TECHNIQUES FOR HANDLING CONVOLUTIONAL DISTORTION WITH 'MISSING DATA' AUTOMATIC SPEECH RECOGNITION

Kalle J. Palomäki[1,2], Guy J. Brown[2] and Jon Barker[2]

[1]Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing

P.O. Box 3000, FIN-02015 HUT, Finland

[2]Department of Computer Science, University of Sheffield, Regent Court,

211 Portobello Street, Sheffield S1 4DP, United Kingdom

Email: kalle.palomaki@hut.fi, g.brown@dcs.shef.ac.uk, j.barker@dcs.shef.ac.uk

Corresponding author:

Kalle Palomäki

Helsinki University of Technology,

Laboratory of Acoustics and Audio Signal Processing

P.O. Box 3000, FIN-02015 HUT, Finland

Email: kalle.palomaki@hut.fi

Tel: +358-9-451 2883

Fax: +358-9-460 224

ABSTRACT

In this study we describe two techniques for handling convolutional distortion with 'missing data' speech recognition using spectral features. The missing data approach to automatic speech recognition (ASR) is motivated by a model of human speech perception, and involves the modification of a hidden Markov model (HMM) classifier to deal with missing or unreliable features. Although the missing data paradigm was proposed as a means of handling additive noise in ASR, we demonstrate that it can also be effective in dealing with convolutional distortion. Firstly, we propose a normalisation technique for handling spectral distortions and changes of input level (possibly in the presence of additive noise). The technique computes a normalising factor only from the most intense regions of the speech spectrum, which are likely to remain intact across various noise conditions. We show that the proposed normalisation method improves performance compared to a conventional missing data approach with spectrally distorted and noise contaminated speech, and in conditions where the sound level of the acoustic input varies. Secondly, we propose a method for handling reverberated speech which attempts to identify time-frequency regions that are not badly contaminated by reverberation and have strong speech energy. This is achieved by using modulation filtering to identify 'reliable' regions of the speech spectrum. We demonstrate that our approach improves recognition performance in cases where reverberation time T_{60} varies from 0.7 sec. to 1.5 sec., compared to a baseline system which uses acoustic features derived from perceptual linear prediction and the modulation filtered spectrogram.

Keywords: speech recognition, missing data, spectral distortion, spectral normalisation, reverberation.

1. INTRODUCTION

Although much research effort has been expended on the development of automatic speech recognition (ASR) systems, their performance still remains far from that of human listeners. In particular, human speech perception is robust when speech is corrupted by noise or by other environmental interference, such as reverberation or a poor transmission line (for example, see Assmann and Summerfield, 2003; Nabelek and Robinson, 1982). In contrast, ASR performance falls dramatically in such conditions (for a comparative review of human and automatic speech recognition performance in noise see Lippmann, 1997). As several researchers have observed (e.g., Cooke *et al*., 2001; Hermansky 1998; Lippmann, 1997), the current limitations of ASR systems might reflect our limited understanding of human speech perception, and especially our inadequate technological replication of the underlying processes.

The robustness of human speech perception can be attributed to two main factors. First, listeners are able to segregate complex acoustic mixtures in order to extract a description of a target sound source (such as the voice of a speaker). Bregman (1990) describes this process as 'auditory scene analysis'. Secondly, human speech perception is robust even when speech is partly masked by noise, or when parts of the acoustic spectrum are removed altogether (for example, by a bandlimited communications channel). Cooke *et al.* (2001) have interpreted this ability in terms of a 'missing data' model of speech recognition, and have adapted a hidden Markov model (HMM) classifier to deal with missing or unreliable features. In their system, a time-frequency 'mask' is employed to indicate whether acoustic features are reliable or corrupted; according to this division the features are treated differently by the recogniser. Typically, the missing data mask is derived from auditory-motivated processing, such as pitch analysis (Barker *et al.*, 2001a; Brown *et al*., 2001) or binaural spatial processing (Palomäki *et al*., 2001; in press). Alternatively, the mask can be set according to local estimates of the signal-to-noise ratio (SNR) (Cooke *et al*., 2001).

The missing data paradigm was conceived by Cooke *et al*. as a means of dealing with *additive* noise in ASR. As a result, little consideration has been given to the ability of missing data ASR systems to handle interference caused by the interaction of a target sound with its environment (such as a

transmission line, audio equipment or reverberant space). In terms of signal theory this is regarded as *convolutional* interference. In this paper, we propose a number of modifications to a missing data ASR system which allow it to perform robustly in the presence of convolutional noise.

A convolutional interference can be characterised by the impulse response of the corresponding system. If the impulse response is relatively short compared to the wavelength of the speech sound, this mainly causes spectral alteration, since convolution in the time domain is equivalent to multiplication in the frequency domain. If, however, the impulse response is relatively long (which is true in the case of reverberation) the interaction is of a different nature. A typical room impulse response consists of sparse early reflections followed by dense late reverberation (higher order reflections), which forms the exponentially decaying tail of the response. The sparse early reflections are highly correlated with the speech signal and actually contribute usefully to speech intelligibility by increasing the loudness of the speech. However, the dense late reverberation is poorly correlated with the original speech signal and therefore behaves more like additive noise. Reverberation also alters the spectral shaping of the speech signal, because different vibrational modes of the room emphasize some frequencies more than the others. Further details of the effect of room acoustics on speech intelligibility can be found in Bradley (1986) and Houtgast and Steeneken (1985).

The conventional way of tackling convolutional interference in ASR has been to use cepstral encoding, and to employ cepstral mean subtraction to remove the spectral distortion. Two common examples of cepstral encoding are mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980) and cepstral features obtained by perceptual linear prediction (PLP) (Hermansky, 1990). Interestingly, both of these approaches are loosely based on known mechanisms of auditory frequency encoding. However, they have been found to perform inadequately with reverberated speech (Hermansky, 1990; Kingsbury, 1998; Kingsbury *et al.*, 1998). Reverberation can also be handled via blind source separation (BSS) using a microphone array, or via blind deconvolution or dereverberation (for an overview see Omologo *et al.*, 1998). In such approaches, the aim is to enhance subjective speech quality rather than to find a robust acoustic encoding. BSS gives good dereverberation performance, but

at least two microphone signals are needed to process a single speech source (for an overview of BSS and independent component analysis see Hyvärinen *et al*., 2001).

Kingsbury and his colleagues (Kingsbury, 1998; Kingsbury *et al*., 1998) have reported that a modulation filtered spectral representation, the modulation spectrogram (MSG), can improve ASR performance with reverberated speech. Spectral bands are processed by a modulation filter, which emphasizes the strongest speech modulations and effectively removes reverberant or noisy regions that are not modulated in the same way as speech signals. This approach is consistent with studies that demonstrate the importance of low frequency modulations in human speech recognition (Houtgast and Steeneken, 1985; Drullman *et al.*, 1994).

In this study we address the problem of handling convolutional distortion in a missing data ASR system which uses spectral speech features. Two conditions are considered; one in which speech is subject to spectral distortion and additive noise, and another in which speech is reverberated. In the first case, we derive a missing data mask from estimates of the SNR in local time-frequency regions, and employ spectral subtraction to remove the noise background. Furthermore, we introduce a new method for normalising spectral features that is compatible with the missing data ASR framework. In reverberant conditions, a modulation filtering scheme is used to generate the missing data mask. This approach exploits temporal modulations of speech in order to find spectro-temporal regions which are not severely contaminated by reverberation.

The current study extends our previous work in several important respects. A related scheme for spectral normalisation was presented in (Palomäki *et al*., in press), but it was applied only to a very specific purpose (speech recognition using a binaural hearing model). Here, we develop and evaluate the normalisation scheme more thoroughly, and evaluate it on a more general speech recognition task with different types of spectral distortion. Our early work on modulation mask estimation (Palomäki *et al*., 2002) suffered from the drawback that the algorithm needed to be hand-tuned to each different reverberation condition. This problem has now been addressed by an adaptive scheme, in which the parameters of the algorithm are set according to an estimate of the degree of reverberation present in the signal. This allows the same system to be used in a wide range of reverberation conditions without the

need for hand-tuning. Finally, in (Palomäki *et al*., 2002) the system was evaluated on a limited number of simulated room impulse responses (RIRs), whereas here we use real RIRs which vary in their T_{60} reverberation time between 0.7 sec. and 1.5 sec. The results obtained with our new method are also compared against Kingsbury's (1998) recogniser for reverberated speech, which uses MSG and PLP features.

Section 2 of the paper describes the overall architecture of the missing data ASR system and the acoustic features used. In Section 3, we present a processing pathway that is optimised for conditions in which speech is subject to spectral distortion and additive noise. A processing pathway for reverberant conditions is described in Section 4. The system is evaluated under a number of noise conditions in Section 5, and compared against a baseline approach. We conclude with a discussion in Section 6.

2. SPEECH RECOGNISER

The missing data speech recognition system is shown schematically in Figure 1. In this section we describe the front-end processing, which extracts spectral features using an auditory model, and explain the missing data ASR approach.

– Figure 1 about here –

2.1. Acoustic features

Typically, HMM-based ASR systems model each state as a mixture of Gaussians with diagonal covariance, and therefore assume that the acoustic features are statistically independent. Cepstral features are widely used because they meet this requirement, since they are an approximately orthogonal encoding of spectral shape (see Gold and Morgan (2000) for a review). Additionally, cepstral mean subtraction can be employed to deal with spectral distortion (Atal, 1974).

However, in the context of missing data ASR there are good reasons for using an acoustic encoding based on spectral features, rather than cepstral coefficients. Firstly, noise that is local in frequency only disrupts local spectral features, whereas it is distributed over a wide range of features in the cepstral domain (Morris, 2002; see also Droppo *et al*., 2003). Furthermore, mask estimation techniques which

are based on our understanding of human perception are most naturally implemented in terms of spectral features, because the peripheral auditory system decomposes sound into frequency bands (Moore, 2003).

– Figure 2 about here –

Here, we derive spectral acoustic features for the recogniser from a simple model of peripheral auditory processing. Cochlear frequency analysis is simulated by a bank of 32 bandpass 'gammatone' filters, with centre frequencies spaced on the equivalent rectangular bandwidth (ERB) scale between 50 Hz and 3850 Hz (for details see Cooke, 1993; Brown and Cooke, 1994). The instantaneous Hilbert envelope is computed at the output of each filter. This is smoothed by a first-order low-pass filter with an 8 ms time constant, sampled at 10 ms intervals, and finally cube root compressed to give a crude simulation of auditory nerve firing rate (a 'rate map'; see Figure 2 for an example). Here, we use the notation $y(i, j)$ to denote the value of the rate map for auditory filter channel *j* at time frame *i*.

2.2. Missing data speech recognition

Automatic speech recognition is a classification problem in which an observed acoustic vector Y must be assigned to a class of speech sound C. Using Bayes' rule, the posterior probability $f(C|Y)$ can be expressed as the product of a likelihood $f(Y|C)$ and a prior $f(C)$, and hence classification can be performed by finding the class C which maximises $f(Y|C)f(C)$. However, when noise is present some elements of the acoustic feature vector Y may be unreliable or missing, and it is not possible to compute $f(Y|C)$ in the usual manner. One solution to this problem is the 'missing data' technique (Cooke *et al*., 2001). This addresses the problem by partitioning Y into reliable and unreliable components, Y_r and Y_u. The reliable components Y_r are directly available to the classifier in the form of the marginal distribution $f(Y_r|C)$. Additionally, the true value of the unreliable features Y_u can often be assumed to lie within a certain range. This provides an additional constraint by bounding the range of possible values over which the unreliable features are integrated. This technique is known as 'bounded marginalisation' (Cooke *et al*., 2001).

Here, we use bounded marginalisation where Y is a vector of simulated auditory nerve firing rates; thus the lower bound of Y_u is zero (since a firing rate cannot be negative) and the upper bound is the observed firing rate. In practice, a 'mask' $m(i,j)$ is used to indicate whether the acoustic evidence in each time-frequency region is reliable. In the simplest case, mask values are taken to be 0 or 1 so that a binary judgement is made as to whether data is reliable or unreliable. Alternatively, the mask elements may be set to real values in the range [0,1] to give soft reliability decisions rather than binary ones (Barker *et al.*, 2000b). In this case the equations for the bounded marginalisation computation are rewritten so as to effectively interpolate between the two interpretations of each acoustic feature (i.e. the interpretation that the feature is reliable, and the interpretation that the feature is unreliable).

In this study, auditory rate maps were used to train a missing data ASR system for recognition of connected digit strings (such as "three five six zero"). Twelve word-level HMMs were trained (a silence model, 'oh', 'zero' and '1' to '9'), each consisting of 16 no-skip, straight-through states with observations modelled by a 7 component diagonal Gaussian mixture.

3. PROCESSING FOR SPECTRAL DISTORTION AND ADDITIVE NOISE

In this section we describe a processing pathway that compensates for spectral distortion and additive noise. Our approach is based on the combination of three techniques; estimation of a missing data mask on the basis of SNR in local time-frequency regions (Section 3.1), spectral subtraction (Section 3.2) and an approach to spectral feature normalisation which is suitable for missing data ASR in the presence of additive noise (Section 3.3).

3.1. SNR mask estimation

If an estimate of the noise spectrum is available, the local SNR in each frequency channel of the rate map at each time frame can be used to derive a missing data mask. Local time-frequency regions with a high SNR (i.e., dominated by speech) are labelled as reliable in the mask, and those with a low SNR are labelled as unreliable.

Following previous work (Cooke *et al.*, 2001) we compute the local SNR from stationary noise estimates, which are obtained by averaging the acoustic spectrum over a short period in which speech

is believed to be absent. Specifically, we estimate the noise spectrum from the first *K=10* frames (i.e., 100 ms) of the rate map,

$$z(j) = \frac{1}{K}\sum_{i=1}^{K} y_e(i, j) \tag{1}$$

where $y_e(i, j) = y(i, j)^3$ and $z(j)$ is the noise estimate for frequency channel *j*. Note that $z(j)$ is estimated from a version of the rate map, $y_e(i, j)$, to which cube root compression has not been applied. The noise estimate is used to calculate a local SNR $s(i, j)$

$$s(i, j) = 20\log_{10}\left(\left[\frac{y_e(i, j) - z(j)}{z(j)}\right]\right) \tag{2}$$

which is subsequently used to estimate the missing data mask. Here, we employ a 'soft' mask in which each value is a real number in the range 0 to 1 (Barker *et al*., 2000b). Such masks can be interpreted as giving the probability that each time-frequency region is dominated by the speech signal. The mask values are produced by passing each local SNR estimate $s(i, j)$ through a sigmoidal function $\sigma(\ )$, i.e.

$$m_s(i, j) = \sigma[s(i, j)] = \frac{1}{1 + \exp\{-\alpha[s(i, j) - \beta]\}} \tag{3}$$

where $m_s(i, j)$ is the mask value for channel *j* at time frame *i*, α is the slope of the sigmoid and β is its centre point. Note that time-frequency regions with a higher local SNR are assigned a higher value in the mask. The values of the parameters α and β were found empirically (Barker *et al*., 2000b). Note that for $\alpha = 0$ all mask values are 0.5, indicating complete uncertainty about the signal and noise. With increasing α the sigmoid (3) becomes steeper, so that the decision between clean and noisy data approaches a binary one. Here, we use $\alpha = 3$ and $\beta = 0.4$.

3.2. Spectral subtraction

The missing data approach aims to identify speech features which are relatively uncontaminated by noise, and to pass these 'reliable' features to the speech recogniser. In practice, even acoustic features which are classified as reliable by the mask estimation process will contain some degree of noise, and

hence there will be a mismatch between the observed acoustics and models trained on clean speech. This mismatch can be reduced by subtracting the noise estimate $z(j)$ from the observed (uncompressed) noisy features $y_e(i, j)$. The 'cleaned' rate map is therefore given by:

$$y_s(i, j) = \sqrt[3]{[y_e(i, j) - z(j)]^+} \tag{4}$$

The operator $[\ ]^+$ denotes half-wave rectification; this ensures that $y_s(i, j)$ contains only positive values. Note that spectral subtraction is performed on the uncompressed rate map, which is subsequently compressed (cube rooted) before passing to the recogniser.

3.3. Normalisation

Conventionally, spectral features are normalised by the mean and variance in each frequency band (for example, see Kingsbury *et al*., 1998). A problem with this approach is that clean regions of the speech signal may be normalised by a mean and variance that are computed when both speech and noise sources are present. This is particularly harmful in missing data ASR, which requires that reliable features presented to the recogniser should be scaled in the same way as the clean speech features used for training.

Here, we take a different approach in which a normalisation factor is computed only from those acoustic features that are likely to be dominated by speech (i.e., uncorrupted by noise). Scaling based only on these regions is likely to reduce the mismatch between the clean training and noisy recognition conditions. Of course, this normalisation technique requires that speech-dominated features can be identified in approximately the same way during training and recognition. Fortunately, this is achievable in practice, as illustrated by the plots of speech-dominated regions for clean and noisy rate maps shown in Figure 3.

– Figure 3 about here –

Here, we use a simple implementation of this scheme in which the acoustic features in each channel are normalised by the mean of the *L* largest features in that channel. We compute the normalisation factor $\eta_s(j)$ for channel *j* as follows,

$$\eta_s(j) = \frac{1}{L} \sum_{i \in \Gamma_s(j)} y_s(i, j) \tag{5}$$

where $y_s(i, j)$ is the 'cleaned' rate map and $\Gamma_s(j)$ is a set containing the indices of the L largest values of $y_s(i, j)$ in channel j. The rationale for Equation 5 is that selection of the L largest values in each channel of the rate map gives a comparable result with clean (training) and noisy (recognition) data, so long as the noise is fairly stationary and the global SNR is favourable. Here, we set L to I/D, where I is the number of time frames in the input and D is a constant (we use D=5).

The value of L must be set empirically, and depends on two conflicting constraints. Firstly, L should be chosen small for good performance in very noisy conditions, since relatively few features in the rate map will be reliable. On the other hand, if L is too small then a stable estimate of the normalisation factor cannot be obtained.

4. PROCESSING FOR REVERBERATION

This section describes a processing pathway for missing data ASR in reverberant conditions (see Figure 1). In the first stage, modulation filtering is used to derive a mask that identifies the speech features that are least contaminated by reverberation. Following this, spectral features are normalised using a modification of the technique described in Section 3.3.

4.1. Reverberation mask estimation

Previously, Kingsbury *et al*. (1998) have shown that modulation filtering can be used to derive robust features for speech recognition in the presence of reverberation. Here, we use modulation filtering in a different way. Specifically, it is used to identify spectro-temporal regions that contain strong speech energy (i.e. regions that are not badly contaminated by reverberation), and hence to derive a 'reverberation mask' for missing data ASR using spectral features. We use a modulation filter $h(n)$ of the following form, where the time index n is measured in frames (see Section 2.1):

$$h(n) = h_{lp}(n) \otimes h_{diff}(n) \tag{6}$$

This is a finite impulse response (FIR) filter consisting of a linear phase lowpass component h_{lp} and a differentiator h_{diff} (the operator $\otimes$ denotes convolution). The lowpass part h_{lp} was designed using the MATLAB `fir2` command (Mathworks, 2003). The filter $h(n)$ has a pass band between D.C. and 17 Hz (see Figure 4), and it is used to derive a modulation-filtered rate map $y_r(i, j)$ by filtering each channel j of $y(i, j)$ as follows:

$$y_r(i, j) = \sum_{k=-\infty}^{\infty} h(k) y(i-k, j) \tag{7}$$

The aim of this filtering scheme is to detect regions of reverberated speech in which direct sound and early reflections dominate, and to mask the areas that contain strong late reverberation. This approach is motivated by observations on human perception of reverberated speech, which emphasize the important role of early reflections on speech intelligibility, and the deleterious effects of late reverberation (Drullman *et al.*, 1994; Houtgast and Steeneken, 1985). The role of the lowpass component h_{lp} is to detect and smooth modulations in the speech range. Following this, the differentiator h_{diff} emphasizes abrupt onsets, which are likely to correspond to direct sound and early reflections.

– Figure 4 about here –

Subsequently, a threshold is applied to the modulation-filtered rate map in order to produce a binary mask for the missing data speech recogniser:

$$m_r(i, j) = \begin{cases} 1 \text{ if } y_r(i, j) > \theta(j) \\ 0 \text{ otherwise} \end{cases} \tag{8}$$

Additionally, the masks are shifted backwards in time to compensate for the delay of the modulation filter $h(n)$. Note that in contrast to the scheme described in Section 3, here we use a binary mask rather than a real-valued mask: initial testing showed that there was no performance gain when using the latter.

The value of the threshold $\theta(j)$ should depend on the degree to which the speech is reverberated. In our previous work $\theta(j)$ was hand-tuned to each reverberation condition (Palomäki *et al.*, 2002), but more recently we have developed a technique for estimating its value directly from an utterance.

Specifically, the threshold is set according to a simple 'blurredness' metric, which exploits the fact that reverberation tends to smooth the rate map by filling the gaps between speech activity with energy originating from reflections. The blurredness metric B is given by

$$B = \sum_{j=1}^{J} \left\{ \frac{\frac{1}{I}\sum_{i=1}^{I} y(i,j)}{\max_i[y(i,j)]} \right\} \tag{9}$$

where I is the number of time frames in the utterance and J=32 is the number of frequency channels. In practice, we have found that it is preferable for $\theta(j)$ to depend not only on B, but also on the mean value over time in channel j of the filtered rate map y_r. Accordingly, we compute the average firing rate $e(j)$ for each filtered rate map channel j as

$$e(j) = \frac{1}{I}\sum_{i=0}^{I} \{y_r(i,j) - \min_i[y_r(i,j)]\} \tag{10}$$

Note that the minimum in the channel is subtracted to ensure that negative values in y_r arising from filtering by Equation 7 are shifted to positive values.

Finally, the threshold $\theta(j)$ is set according to a sigmoidal function of the average firing rate $e(j)$ and blurredness B,

$$\theta(j) = e(j) \cdot \frac{\lambda}{1 + \exp(-\gamma(B - \delta))} \tag{11}$$

where γ = 16 is the slope, δ = 0.42 is the centre point and λ=1.3 determines the width of the sigmoid. These parameters were determined by a series of experiments on a validation set consisting of 300 utterances (different from the training and test sets), which were processed with two different RIRs. A sigmoidal shape was chosen for Equation 11 in order to allow saturation of the threshold at high blurredness values (i.e., long reverberation times).

– Figures 5 and 6 about here –

The reverberation mask estimation process is illustrated in Figures 5 and 6. Figure 5 shows the distribution of the blurredness metric computed for 300 utterances, when no reverberation is present and when the T_{60} reverberation time is 0.7 sec. and 1.5 sec. Note that the distribution shifts to the right (i.e., the mean blurredness increases) with increasing reverberation time.

Figure 6 demonstrates the mask estimation process for a single frequency channel with a centre frequency of 103 Hz. The top panel (A) shows the rate map values in this channel, which are smoothed with a lowpass filter h_{lp} (B) and then differentiated by filtering with h_{diff} (C). Also in panel C, the threshold $\theta(j)$ obtained from Equations 9-11 is shown as a solid line. Finally, the bottom panel (D) shows the reliable regions (solid line) and unreliable regions (dotted line) of the rate map selected by Equation 8. Note that these regions tend to be high in energy, and usually correspond to the first part of a sustained acoustic input (i.e., late reflections are suppressed).

4.2. Normalisation

In reverberant conditions, we do not use a noise estimate; rather, we select the *L* largest values from the regions of $y(i, j)$ which are marked as clean according to the reverberation mask. Specifically, we define a normalisation factor $\eta_r(j)$ as follows:

$$\eta_r(j) = \frac{1}{L} \sum_{i \in \Gamma_r(j)} y_c(i, j) \tag{12a}$$

$$y_c(i, j) = m_r(i, j) \cdot y(i, j) \tag{12b}$$

Here, $m_r(i, j)$ is the binary reverberation mask and $\Gamma_r(j)$ is the set containing the indices of the *L* largest values of $y_c(i, j)$ in channel *j*. Generally L is set as described in Section 3.3. In cases where the value of L computed in this way is less than the number of reliable regions, L is set to the number of reliable regions exactly. Moreover, if channel *j* does not contain any speech dominated features, i.e. when $\Gamma_r(j) = \varnothing$, the scaling factor $\eta_r(j)$ is interpolated from adjacent channels (or extrapolated in the case of the lowest and highest frequency channels).

5. EVALUATION

5.1. Corpus and recogniser configuration

The missing data ASR system was evaluated using a subset of the Aurora connected digits recognition task (Pearce and Hirsch, 2000). The sampling rate of all speech data was 8 kHz. Auditory rate maps were obtained for the training section of the Aurora corpus, and were used to train 12 word-level HMMs (see Section 2.2). In the first experiment (see below), the performance of the missing data ASR system was compared against a baseline HMM recogniser which employed a feature vector of 13 mean-normalised mel-cepstral coefficients (MFCCs) with first and second order temporal derivatives. The MFCC baseline recogniser was configured in a similar manner to the missing data system, except that only 3 mixture components were used to model each state (as opposed to 7 components for the system trained on rate maps) since it was noted that more mixtures caused the MFCC-based models to overfit to clean speech. All models were trained with clean (noiseless and unreverberated) signals. In the experiments involving spectral distortion and reverberation, test utterances were convolved with a microphone or room impulse response, respectively. All of the utterances were presumed to start from silence.

5.2. Baseline hybrid HMM-MLP recogniser

In the following experiments we compare our system against a hybrid HMM-MLP (hidden Markov model multi-layer perceptron) recogniser described by Kingsbury (1998). Kingsbury's system uses two streams of acoustic features which provide robust encoding of speech in the presence of reverberation; cepstral features (plus their deltas and double deltas) obtained by perceptual linear prediction (PLP), together with modulation filtered spectrogram (MSG) features. Here, we have adapted Kingsbury's system for comparison with our missing data recogniser, maintaining the original parameters of his system wherever possible. The system was implemented using the STRUT (1997) speech recognition toolkit. On the test corpus, we present results for three configurations of the hybrid recogniser, firstly using PLP features alone, secondly using MSG features alone, and finally by combining likelihood estimates from PLP and MSG features.

Following Kingsbury's approach, four different MLPs were trained for likelihood estimation. The first two of these were used for tests with PLP and MSG features alone, and the second two were used for the combined features. The MLP network topologies were $189 \times 488 \times 25$ (input layer $\times$ hidden layer $\times$ output layer) for PLP features alone and $189 \times 328 \times 25$ for MSG features alone. For the recogniser using both features, the number of units in the hidden layer of each network was halved, as described by Kingsbury (1998).

Acoustic models for 23 phonemes, silence and unknown (required by the STRUT tools) were obtained from the training part of the Aurora corpus (see also Hermansky *et al*., 2000). Durational information was included in the HMM model for each phone by matching the number of states in the model to half the average duration of the phone, computed from the training set (see page 45 of Kingsbury (1998) for details).

5.3. Experiment 1: Spectral distortion with additive noise

In the first experiment, the performance of the spectral normalisation method was evaluated using the Aurora 2 task. The Aurora corpus contains three different test sets, labelled A, B and C. Test sets A and B are comprised of different utterances and also differ from each other due to the type of additive noise; subway noise in the former and car noise in the latter. Also, test sets A and B have transmission line characteristics defined by G.712 (ITU-T, 1996a), which is the same characteristic applied to the training part of the corpus. Therefore, test sets A and B are not regarded as spectrally distorted.

For testing the effect of transmission line distortion, test set C is provided. Test set C is a subset of the speech and noise mixtures from sets A and B, but in addition the signals are filtered with the MIRS telephone front-end (ITU-T, 1996b). MIRS differs in its spectral characteristic to G.712; the latter has a flat response in the telephone band of 300 Hz – 3.4 kHz, whereas MIRS has a rising gain at higher frequencies and some attenuation at low frequencies. MIRS defines an official recommendation for the frequency characteristic of a telecommunication channel sender and receiver, including the microphone and speaker respectively.

In order to evaluate the effect of spectral distortion we used test signals which shared common noise types in the spectrally matching test sets (A and B) and spectrally mismatching case (C). We also created two additional spectrally distorted test conditions by convolving samples (speech with subway noise, test set A; and speech with street noise test set B) with impulse responses of poor quality microphones. The impulse responses of these microphones are depicted in Figure 7.

– Figure 7, Table 1 and Table 2 about here –

To put the performance of our system in perspective, it is tested against a missing data system that does not use any spectral normalisation (Barker *et al.*, 2000b). We also compare the performance of missing data systems against the MFCC baseline system, generated as recommended within the Aurora framework (Pearce and Hirsch, 2000). The results of the experiment are shown in Tables 1 and 2. In the spectrally non-distorted test case (test set A subway noise, and test set B street noise), the performance of the two missing data systems was comparable, with both performing better than the MFCC baseline at low SNR (Table 1). In these particular (non-distorted) test cases, the performance of the proposed missing data system with spectral normalisation was slightly lower than that of the system without normalisation. However, when tested with spectrally distorted input (Table 2) the advantages of the proposed normalisation technique become evident. The differences in performance are most noticeable in the worst spectral distortion condition (microphone 2) and at low SNRs.

5.4. Experiment 2: Random gain modulations with additive noise

In the Aurora test corpus the energy of each speech sample was equalised before artificially adding noise (Pearce and Hirsch, 2000). Clearly, such equalisation is not representative of natural acoustic environments, in which speech intensity depends upon the signal path (e.g., the distance between the speaker and the microphone) and on the loudness of speech production itself. In previous missing data work (e.g., Barker *et al.*, 2000a,b; 2001a,b; Cooke *et al.*, 2001) this issue has not been addressed; it is therefore unlikely that the results obtained on energy-equalised corpora in these studies will generalise to real acoustic environments. Here, we demonstrate that our proposed spectral normalisation scheme also improves robustness when the acoustic input is subject to variations in overall sound level.

For testing purposes we generated a random gain for each utterance in the test set. This gain was held constant for the duration of the utterance. It should be noted that the same seed was used to randomise gains in each experimental condition; hence the corresponding speech samples were scaled with the same random value in each condition, in order to allow a direct comparison.

– Table 3 about here –

Gain modulation tests are shown in Table 3 for missing data systems with and without spectral normalisation, and for the MFCC baseline system. The gain on the input was varied randomly between -10 dB and 10 dB. Comparison with Table 1 indicates that the performance of the missing data recogniser without spectral normalisation is degraded by gain modulation, even in the clean condition. In comparison, the missing data system with spectral normalisation is unaffected by gain modulation.

5.5. Experiment 3: Reverberation

The degree of reverberation in an enclosed space is often characterized using a simple measure called reverberation time T_{60}, which is defined as the time required for the reverberation level to drop 60 dB below that of the original sound onset. For example, the recommended T_{60} for a speech hall is 0.4 sec., whereas a richer acoustic environment (and hence a longer T_{60}) is required for music; a typical value for a concert hall is 2.0 sec.

For testing the model performance under reverberant conditions the speech samples were convolved with impulse responses of rooms with different reverberation characteristics. A total of 6 impulse responses were used in the testing. Four of these responses were originally used by Kingsbury (1998). They were recorded in a varechoic chamber with two different settings of the wall panels. For the first wall panel setting the T_{60} was 0.7 sec. and the distances between the source and microphone were 2.35 m and 3.05 m. For the second wall panel setting, the T_{60} was 1.2 sec. and source-microphone distances were 2.0 m and 3.05 m. Another two impulse responses (not used by Kingsbury) were measured in a larger room, having a T_{60} of 1.5 sec. and source-microphone distances of 6.1 m and 18.3 m.

– Table 4 about here –

The results of this experiment are shown in Table 4. The missing data system with reverberation mask estimation, described in Section 4.1, outperformed the MSG+PLP baseline in the most reverberant test cases. However, the performance of the MSG+PLP system was better than that of the missing data system for the shortest T_{60} condition, and in clean conditions (no reverberation). The hybrid HMM-MLP recogniser using MSG+PLP features always performed better than configurations of this system which used MSG or PLP features alone.

6. DISCUSSION

In this paper we have described techniques for handling convolutional distortion in 'missing data' speech recognition, an issue which has been largely unaddressed to date. As the convolutional interference can be quite different in nature depending upon the length of the impulse response concerned, we propose two approaches; one to handle spectral distortion due to a transmission line or audio equipment, and another to handle room reverberation interference. In summary, the results show substantial performance improvements compared to a standard missing data recogniser when speech is contaminated by additive noise and spectrally distorted or when the intensity of the input speech varies. The performance of the missing data approach is superior to that of a MFCC baseline system at low SNRs. We also developed a missing data mask estimation system for reverberant speech recognition, based on detection of the strongest modulation frequencies of speech. Our system performs rather better than a hybrid HMM-MLP recogniser employing MSG and PLP features (Kingsbury, 1998; Kingsbury *et al*., 1998), for T_{60} reverberation times of 1.2 sec. and greater.

The reverberation masking system proposed here has some parallels with RASTA-PLP (Hermansky and Morgan, 1994) and MSG (Kingsbury, 1998), which are used for producing noise robust feature vectors. Both of these techniques have a processing chain that firstly divides the signal into frequency bands and then (after downsampling and compression) applies a band-pass filter to emphasise the most noise-tolerant speech signal regions. RASTA-PLP and MSG have both been applied to robust ASR in reverberation, with a combination of likelihood estimates from MSG and PLP being most successful (Kingsbury, 1998).

Both MSG and the proposed modulation filtering approach to mask estimation exploit the fact that the strongest modulations of speech occur at modulation frequencies roughly between 0 Hz and 10 Hz. We believe, however, that our approach has some advantages. When noise robust techniques such as MSG are used, the same acoustic features must be used during training and recognition. This, in turn, might lead to a compromise because the use of particular acoustic features may effectively tune the ASR system to certain acoustic conditions. For example, when using modulation-filtered features, the configuration of the front-end needed for optimal performance in reverberant conditions may be less than optimal for recognition of clean speech.

In principle, the missing data approach can overcome this problem because unreliable regions are filtered out by the mask estimation processing during recognition; acoustic models are trained on clean speech, and hence there is no need to re-train for different conditions. In practice, however, the baseline system outperformed our missing data system in the least reverberated cases. This may be because our method of estimating the amount of reverberation present in a speech sample is not sensitive enough to distinguish between anechoic and mildly reverberant conditions; future work will address this issue. In the most reverberant cases, however, the missing data processing has a clear advantage compared to the MSG+PLP system.

A benefit of the missing data approach is that it does not make assumptions about the type of noise present. Therefore, a missing data recogniser can be adapted to different noise conditions simply by changing the mask estimation rule; any assumptions about the noise type are restricted to the mask estimation process. Hence, different types of front-end can be ‘switched in’. For example, here we have described two front-ends for the same recogniser; one that is robust for additive noise and another that is robust for reverberation. This approach may offer advantages for speech recognition in mobile devices, since the mask estimation process could be dynamically altered to compensate for different acoustic conditions as they arise. Future work will investigate this possibility.

ACKNOWLEDGMENTS

KJP was funded by the EC TMR SPHEAR project, the Academy of Finland (project number 1277811) and was partially supported by a Finnish Nokia säätiö grant. GJB was funded by EPSRC grant GR/R47400/01. The authors owe many thanks to Dan Ellis, Brian Kingsbury and Heidi Christensen for their kind help with implementing the MSG+PLP baseline system. Dan Ellis and Brian Kingsbury also made some of the real room impulse responses available to us. The authors also wish to thank Jim West, Michael Gatlin and Carlos Avendano who originally collected these responses.

REFERENCES

Assmann, P., Summerfield, Q., 2003. The perception of speech under adverse acoustic conditions, in: Greenberg, S., Ainsworth, W. (Eds.) Speech processing in the auditory system (Springer handbook of auditory research vol. 18), Springer-Verlag.

Atal, B. S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am., 55, 1304-1312.

Barker, J., Cooke, M. P., Ellis, D. P. W., 2000a. Decoding speech in the presence of other sound sources. Proc. ICSLP-2000, IV, 270-273.

Barker, J., Josifovski, L., Cooke, M. P., Green, P. D., 2000b. Soft decisions in missing data techniques for robust automatic speech recognition. Proc. ICSLP-2000, I, 373-376.

Barker, J., Cooke, M. P., Green, P. D., 2001a. Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. EUROSPEECH-2001, 213-217.

Barker, J., Green, P. D., Cooke, M. P., 2001b. Linking auditory scene analysis and robust ASR by missing data techniques, Proceedings of the Workshop on Innovations in Speech Processing (WISP-2001), Stratford-upon-Avon, UK, 2nd-3rd April.

Bradley, J. S., 1986. Predictors of speech intelligibility in rooms. J. Acoust. Soc. Am., 80, 837-845.

Bregman, A. S., 1990. Auditory scene analysis. MIT Press, Cambridge, MA.

Brown, G. J., Cooke, M. P., 1994. Computational auditory scene analysis. Comp. Speech Lang., 8, 297-336.

Brown, G. J., Barker, J., Wang, D. L., 2001. A neural oscillator sound separator for missing data speech recognition. Proc. IJCNN-2001, 2907-2912.

Cooke, M. P., 1993. Modelling auditory processing and organization. Cambridge University Press, Cambridge, UK.

Cooke, M. P., Green, P. D., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Comm., 34, 267-285.

Davis, S. P., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process., ASSP-28, 357-366.

Droppo, J., Acero, A., Deng, L., 2002. Uncertainty decoding with SPLICE for noise robust speech recognition. Proc. ICASSP-2002, I, 57-60.

Drullman, R., Festen, J. M., Plomp, R., 1994. Effects of temporal envelope smearing on speech reception. J. Acoust. Soc. Am., 95, 1053-1064.

Gold, B., Morgan, N., 2000. Speech and audio signal processing, John Wiley and Sons, Inc., NY.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am., 87, 1738-1752.

Hermansky, H., 1998. Should recognisers have ears? Speech Comm., 25, 3-27.

Hermansky, H., Ellis, D. P. W., Sharma, S., 2000. Tandem connectionist feature extraction for conventional HMM systems. Proc. ICASSP-2000, III, 1635-1638.

Hermansky, H., Morgan, N., 1994. RASTA Processing of Speech. IEEE Trans. Speech and Audio Proc., 2, 578-589.

Houtgast, T., Steeneken, H. J. M., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Am., 77, 1069-1077.

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent component analysis, John Wiley and Sons, NY.

ITU-T recommendation G.712, 1996a. Transmission performance characteristics of pulse code modulated channels. International Telecommunications Union, Geneva.

ITU-T recommendation P.830, 1996b. Subjective performance assessment of telephone band and wide band digital codecs. International Telecommunications Union, Geneva.

Kingsbury, B. E. D., 1998. Perceptually inspired signal-processing strategies for robust speech recognition in reverberant environments. PhD thesis, Univ. California, Berkeley.

Kingsbury, B. E. D., Morgan, N., Greenberg, S., 1998. Robust speech recognition using the modulation spectrogram. Speech Comm., 25, 117-132.

Lippmann, R. P., 1997. Speech recognition by machines and humans. Speech Comm., 22, 1-15.

Mathworks, Inc., 2003. MATLAB release 13 reference manual. Natick, MA.

Moore, B. C. J. (2003) An introduction to the psychology of hearing, fifth ed. Academic press, Cambridge, UK.

Morris, A. C., 2002. Analysis of noise PDF transformation in secondary feature processing. IDIAP Research Report 02-29, IDIAP, Martigny, Switzerland.

Nabelek, A. K., Robinson, P. K., 1982. Monaural and binaural speech perception in reverberation for listeners of various ages. J. Acoust. Soc. Am., 71, 1242-1248.

Omologo, M., Svaizer, P., Matassoni, M., 1998. Environmental conditions and acoustic transduction in hands-free speech recognition. Speech Comm., 25, 75-95.

Palomäki, K. J., Brown, G. J., Barker, J., 2002. Missing data speech recognition in reverberant conditions. Proc. ICASSP-2002, I, 65-68.

Palomäki, K. J., Brown, G. J., Wang, D. L., 2001. A binaural auditory model for missing data speech recognition in noisy and reverberant conditions. Proc. CRAC EUROSPEECH-2001 satellite workshop, Aalborg, 2nd September.

Palomäki, K. J., Brown, G. J., Wang, D. L., in press. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. Accepted for publication in Speech Comm.

Pearce, D., Hirsch, H. -G., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proc. ICSLP-2000, 4, 29-32.

STRUT Version 2.4, 1997. Step by step guide to using the speech training and recognition unified tool STRUT, http://www.tcts.fpms.ac.be/asr/project/strut/.

FIGURE LEGENDS

Figure 1. Schematic diagram of the model. In the processing pathway described in Section 3, a mask m_s is derived from local SNR estimates, and this is passed to the recogniser together with a 'cleaned' rate map y_s, which is normalised by a factor η_s. In the pathway described in Section 4, a reverberation mask m_r is estimated and this is passed to the recogniser together with the rate map *y*, normalised by a factor η_r.

Figure 2. A. Rate map for the male utterance "zero one zero five nine" without added noise. B. Rate map for the same utterance in the presence of noise with an SNR of 5 dB. C. Soft SNR mask (black pixels are reliable, white pixels are unreliable).

Figure 3. Selection of time-frequency regions for spectral normalisation, for the male utterance "seven two one nine" when (A) clean and (B) mixed with subway noise at a SNR of 5 dB. Black areas correspond to the regions selected for scaling according to the L-largest rule.

Figure 4. Frequency response of the modulation filter, $h(n)$.

Figure 5. Distributions of blurredness *B* for three reverberation conditions, computed from a test set of 300 utterances.

Figure 6. Demonstration of modulation filtering-based mask estimation. A. Output of the rate map channel with a centre frequency of 103 Hz. B. Rate map channel filtered by the lowpass part $h_{lp}(n)$ of the modulation filter. C. Rate map channel filtered by the whole modulation filter $h(n)$. The horizontal line indicates the value of the threshold θ. D. Estimated reliable regions (solid line) and unreliable regions (dotted line).

Figure 7. Frequency responses of the two microphone responses used in the second experiment.

TABLE LEGENDS

Table 1. Speech recognition performance for non-distorted test cases. Each row shows the results for three different recognisers: missing data recogniser with spectral normalisation (MD-SN), missing data recogniser without spectral normalisation (MD) and Aurora MFCC baseline (MFCC). The test cases are subway noise and street noise, added at SNRs between -5 dB and 20 dB. Results for clean speech are also shown.

Table 2. Speech recognition performance for spectrally distorted test cases. Each row shows the results for three different recognisers: missing data recogniser with spectral normalisation (MD-SN), missing data recogniser without spectral normalisation (MD) and Aurora MFCC baseline (MFCC). The test cases are (from top to bottom) MIRS characteristic, first microphone characteristic and second microphone characteristic. In each condition, the filtering characteristic was applied after mixing with subway noise or street noise, at SNRs between -5 dB and 20 dB. The 'clean' column indicates performance when the respective filtering characteristic was applied to speech without added noise.

Table 3. Speech recognition performance in the gain modulation test. Each row of the table shows the results for three different recognisers: missing data recogniser with spectral normalisation (MD-SN), missing data recogniser without spectral normalisation (MD) and Aurora MFCC baseline (MFCC). Test conditions are gain modulations with peak amplitude change of ±10 dB after mixing with subway noise or with street noise. For each noise condition, results are shown for SNRs between -5 dB and 20 dB, and for clean speech (i.e., gain modulation but no added noise).

Table 4. Speech recognition performance in the reverberation task. Results are shown for four systems in six reverberation conditions, and for unreverberated speech. Columns indicate the performance of the hybrid HMM-MLP recogniser using PLP features alone (HYBRID PLP), modulation spectrogram features alone (HYBRID MSG) and both features together (HYBRID MSG+PLP), and for the missing data system (MD).

Table 1

Noise type	Method	-5	0	5	10	15	20	Clean
Subway	MD-SN	28.7	54.0	74.8	86.8	92.1	94.9	97.5
	MD	30.0	54.2	75.3	85.9	92.6	95.7	98.8
	MFCC	12.6	27.3	53.4	78.7	92.9	97.0	98.8
Street	MD-SN	24.5	51.6	73.2	85.1	91.6	94.3	97.2
	MD	28.7	52.9	73.2	84.9	91.8	94.9	98.6
	MFCC	10.1	18.7	38.2	66.8	88.3	95.8	99.0

Table 2

Noise type		-5	0	5	10	15	20	Clean
MIRS Subway	MD-SN	28.4	55.0	75.8	85.8	91.7	94.5	97.3
	MD	20.7	44.3	67.3	81.5	89.6	92.9	97.6
	MFCC	12.1	26.0	52.8	75.2	87.6	94.5	99.0
MIRS Street	MD-SN	25.8	51.7	73.2	83.6	91.4	94.3	96.9
	MD	19.4	40.8	63.9	78.9	87.2	91.7	96.8
	MFCC	10.7	21.6	48.9	75.2	89.7	95.1	99.0
Microphone 1 Subway	MD-SN	26.8	52.4	72.8	85.1	91.2	94.4	97.3
	MD	22.2	45.4	69.1	83.6	91.1	94.2	98.3
	MFCC	8.9	17.6	48.3	76.6	90.9	96.0	98.7
Microphone 1 Street	MD-SN	23.9	50.9	71.7	84.5	90.6	93.6	96.8
	MD	22.2	44.3	67.9	81.2	89.7	94.2	97.7
	MFCC	9.4	15.1	35.4	66.2	87.9	95.8	98.8
Microphone 2 Subway	MD-SN	23.9	48.4	69.3	82.1	90.1	93.6	97.1
	MD	14.4	28.0	46.2	60.2	72.5	80.3	88.7
	MFCC	7.7	7.3	8.0	14.7	28.5	50.5	93.8
Microphone 2 Street	MD-SN	21.8	47.5	70.1	83.6	90.0	93.0	95.9
	MD	17.6	32.9	49.2	62.1	72.9	80.0	87.8
	MFCC	9.0	12.8	23.1	37.2	55.6	71.8	94.1

Table 3

Noise type	Method	-5	0	5	10	15	20	Clean
Subway	MD-SN	28.6	54.2	75.0	86.5	91.9	94.8	97.5
	MD	24.0	47.4	67.9	79.6	87.0	91.1	96.2
	MFCC	12.3	27.2	52.7	75.0	90.3	95.9	98.8
Street	MD-SN	25.2	51.4	73.1	84.9	91.8	94.2	97.1
	MD	24.8	46.0	66.0	78.6	86.9	90.4	96.1
	MFCC	10.6	19.0	38.5	64.1	84.1	93.6	99.0

Table 4

T_{60} and source-receiver distance	HYBRID PLP	HYBRID MSG	HYBRID MSG+PLP	MD
1.5 sec., 18.3 m	53.3	53.5	59.8	63.2
1.5 sec., 6.1 m	55.2	62.0	64.0	67.6
1.2 sec., 3.05 m	59.1	66.6	69.5	75.6
1.2 sec., 2.0 m	60.2	71.3	71.5	77.6
0.7 sec., 3.05 m	88.0	93.0	93.5	91.9
0.7 sec., 2.35 m	89.5	94.0	95.1	93.0
Unreverberated	98.2	98.0	98.5	97.0

Figure 1

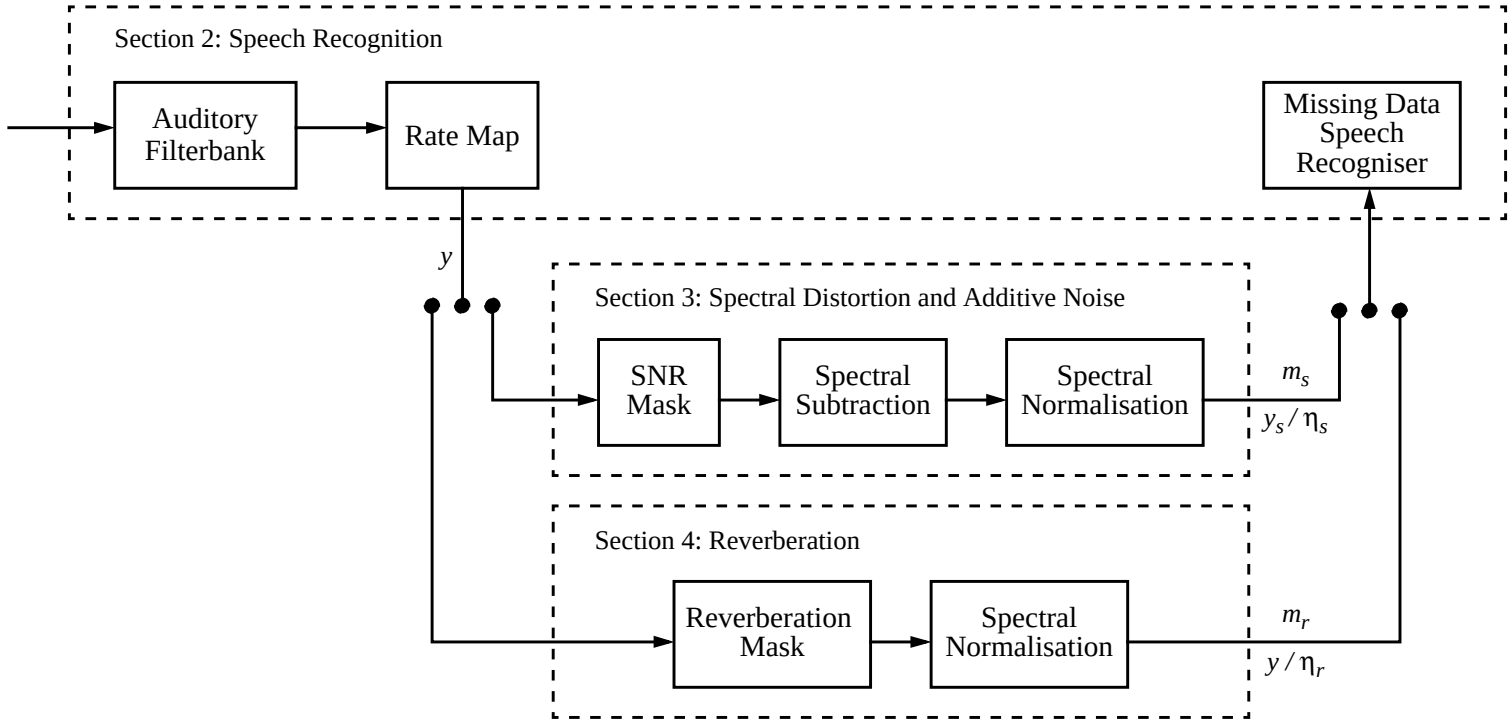

Section 2: Speech Recognition
Auditory Filterbank
Rate Map
Missing Data Speech Recogniser
y
Section 3: Spectral Distortion and Additive Noise
SNR Mask
Spectral Subtraction
Spectral Normalisation
m_s
y_s / η_s
Section 4: Reverberation
Reverberation Mask
Spectral Normalisation
m_r
y / η_r

Figure 2

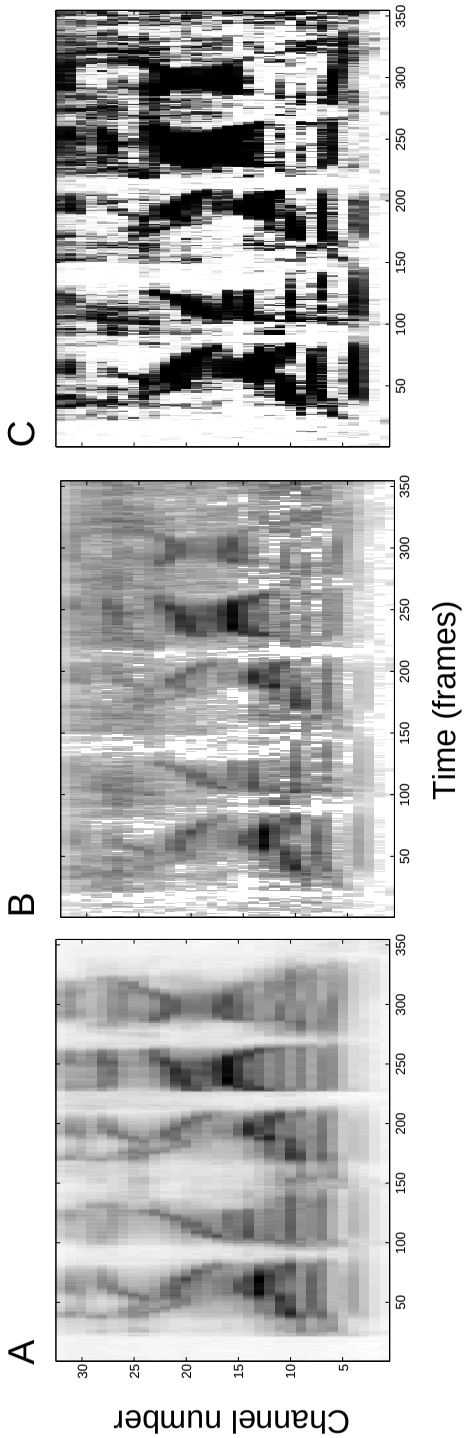

A
B
C
Channel number
30
25
20
15
10
5
50
100
150
200
250
300
350
Time (frames)

Figure 3

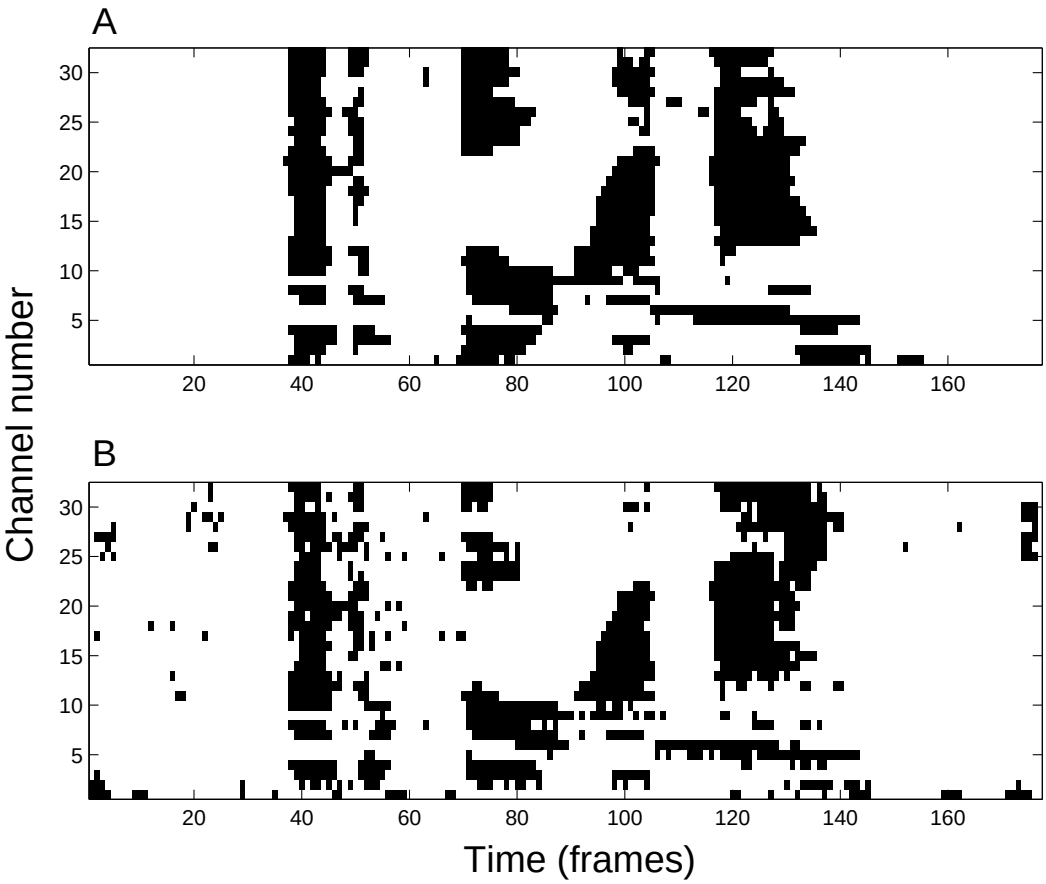


Figure 4

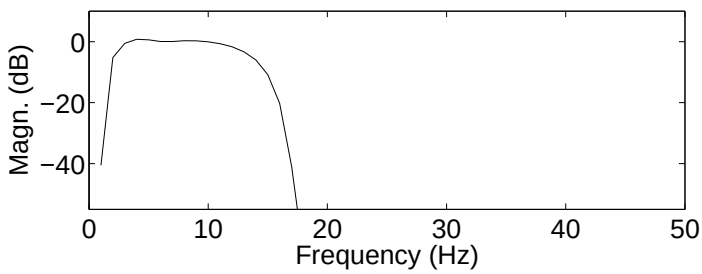


Figure 5

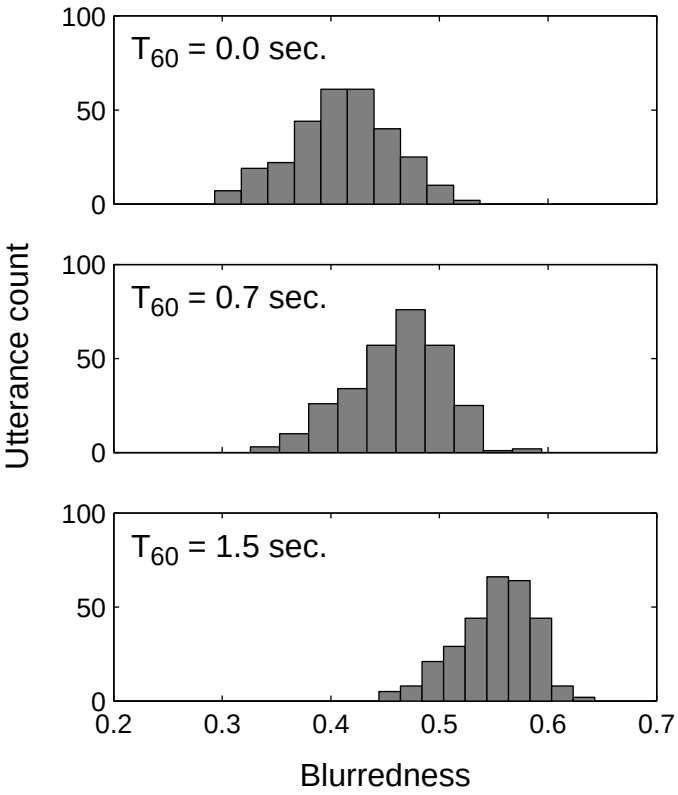

100
50
0
T_{60} = 0.0 sec.
100
50
0
T_{60} = 0.7 sec.
100
50
0
T_{60} = 1.5 sec.
Utterance count
0.2
0.3
0.4
0.5
0.6
0.7
Blurredness

Figure 6

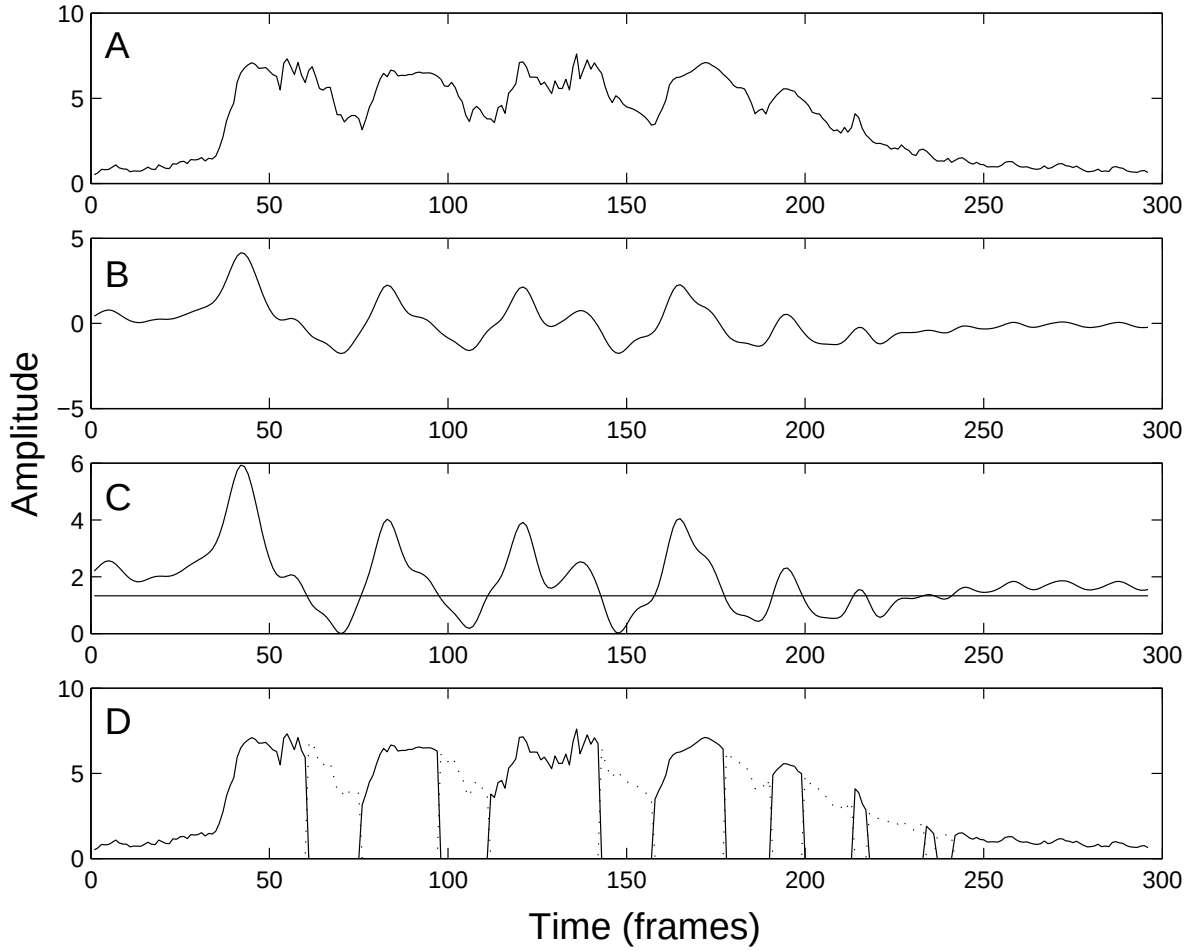
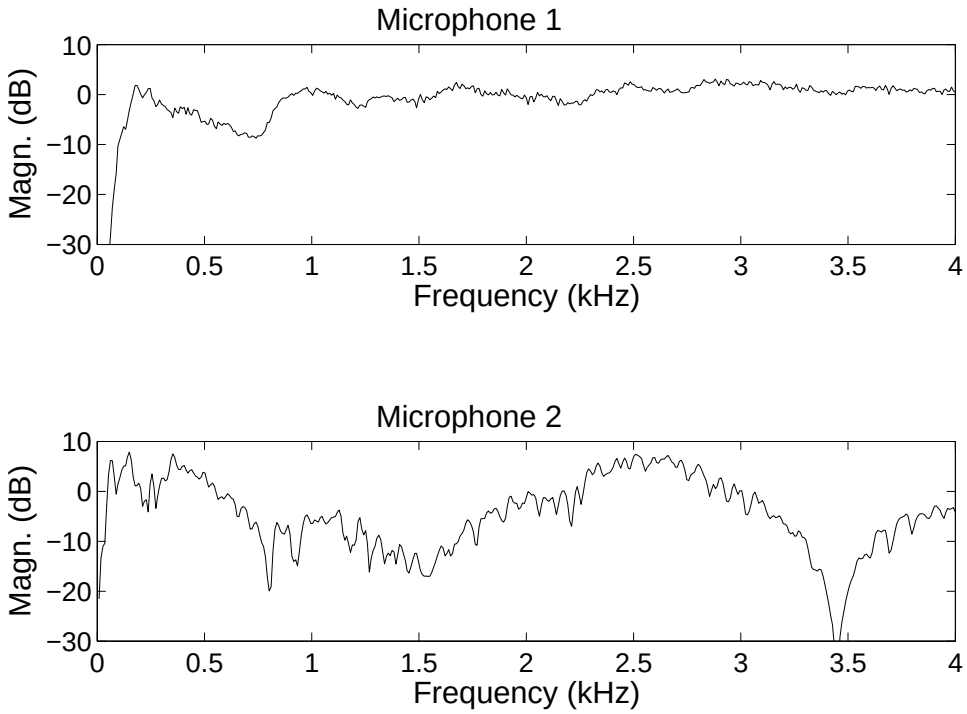


Figure 7


Microphone 1
Magn. (dB)
10
0
−10
−20
−30
0
0.5
1
1.5
2
2.5
3
3.5
4
Frequency (kHz)
Microphone 2
Magn. (dB)
10
0
−10
−20
−30
0
0.5
1
1.5
2
2.5
3
3.5
4
Frequency (kHz)