Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 43 (2004) 123–142
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Techniques for handling convolutional distortion with 'missing data' automatic speech recognition

Kalle J. Palomäki [a,b,c,*], Guy J. Brown [b], Jon P. Barker [b]

[a] Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, P.O. Box 3000, FIN-02015 HUT, Finland
[b] Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK
[c] Apperception & Cortical Dynamics (ACD), Department of Psychology, University of Helsinki, P.O.B. 9, FIN-00014, Finland

Received 11 August 2003; received in revised form 21 January 2004; accepted 27 February 2004

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 gain of the input signal 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 the reverberation time T_{60} exceeds 0.7 s, 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

* Corresponding author. Address: Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, P.O. Box 3000, FIN-02015 HUT, Finland. Tel.: +358-9-451-2883; fax: +358-9-460-224.
E-mail addresses: kalle.palomaki@hut.fi (K.J. Palomäki), g.brown@dcs.shef.ac.uk (G.J. Brown), j.barker@dcs.shef.ac.uk (J.P. Barker).

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.02.005

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 length of the impulse response is short compared to the analysis window, then the interference mainly causes spectral alteration (see Avendano, 1997, Chapter 5). This follows because convolution in the time domain is equivalent to multiplication in the frequency domain (see Oppenheim and Schafer (1989) for a description of the convolution theorem of the Fourier transform). The analysis window used in speech processing is usually longer than 10 ms, which roughly corresponds to the pitch period of an average adult male voice. Examples of practical systems having short impulse responses are transmission lines, microphones and loudspeakers.

In the case of room reverberation the interaction is of a different nature, because the impulse response of a room is relatively long (from approximately 0.2 up to 5 s) compared to the window used for speech analysis. 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 often contribute usefully to speech intelligibility by increasing the loudness of the speech. However, early reflections can also cause some spectral deviation due to comb filtering caused by successive reflections and the varying frequency characteristics of surface absorption. In contrast, the dense late reverberation is poorly correlated with the original speech signal and therefore behaves more like additive noise. Indeed, early versus late reverberation has successfully been used as a predictor of speech intelligibility in rooms (Bradley, 1986). It is common to define a critical delay time for early and late reverberation, such that reflections arriving before the delay are beneficial to auditory perception whereas reflections arriving after it will have a detrimental effect. The European norm ISO 3382 (1997) suggests critical delays of 50 ms for speech and 80 ms for music perception. Gölzer and Kleinschmidt (2003)

investigated the role of early and late reflections in conjunction with ASR. They suggested that reflections have a conducive effect on speech recognition accuracy up to a critical delay of 25–50 ms, assuming that late reverberation is strongly present in the room impulse response. Further details of the effect of room acoustics on speech intelligibility can be found in (Bradley, 1986; 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 (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) and in ASR (Kandera et al., 1999).

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 and 1.5 s. The results obtained with our new method are also compared against Kingsbury (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 Fig. 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.

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; Rosenberg et al., 1994).

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., 2002). 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).

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 uniformly on the equivalent rectangular bandwidth (ERB) scale between 50 Hz (1.837 ERB) and 3850 Hz (26.772 ERB). Spacing between adjacent filter channels is therefore 0.804 ERB, which is close to the 3-dB bandwidth of the gammatone filter (0.887

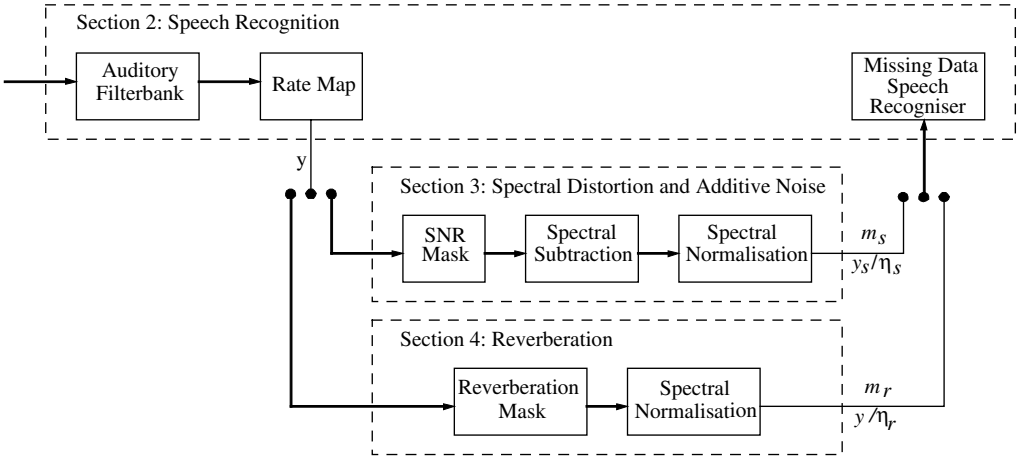

Fig. 1. Schematic diagram of the system. In the processing pathway described in Section 3, a mask m_s is derived from local SNR estimates, and this is passed to the recogniser together with a 'cleaned' rate map y_s, which is normalised by a factor η_s. In the pathway described in Section 4, a reverberation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

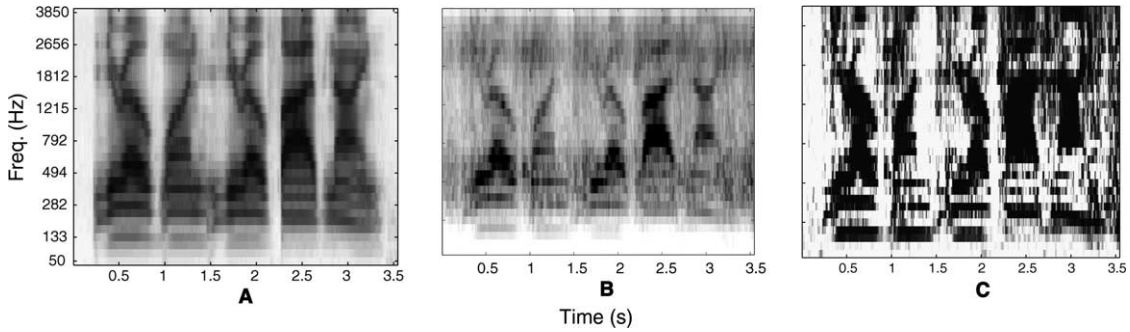

Fig. 2. (A) Rate map for the male utterance "zero one zero five nine" without added noise. (B) Rate map for the same utterance in the presence of noise with an SNR of 5 dB. (C) Soft SNR mask (black pixels are reliable, white pixels are unreliable).

ERB; see Patterson et al., 1988). For details of the digital implementation of the gammatone filter 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 lowpass filter with an 8-ms time constant, sampled at 10 ms intervals, and finally compressed by raising it to the power of 0.3 to give a crude simulation of auditory nerve firing rate (a 'rate map'; see Fig. 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 and assuming the prior $f(Y)$ to be constant, 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; for an alternative approach see Raj et al., in press). Here, we use bounded marginalisation where Y is a vector of simulated auditory nerve firing rates; thus the lower bound $Y_{u,low}$ is zero (since a firing rate cannot be negative) and the upper bound $Y_{u,high}$ is the observed firing rate. Assuming diagonal Gaussian mixture components (indexed by k), this leads to the following integral over the unreliable parts Y_u:

$$\int_{Y_{u,low}}^{Y_{u,high}} f(Y_u|k,C)\,dY_u \tag{1}$$

We cannot calculate the likelihood $f(Y|C)$ as some of the components of Y, though bounded, are not precisely known. Instead, we calculate, $\overline{f(Y|C)}$, the average of the likelihood $f(Y_r, Y_u|C)$ over the range of allowable Y_u values. This can be expressed as,

$$\overline{f(Y|C)} = \sum_{k=1}^{M} P(k|C) f(Y_r|k,C) \frac{1}{Y_{u,high} - Y_{u,low}} \times \int_{Y_{u,low}}^{Y_{u,high}} f(Y_u|k,C)\,dY_u \tag{2}$$

where $P(k|C)$ are the Gaussian mixture coefficients. For justification of Eq. (2) see Barker et al. (2000b). 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, a binary judgement is made as to whether

data is reliable (1) or unreliable (0). 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. Note that the missing data algorithm is only applied during the testing phase, not during training.

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,

$$z(j) = \frac{1}{K}\sum_{i=1}^{K} y_{\mathrm{e}}(i,j) \tag{3}$$

where $y_{\mathrm{e}}(i,j) = y(i,j)^{3.333}$ and $z(j)$ is the noise estimate for frequency channel j. Note that $z(j)$ is estimated from a version of the rate map, $y_{\mathrm{e}}(i,j)$, to which compression has not been applied. The noise estimate is used to calculate a local SNR $s(i,j)$,

$$s(i,j) = 20\log_{10}\left(\frac{y_{\mathrm{e}}(i,j) - z(j)}{z(j)}\right) \tag{4}$$

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 0–1 (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.,

$$m_{\mathrm{s}}(i,j) = \sigma[s(i,j)] = \frac{1}{1+\exp\{-\alpha[s(i,j)-\beta]\}} \tag{5}$$

where $m_{\mathrm{s}}(i,j)$ 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 to 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 (5) 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:

$$y_s(i,j) = \{[y_e(i,j) - z(j)]^+\}^{0.3} \quad (6)$$

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 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, 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 Fig. 3.

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:

$$\eta_s(j) = \frac{1}{L} \sum_{i \in \Gamma_s(j)} y_s(i,j) \quad (7)$$

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 Eq. (7) is that selection of the L largest values in each channel of the rate map gives a comparable result with clean

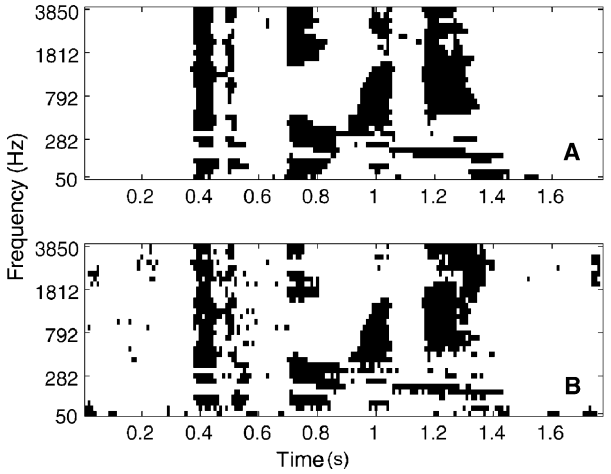

Fig. 3. Selection of time–frequency regions for spectral normalisation, for the male utterance "seven two one nine" when (A) clean and (B) mixed with subway noise at a SNR of 5 dB. Black areas correspond to the regions selected for scaling according to the L-largest rule.

Table 1
Speech recognition accuracy (percent) for different values of the parameter D, for SNRs between 0 and 20 dB, and for clean speech

D	0	10	20	Clean
2	51.5	85.3	94.2	98.0
3	53.8	86.4	95.0	97.6
5	54.0	86.8	94.9	97.5
7	53.8	86.1	94.4	97.4
10	52.7	85.8	94.2	97.3

(training) and noisy (recognition) data, so long as the noise is fairly stationary and the global SNR is favourable.

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.

Note that normalisation proceeds on an utterance-by-utterance basis, and could therefore be affected by utterance length. Here we have attempted to account for this by normalising by a factor $L = I/D$, where I is the number of time frames in the input and D is a constant parameter. The training section of the corpus used here contained utterances with a minimum, maximum and mean duration of 0.59, 5.15 and 1.76 s respectively. Values of D between 2 and 10 were considered, with $D = 5$ found to give the best performance (Table 1).

Fig. 3 illustrates the feature selection process for $D = 5$. It shows that very similar regions of strong speech can be spotted in clean and noisy rate maps using the proposed technique.

4. Processing for reverberation

This section describes a processing pathway for missing data ASR in reverberant conditions (see Fig. 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):

$$h(n) = h_{\mathrm{lp}}(n) \otimes h_{\mathrm{diff}}(n) \tag{8}$$

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). Parameters of the differentiator part are the following, $h_{\mathrm{diff}}(0) = 1$, $h_{\mathrm{diff}}(1) = -0.999$ and $h_{\mathrm{diff}}(n) = 0$ when $n \neq \{0, 1\}$. The lowpass part h_{lp} was designed using the MATLAB `fir2` function [1] (Mathworks, 2003). Amplification in the lowpass part was greater than zero in order to set the gain in the pass-band of the combined filter to approximately 0 dB. The resulting modulation filter $h(n)$ is bandpass (see Fig. 4), with 3 dB cutoff frequencies at 1.5 and 8.2 Hz. The limiting zeros are placed at 0 and 11.7 Hz. The filter $h(n)$ is used to derive a modulation-filtered rate map $y_{\mathrm{r}}(i, j)$ by filtering each channel j of $y_{\mathrm{r}}(i, j)$ as follows:

$$y_{\mathrm{r}}(i, j) = \sum_{k=-\infty}^{\infty} h(k) y(i - k, j) \tag{9}$$

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

[1] The specific function call was `fir2(100,[0 0.04 0.07 0.1 0.11 0.21 0.5 1],[9.3 7.44 4.65 2.883 3.162 0 0 0])`, using MATLAB 6.5 release 13.

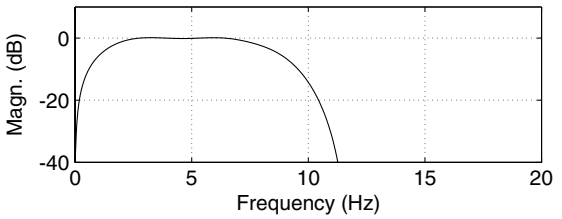

Fig. 4. Frequency response of the modulation filter, $h(n)$.

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.

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

$$m_{\mathrm{r}}(i,j) = \begin{cases} 1 & \text{if } y_{\mathrm{r}}(i,j) > \theta(j) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Additionally, the masks are shifted backwards in time to compensate for the group delay of the modulation filter $h(n)$. The filter has a linear phase impulse response which is 102 frames in length, and has a group delay (constant across frequencies) of 50 frames. Rather than correct for the group delay exactly, a backward shift of 48 frames is applied. Maintaining a two-frame offset was found to be beneficial, probably because it discards reverberation contaminated frames that occur just before speech onsets. 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

$$B = \sum_{j=1}^{J} \left\{ \frac{\frac{1}{I}\sum_{i=1}^{I} y(i,j)}{\max_i[y(i,j)]} \right\} \tag{11}$$

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

$$e(j) = \frac{1}{I}\sum_{i=1}^{I}\{y_{\mathrm{r}}(i,j) - \min_i[y_{\mathrm{r}}(i,j)]\} \tag{12}$$

Note that the minimum in the channel is subtracted to ensure that negative values in y_{r} arising from filtering by Eq. (9) 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,

$$\theta(j) = e(j) \cdot \frac{\lambda}{1 + \exp(-\gamma(B - \delta))} \tag{13}$$

where $\gamma = 16$ is the slope, $\delta = 0.42$ is the centre point and $\lambda = 1.24$ 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 Eq. (13) in order to allow saturation of the threshold at high blurredness values (i.e., long reverberation times).

The bandwidth of the modulation filter $h(n)$ was also derived through experimentation on a validation set of 300 utterances. Each utterance in the validation set was convolved with three RIRs, one for each room used in the evaluation. For these, reverberation times T_{60} and microphone to source distances were 0.7 s and 3.05 m, 1.2 s and 3.05 m, and 1.5 s and 18.3 m. Subsequent recognition testing on the reverberated validation set

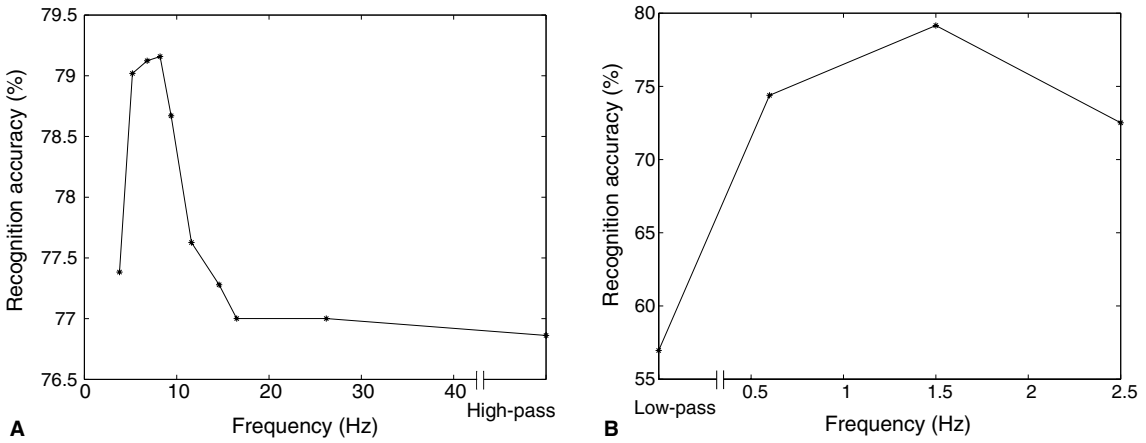

Fig. 5. Tuning the modulation bandpass filter. The graphs show recognition accuracies averaged over 300 test utterances, each convolved with three room impulse responses. (A) The 3 dB highpass cutoff frequency was fixed at 1.5 Hz and the lowpass cutoff was varied. The last data point in the graph represents a plain highpass filter. (B) The lowpass cutoff frequency was fixed at 8.2 Hz and the highpass cutoff was varied. The first data point represents a plain lowpass filter.

yielded a recognition accuracy for each reverberation condition. The average recognition accuracy over these three conditions was investigated for a number of different filter parameters. For simplicity, the threshold was chosen as $\theta(j) = e(j)\theta'$, where θ' was hand-tuned to be optimal for each reverberation condition. Fig. 5 shows the recognition accuracy obtained for various configurations of the modulation filter. The left panel (A) demonstrates the effect of varying the lowpass

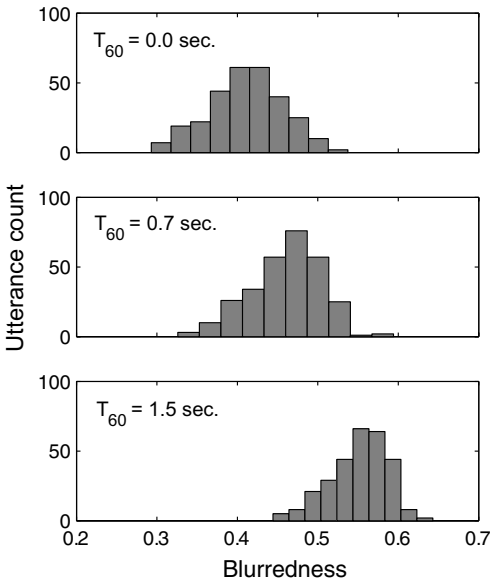

Fig. 6. Distributions of blurredness B for three reverberation conditions, computed from a test set of 300 utterances.

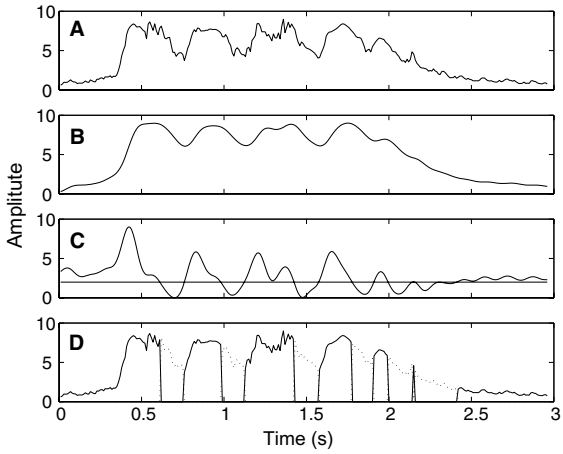

Fig. 7. Demonstration of modulation filtering-based mask estimation. (A) Output of the rate map channel with a centre frequency of 103 Hz. (B) Rate map channel filtered by the lowpass part $h_{\mathrm{lp}}(n)$ of the modulation filter. (C) Rate map channel filtered by the whole modulation filter $h(n)$. The horizontal line indicates the value of the threshold θ. (D) Estimated reliable regions (solid line) and unreliable regions (dotted line). Amplitude maxima in each graph are normalised to fit to the same scale.

cutoff frequency when the highpass cutoff is fixed at 1.5 Hz. The right panel (B) shows the effect of varying the highpass cutoff frequency when the lowpass cutoff frequency is fixed at 8.2 Hz. From these experiments it is clear that recognition performance is mainly determined by the highpass cutoff frequency.

The reverberation mask estimation process is illustrated in Figs. 6 and 7. Fig. 6 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 and 1.5 s. Note that the distribution shifts to the right (i.e., the mean blurredness increases) with increasing reverberation time.

Fig. 7 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 Eqs. (11)–(13) 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 Eq. (10). 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 $\eta_{\mathrm{r}}(j)$ as follows:

$$\eta_{\mathrm{r}}(j) = \frac{1}{L} \sum_{i \in \Gamma_{\mathrm{r}}(j)} y_{\mathrm{c}}(i,j) \tag{14a}$$

$$y_{\mathrm{c}}(i,j) = m_{\mathrm{r}}(i,j) \cdot y(i,j) \tag{14b}$$

Here, $m_{\mathrm{r}}(i,j)$ is the binary reverberation mask and $\Gamma_{\mathrm{r}}(j)$ is the set containing the indices of the L largest values of $y_{\mathrm{c}}(i,j)$ 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 $\Gamma_{\mathrm{r}}(j) = \varnothing$, the scaling factor $\eta_{\mathrm{r}}(j)$ 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 2.0 connected English language digits recognition task (Pearce and Hirsch, 2000). The sampling rate of all speech data was 8 kHz. Auditory rate maps were obtained for the clean training section of the Aurora corpus, and were used to train 12 word-level HMMs (see Section 2.2). The training section contained 8440 clean (noiseless) utterances. For the evaluation of the missing data system we trained two kinds of recognisers, one which used spectrally normalised rate map features (denoted MD-SN) and another which used non-normalised rate map features (denoted MD). It should be emphasised that the MD-SN recogniser was always tested with spectrally normalised features, and the MD recogniser was always tested with non-normalised features.

In the first experiment (see below), the performance of the missing data ASR system was compared against an Aurora 2.0 baseline recogniser. Mel frequency cepstral (MFCC) feature vectors were produced using scripts provided with Aurora 2.0. HMMs were also trained using scripts provided with Aurora. The MFCC feature vectors consisted of 12 mel-cepstral coefficients (the zeroth term was excluded) and logarithmic frame energy. In addition, first and second order temporal derivatives were included, giving a total of 39 features per vector. Aurora 2.0 provides two alternatives for generating MFCC features. We used the version which is based on the HTK implementation (Young et al., 2001) and includes cepstral liftering. The results obtained using these features differ slightly from those for the second version, which are reported by Pearce and Hirsch (2000).

The training configuration of the MFCC baseline system differed from that of the missing data system, in that the MFCC system used only three mixture components to model each state whereas seven components were used for the missing data system trained on rate maps. It was noted that using more mixtures caused the MFCC-based models to overfit to clean speech. All models were trained with clean (noiseless and unreverberated) signals.

For the recogniser evaluation, the Aurora test sets were used as specified in the following subsections. The data corresponding to these tests is presented in Tables 2–5. Each data point in these tables corresponds to a recognition accuracy for one noise condition averaged over 1001 utterances. In the experiments involving reverberation or only spectral distortion, test utterances were convolved with either a room or microphone 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, following the configuration presented in (Kingsbury, 1998, pp. 148–152). 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. Modulation filtered spectrogram features were obtained using the SPRACHcore computer program (version nogui-2001-05-14). However, the program required some modification in order to produce the desired features. Specifically, the modulation filters were configured as a 8 Hz lowpass filter and 8–16 Hz bandpass filter (see Kingsbury, 1998, pp. 148–149), corresponding to filter files lo0_hi8_n21_dn5.sos and lo8_hi16_n21.sos respectively. Also, bandpass features in adjacent frequency bands were summed (Kingsbury, 1998, p. 149, Fig. 5.2).

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 $189\times488\times25$ (input layer $\times$ hidden layer $\times$ output layer) for PLP features alone and $189\times328\times25$ for MSG features alone. For the

Table 2
Speech recognition accuracy (percent) and mean estimator for the missing data mask for non-distorted test cases

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
	MASK-AVE	0.12	0.16	0.20	0.26	0.31	0.36	0.62
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
	MASK-AVE	0.12	0.15	0.20	0.24	0.30	0.36	0.62

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 mean estimator for the mask is indicated by MASK-AVE. The test cases are subway noise and street noise, added at SNRs between −5 and 20 dB. Results for clean speech are also shown.

Table 3
Speech recognition accuracy (percent) for spectrally distorted test cases

Noise type	Method	−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

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 and 20 dB. The 'clean' column indicates performance when the respective filtering characteristic was applied to speech without added noise.

Table 4
Speech recognition accuracy (percent) in the gain modulation test

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

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 and 20 dB, and for clean speech (i.e., gain modulation but no added noise).

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 2.0 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 Kingsbury (1998, p. 45) for details).

Table 5
Speech recognition accuracy (percent) in the reverberation task

T_{60} and source–receiver distance	Hybrid PLP	Hybrid MSG	Hybrid MSG + PLP	MD-SN	MASK-AVE
1.5 s, 18.3 m	53.3	53.5	59.8	64.3	0.36
1.5 s, 6.1 m	55.2	62.0	64.0	67.8	0.36
1.2 s, 3.05 m	59.1	66.6	69.5	76.6	0.40
1.2 s, 2.0 m	60.2	71.3	71.5	78.4	0.41
0.7 s, 3.05 m	88.0	93.0	93.5	92.4	0.57
0.7 s, 2.35 m	89.5	94.0	95.1	93.1	0.60
Unreverberated	98.2	98.0	98.5	97.0	0.80

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-SN). The mean estimator for the missing data mask is indicated by MASK-AVE.

Our system differed in some respects from Kingsbury's (1998) original system. Firstly, different corpora were used; Kingsbury used the Numbers-95 corpus (Cole et al., 1995), whereas we used Aurora 2.0. Secondly, the vocabulary size of the two corpora were different; the Numbers-95 corpus contains 30 words whereas Aurora has only 11 words. Thirdly, we used a smaller number of 23 phonemes, whereas Kingsbury used 32 (Kingsbury, 1998, p. 112, 158). Finally, Kingsbury used a simple bigram language model which we did not use. This choice was made in order to allow a fair comparison against the missing data recogniser, which does not use a language model.

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 2.0 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 types of additive noise. We chose to use parts of these test sets that contain speech with added subway noise A1 (test set A, noise 1) and car noise B2 (test set B, noise 2). 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—including utterances with added subway and car noises—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 (subway and car noises) in the spectrally matching test sets (A1 and B2) and spectrally mismatching case (C1 and C2). 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 Fig. 8.

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). The mask estimation process remains the same for these two systems. We also compare the performance of missing data

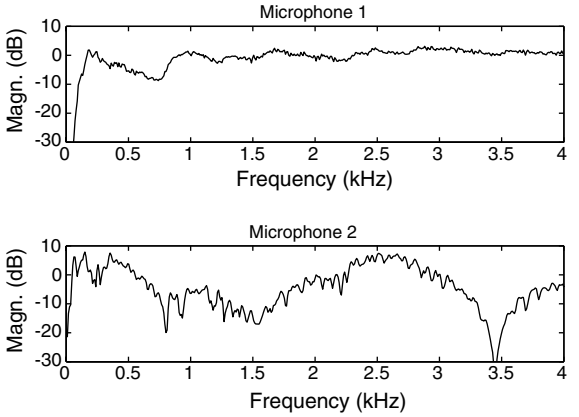

Fig. 8. Frequency responses of the two microphones used in the second experiment.

systems against the MFCC baseline system, generated as recommended within the Aurora 2.0 framework (see Section 5.1). The results of the experiment are shown in Tables 2 and 3. In addition, Table 2 shows the mean estimator for the mask values computed over all (real-valued) masks for the corresponding test cases. As expected, the mean estimator decreases with decreasing SNR, indicating that there is less reliable evidence of the speech signal as the SNR falls. 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 SNRs (Table 2). 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 3) 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 2.0 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-world acoustic environments. Here, we demonstrate that our proposed spectral normalisation scheme also improves robustness when the input signal is subject to overall gain modulation.

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.

Gain modulation tests are shown in Table 4 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 and 10 dB. Comparison with Table 2 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 s, 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 s.

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 using a chirp exited system identification program (Kingsbury, 1998, p. 90). We have verified the quality of the responses by examining Schröder plots of the responses on a dB scale. All of the response have been truncated before they reach the background noise level, and hence they include only the linearly decreasing phase (in dB) of the response. For the first wall panel setting the T_{60} was 0.7 s, the distances between the source and microphone were 2.35 and 3.05 m, and the tail of response was truncated at −55 dB below the level of direct sound. For the second wall panel setting, the T_{60} was 1.2 s, source-microphone distances were 2.0 and 3.05 m, and the tail of response was truncated at −35 dB below the level of direct sound. Another two impulse responses (not used by Kingsbury) were measured in a larger room, having a T_{60} of 1.5 s and source-microphone distances of 6.1 and 18.3 m. The tails of these responses were truncated at −50 dB below the level of direct sound.

The results of this experiment are shown in Table 5. 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.

Table 5 also shows the mean estimator of the mask value for each reverberation condition. In this experiment binary masks were used (see Section 4.1), and hence the mean estimator can be interpreted as the proportion of time–frequency regions which are marked as reliable. As expected, the mean mask value falls with increasing reverberation time.

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.

The missing data systems did not perform as well as the MFCC baseline system in the clean and 20 dB SNR conditions (Tables 2–4). This can possibly be explained as follows. Firstly, both the baseline system and missing data systems model observations using Gaussian mixtures with diagonal covariance, i.e., they assume statistical independence between features. This assumption holds better for the case of MFCCs than for the case of rate map features. One solution to this problem would be to use a full covariance structure, but this has been found to be computationally prohibitive when used in the missing data framework (Morris et al., 1998). Secondly, for clean speech there is a small decrease in the performance of the missing data recogniser when spectral normalisation is applied (for example, see Table 2). Inevitably, any data normalization scheme makes some generalizations over the data in order to reduce variability between different noise conditions. Whilst this improves performance in convolutional noise, some sensitivity to the true variability of clean speech samples may be lost.

The utterance-by-utterance spectral normalisation technique described here probably requires further development before it is suitable for real-world speech recognition applications. For example, long silent pauses in speech would cause the normalisation factor to be biased, since it is

inversely proportional to the utterance length (see Eq. (7)). Also, the proposed scheme seeks to identify the strongest speech regions with a simple algorithm that is blind to the content of the audio signal. It is therefore possible that intense noise regions during a speech pause will be selected instead of clean speech. We note, however, that the above mentioned problems may be shared with other utterance-by-utterance normalisation schemes. One common example of this is the use of MFCCs with cepstral mean subtraction (Atal, 1974; Rosenberg et al., 1994).One possible solution would be to combine a frame-by-frame adaptive normalisation scheme for constant speech flow (e.g. Kingsbury, 1998) with the proposed spectral normalisation technique.

Our concern in this paper was convolutional interference, rather than additive noise. Hence, the current mechanism for dealing with additive noise is simplistic: SNR masks are derived from a stationary noise estimate made during a silent period at the beginning of each utterance. More sophisticated methods for adaptive SNR estimation have been proposed (e.g., Kleinschmidt and Hohmann, 2003; Dupont and Ris, 1999; Hirsch and Erlicher, 1995; Martin, 1993), and it would be straightforward to integrate such algorithms into a missing data recogniser. Future work will address this issue.

In the random gain experiment, the gain varied randomly between different utterances but did not vary within each utterance. Clearly, the conditions of this experiment favour the utterance-by-utterance normalisation scheme proposed here. It should be noted that gain deviations larger than those observed in the Aurora corpus could also occur within each utterance, due to changes in the speaker-microphone distance. Nonetheless, our random gain experiment could be considered an approximate model for the gain changes that occur when a telephone centre receives successive calls from different callers.

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 s and greater. In the least reverberated cases, however, the baseline system outperformed our missing data system. 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. We also note that our comparison of the missing data and baseline systems uses a relatively crude metric; overall recognition accuracy. Further insight could be gained into the relative performance of these techniques by examining the kinds of confusion made at the word and phone levels by each system. This will be addressed in future work.

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 bandpass 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 3 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; hence the filter parameters must be chosen prior to training. This, in turn, might lead to a compromise because the use of particular features may effectively tune the ASR system to certain acoustic conditions. 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 front-end configurations. We note, however, that our system also includes parameters whose values must be set prior to

training, such as the parameter D in the spectral normalisation algorithm.

In designing a modulation filter for this study, we sought only to find the best-performing filter which had a smooth bandpass characteristic. Conceivably, there may be a better performing filter of more complex design. For example, Kingsbury (1998) best performing solution was to produce MSGs from two differently modulation filtered spectral features, those that were bandpass filtered (8–16 Hz) and others that were low-pass filtered (8 Hz). Another open question is whether rate maps are the optimal substrate for the modulation filtering approach presented here.

The experimental results presented in this paper were generated for a small vocabulary of only 11 words. Clearly, transferability of the current system to a larger vocabulary task is an important question. Raj (2000), Raj et al. (in press) and Luo and Du (2003) have successfully applied missing data techniques to large vocabulary tasks.

Finally, we note that 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, allowing different types of front-end to 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. In practice, switching between systems would require some kind of situation classification algorithm (e.g., Eronen et al., 2003; Li et al., 2001; Martin, 1999; Peltonen et al., 2002; Akbacak and Hansen, 2003). Future work will investigate this possibility.

Acknowledgements

We thank two anonymous reviewers for their incisive comments. 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

Akbacak, M., Hansen, J.H.L., 2003. Environmental sniffing: noise knowledge estimation for robust speech systems. Proc. ICASSP-2003 II, 113–116.

Avendano, C., 1997. Temporal processing of speech in a time-feature space, PhD thesis, Oregon graduate institute.

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. Proc. Eurospeech-2001, 213–217.

Barker, J., Green, P.D., Cooke, M.P., 2001b. Linking auditory scene analysis and robust ASR by missing data techniques, In: 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.

Cole, R.A., Noel, M., Lander, T., Durham, T., 1995. New telephone speech corpora at CSLU. Proc. Eurospeech-1995 I, 821–824.

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. Amer. 95, 1053–1064.

Dupont, S., Ris C., 1999. Assessing local noise level estimation methods. In: Proc. of Workshop on Robust Methods for Speech Recognition in Adverse Environments, pp. 115–118.

Eronen, A., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T., Lorho, G., Huopaniemi, J., 2003. Audio context awareness—acoustic modeling and perceptual evaluation. Proc. ICASSP-2003 V, 529–532.

Gold, B., Morgan, N., 2000. Speech and Audio Signal Processing. John Wiley and Sons, Inc., New York.

Gölzer, H., Kleinschmidt, M., 2003. Importance of early and late reflections for automatic speech recognition in reverberant environments. Proc. Elektronische Sprachsignalverarbeitung (ESSV).

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., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Proc. 2, 578–589.

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

Hirsch, H.G., Erlicher, C., 1995. Noise estimation techniques for robust speech recognition. Proc. ICASSP-1995 I, 153–156.

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, New York.

ISO 3382, 1997. Acoustics—Measurement of the Reverberation Time of Rooms with Reference to Other Acoustical Parameters, second ed. ISO, Geneva, Switzerland.

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.

Kandera, N., Arai, T., Hermansky, H., Pavel, M., 1999. On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Comm. 28, 43–55.

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

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

Kleinschmidt, M., Hohmann, V., 2003. Sub-band SNR estimation using auditory feature processing. Speech Comm. 39 (1–2), 47–64.

Li, D., Sethi, I.K., Dimitrova, N., McGee, T., 2001. Classification of general audio data for content-based retrieval. Pattern Recognition Lett. 22, 533–544.

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

Luo, Y., Du, L., 2003. Single gauss model set-based data imputation method for complex ASR task. Proc. ISCAS 2003 II, 564–567.

Martin, K.D., 1999. Sound source recognition. PhD thesis, MIT, Massachusetts.

Martin, R., 1993. An efficient algorithm to estimate the instantaneous SNR of speech signals. Proc. Eurospeech-1993, 37–40.

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.

Morris, A.C., Cooke, M.P., Green, P.D., 1998. Some solutions to the missing feature problem in the classification, with application to noise-robust ASR. Proc. ICASSP-1998 II, 737–740.

Nabelek, A.K., Robinson, P.K., 1982. Monaural and binaural speech perception in reverberation for listeners of various ages. J. Acoust. Soc. Amer. 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.

Oppenheim, A.V., Schafer, R.W., 1989. Discrete Time Signal Processing. Prentice Hall.

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

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., in press. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. Speech Comm.

Patterson, R.D., Holdsworth, J.W., Nimmo-Smith, I., Rice, P., 1988. SVOS Final Report: The Auditory Fiterbank. APL Report 2341.

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.

Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J., Sorsa, T., 2002. Computational auditory scene recognition. ICASSP-2002 II, 1941–1944.

Raj, B., 2000. Reconstruction of incomplete spectrograms for robust speech recognition. PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Raj, B., Seltzer, M.L., Stern, R.M., in press. Reconstruction of Missing Features for Robust Speech Recognition, Speech Comm.

Rosenberg, A.E., Lee, C.-H., Soong, F.K., 1994. Cepstral channel normalisation techniques for HMM-based speaker verification. Proc. ICSLP 94 4, 1835–1838.

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

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D.,Valtchev, V., Woodland, P., 2001. The HTK Book, Revised for Version 3.1, Cambridge University Engineering Department, Available from <htk.eng.cam.ac.uk/>.