Speech Signal Process.,Albuquerque,NM,pp.429-432.1990.

Y.Anglade,D.Fohr and J.C.Junqua,'Speech descrimination in adverse conditions using acoustic knowledge and selectively trained Neural
Networks'Proc.IEE.Internat.Conf.Acoust.Speech.Signal.Process.,Vol.II,pp.985-988.1993.

A.Erell and M.Weintraub,'Filter-bank-energy estimation using mixture and Markov models recognition of noisy speech',IEEE.Trans.Speech.Audio.Process.,Vol.SAP-1,No.1,1993.

N.Nocerino,F.K.Soong,L.R.Rabiner and D.H.Klatt,'Comparative study of several distortion measures for speech recognition',Proc.IEEE,Internat.Conf.Acoust.Speech Signal Process.,pp-25-28,1985.

S.Nakamura,T.Akabane, and S.Hamaguchi,'Robust word spotting in adverse car environments',Proc.European Conf. Speech Technology,Berlin,Vol.2,pp.1045-1049.1993.

D.Mansour and B.H.Juang,'The short-term modified coherence representation and its application for noisy speech recognition',IEEE.Trans.Acoust.Speech.Signal.Process.,Vol.37,pp.795-804.1989.

J.G.Mena,L.S.Sandoval,R.G.Gomez,'A comparative study of feature extraction methods for noisy speech recognition',Speech Signal Processing V: Theories and Applications ed. by L.torres, E.Masgrau, and M.A.Lagunas,pp.1191-1194.1990.

Y.Gao,T.Huang,S.Chen,J-P,Haton,'Auditory model based on speech processing',Internat.conf.Speech.Lang.Process.,Banff,Alberta,Canada,Vol.1,pp.73-76,1992.

H.Hermanski,B.A.Hanson,H.Wakita,'Perceptually based linear predictive analysis of speech',Proc.IEEE Internat.Conf.Acoust.Speech Signal Process.,Tampaa,Vol.1,oo.509-512.1985.

Y.M.Cheng and D.O'Shaughnessy,'Speech Enhancement based conceptually on auditory evidence',IEEE.Trans.Acoust.Speech.Signal.Process.,Vol.39,No.9,pp.1943-1954.1991.

O.Ghitza,'Auditory nerve representation as a front-end for speech recognition in a noisy environment',Comp.Speech and Language,Vol.1,pp.109-130,1986.

O.Ghitza,'Auditory nerve representation as a basis for speech processing',in ADvances in Speech Signal Processing by S.Furui and M.M.Sondhi (Marcel Dekker,New York),Chapter 15,pp.453-485.

C.Dobrin,P.Haavisto,K.Laurila,J.Astola,'Speech Recognition Experiments in a Noisy Environment using Auditory System Modelling',Eurospeech 95 V1.p131,1995.

Z.L.Wu,J-L.Schwartz,P.Escudier,'Modelling spectral processing in the central auditory system',Proceedings ICASSP,pp.373-376.1990.

B.Dautrich,L.Rabiner,T.Martin,'On the effects of varying filter-bank parameters on isolated word recognition',IEEE Trans. ASSP,ASSP-31(4),pp.793-806,1983.

S.Furui,'Toward robust speech recognition under adverse conditions',ESCA Workshop Proc. Speech Processing in Adverse Conditions,Cannes,France,pp.31-41,1992.

H.Tsuboi,H.Kanazawa,Y.Takebayashi,'An accelerator for high-speech spoken word spotting and noise immunity learning system', In ICSLP,pp.273-276.1990.

R.Sankar,S.Patravali,'Noise immunisation using neural net for speech recognition',In ICSLP,pp.II.685-II.688.1994.

T.Kiatmura,S.Ando,E.Hayahara,'Speaker independant spoken digit recognition in noise environments using dynamic spectral features and neural networks',Internat.Conf.on Speech and Language Processing,October 1992,Vol.1,pp.699-702.

R.P.Lipmann,E.A.Martin,D.B.Paul,'Multi-style training for robust isolated word speech recognition',Proc.IEEE Internat. Conf. Acoust. Speech Signal Process.,April 1987,pp.705-708.

S.Mizuta,K.Nakajima,'Optimal discriminative training for HMM's to recognise noisy speech',Internat.Conf.on Speech and Language Processing,Vol.II,October 1992,,pp.1519-1522.

C.H.Lee,C.H.Lin,B.H.Juang,'A study on speaker adaption of the parameters of continuous density hidden Markov models',IEEE.Trans.Signal Process.,Vol.39,No.4,April 1991,pp.806-814.

J.D.Ferguson,'Variable duration models for speech',Proc.Symp.on the Applications of Hidden Markov Models to Text to Speech,IDA-CRD,pp.143-179,1980.

M.J.Russell,M.Cook,'Experimental evaluation of duration modelling techniques for automatic speech recognition',Proc.IEEE Internat.Conf.Acoust.Speech Signal.Process.,pp.2376-2379,1987.

N.Nicol,S.Euler,M.Falkhausen,H.Reininger,D.Wolf,J.Zinke,'Improving the robustness of automatic speech recognisers using state duration information',ESCA Workshop Proc. Speech Processing in Adverse Conditions,pp.183-186.1992.

A.P.Varga,R.K.Moore,'Hidden Markov Model decomposition of speech and noise',Proc.IEEE.Internat.Conf.Acoust.Speech.Signal.Process.,pp.845-848.1990.

M.J.F.Gales,S.J.Young,'An improved approach to the hidden Markov model decomposition of speech and noise',Proc. IEEE. Conf. Acoust. Speech Signal Process.,Vol.1,pp.223-236.April,1992.

M.J.F.Gales,S.J.Young,'Cepstral parameter compensation for HMM recognition in noise',Speech Communication,Vol.12,No.3,pp.231-239.1993.

T.Kobayashi,R.Mine,K.Shirai,'Markov model based noise modelling and its application to noisy speech recognition using the dynamical features of speech',Proc.IEEE.Internat.Conf.Acoust.Speech Signal Process.,April 1994,pp.57-60.

S.V.Vaseghi,B.P.Milner,J.J.Humphries,'Noisy speech recognition using cepstral-time features and spectral-time features',Proc.IEEE.Internat.Conf.Speech Technology,Berlin,1993,Vol.II,pp.65-68.

C.Mokbel,L.Barbier,Y.Kerlou,G.Chollet,'Word Recognition in the car; adapting recognisers to new environments',Internat.Conf.Speech Language Processing,Alberta,October 1992,Vol.1,pp.701-710.

S.Tamura,M.Nakamura,'Improvements to noise reduction neural networks',ICASSP 1990,pp.825-828.

D.Van Compernolle,'Noise adaption in hidden Markov model speech recognition system',Computer Speech and Language,3(2),pp.151-167.

M.Berouti,R.Schwartz,J.Makhoul,'Enhancement of speech corrupted by acoustic noise',ICASSP 1979,pp.280-211.

S.Boll,'Suppression of acoustic noise in speech using spectral subtraction',IEEE.Trans. ASSP, ASSP-27(2),pp.113-120.1979.

P.Lockwood and J.Boudy,'Experiments with a Non-linear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars',Eurospeech 1991,pp.79-82.

J.Nolazco-Flores,S.Young,'Continuous speech recognition in noise using spectral subtraction and HMM

adaptation',ICASSP 1994,pp.I.409-412.

D.Klatt,'A digital filter bank for spectral matching',ICASSP 1976,pp.573-576.

J.Holmes,N.Sedgewick,'Noise compensation for speech recognition using probabalistic models',ICASSP 1986,pp.741-744.

A.Varga,K.Ponting,'Control experiments on noise compensation in hidden Markov model based continuous word recognition',EUROSPEECH 1989,pp.167-170.

B.Mellor,A.Varga,'Noise masking in a transform domain',ICASSP 1993,pp.II.87-90.

J.T.Graf,N.Hubing,'Dynamic Time Warping for the enhancement of speech degraded by white Gaussian noise',Proc.IEEE Internat. Conf. Acoust. Speech Signal Processing.,1993,Vol.II,pp.339-342.

J.S.Lim,A.V.Oppenheim,'Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition',IEEE.Trans. Acoust. Speech Signal Processing.,Vol.26,No.4,pp.354-358.1978.

D.O'Shaughnessy,'Speech enhancement using vector quantisation and a formant distance measure',Proc. IEEE Internat. Conf. Speech Signal Process.,1988,pp.549-552.

Y.Gong,'Base transformation for environment adaption in continuous speech recognition',Proc.European Conf. Speech Communication and Technology,Berlin,1993,Vol.3,pp.2227-2230.

W.C.Treuniet,Y.Gong,'Noise independant speech recognition for a variety of noise types',IEEE Internat. Conf. Acoust. Speech Signal Process.,1994,Vol.1,pp.437-440.

D.Van Compernolle,'Spectral estimation using a log-distance error criterion applied to speech recognition',ICASSP 1989,pp.845-848.

A.Erell,M.Weintraub,'Filter-bank energy estimation using mixture and Markov models for recognition of noise speech', IEEE Trans. on Speech and Audio Processing,1(1):pp.68-76,1993.

Y.Ephraim,'A minimum mean square error approach for speech enhancement',ICASSP 1990,pp.829-832.

A.S.Bregman,'Auditory Scene Analysis',MIT Press,London.1990.

M.S.Brandstein,J.E.Adcock,H.F.Silverman,'A practical time-delay estimator for localising speech sources with a microphone array',Computer Speech and Language 9,pp.153-169.1995.

F.Berthommier,G.Meyer,'Source separation by a functional model of amplitude modulation',Eurospeech 95,pp.135-138.

G.Langner, C.E.Schreiner,'Periodicity coding in the inferior collinculus of the cat I: Neuronal mechanisms',J.Neurophysiol.,60,pp.1799-1822.1988.

M.Cooke,'Modelling Auditory processing and organisation',PhD Thesis, Distinguished Dissertations in Computer Science, Cambridge University Press,1993.

G.J.Brown,M.Cooke,'Computational auditory scene analysis',Computer Speech and Language 8,pp.297-336.1994.

Makhoul and Richard Schwartz, 'State of the Art in Continuous Speech Recognition', Proc.Natl.Acad.Sci.USA, Vol.92,pp.9956-9963,1995.

Yifan Gong,'Speech Recognition in Noisy Environments: A Survey',Speech Communication 16,pp.261-291,1995.

Jean-Claude Junqua and Jean-Paul Haton,'Robustness in Automatic Speech Recognition - Fundamentals and Applications',Kluwer Academic Publishers, 1996.