Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 43 (2004) 361–378
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation

Kalle J. Palomäki [a,b,c,*], Guy J. Brown [a], DeLiang Wang [d]

[a] *Department of Computer Science, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, United Kingdom*
[b] *Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, P.O.B. 3000, FIN-02015 HUT, Finland*
[c] *Apperception and Cortical Dynamics (ACD), Department of Psychology, University of Helsinki, P.O.B. 9, FIN-00014, Finland*
[d] *Department of Computer and Information Science and Center for Cognitive Science, The Ohio State University, Columbus, OH 43210, USA*

Received 10 May 2002; received in revised form 12 May 2003; accepted 31 March 2004

Abstract

In this study we describe a binaural auditory model for recognition of speech in the presence of spatially separated noise intrusions, under small-room reverberation conditions. The principle underlying the model is to identify time–frequency regions which constitute reliable evidence of the speech signal. This is achieved both by determining the spatial location of the speech source, and by grouping the reliable regions according to common azimuth. Reliable time–frequency regions are passed to a 'missing data' speech recogniser, which performs decoding based on this partial description of the speech signal.

In order to obtain robust estimates of spatial location in reverberant conditions, we incorporate some aspects of precedence effect processing into the auditory model. We show that the binaural auditory model improves speech recognition performance in small room reverberation conditions in the presence of spatially separated noise, particularly for conditions in which the spatial separation is 20° or larger. We also demonstrate that the binaural system outperforms a single channel approach, notably in cases where the target speech and noise intrusion have substantial spectral overlap.

Keywords: Binaural model; Speech recognition; Precedence effect; Missing data

1. Introduction

Human speech perception is remarkably robust. Listeners can follow a conversation in the presence of background noise, even in cases where two or

* Corresponding author. Address: Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, P.O.B. 3000, FIN-02015 HUT, Finland. Tel.: +358 9 4512883; fax: +358 9 460224.
E-mail addresses: kalle.palomaki@hut.fi (K.J. Palomäki), g.brown@dcs.shef.ac.uk (G.J. Brown), dwang@cis.ohio-state.edu (DeLiang Wang).

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.03.005

more speakers are simultaneously active (Yost, 1997; Hawley et al., 1999). Similarly, speech perception adapts quickly to the characteristics of an acoustic environment, and can tolerate the spectral distortion introduced by moderate reverberation (Nabelek and Robinson, 1982) or by a transmission channel (Watkins, 1991).

In contrast to human performance, automatic speech recognition (ASR) in noisy or reverberant acoustic environments remains very problematic. It is reasonable to argue, therefore, that ASR performance could be improved by adopting an approach that models the known mechanisms of auditory processing more closely (for example, see Hermansky, 1998). Additionally, such auditory models may contribute to our understanding of human hearing by clarifying the computational processes involved in speech perception.

The robustness of human speech perception has its foundation, at least in part, in the ability of the auditory system to perceptually segregate a target sound from an acoustic mixture. The process by which listeners parse a mixture of sounds in order to retrieve a description of a particular sound source has been termed *auditory scene analysis* (ASA) by Bregman (1990). Conceptually, ASA may be regarded as a two-stage process. In the first stage, the sound reaching the ears is decomposed into sensory elements. In the second stage—termed *auditory grouping*—elements that are likely to have arisen from the same environmental event are combined to form a perceptual stream. Streams are subjected to higher-level processing, such as language understanding.

Auditory grouping is known to exploit acoustic cues which are related to common spectro-temporal properties of sound (see Darwin and Carlyon, 1995; for a review). Additionally, ASA uses information about the spatial location of sound sources, which is principally encoded by interaural time difference (ITD) and interaural level difference (ILD) cues at the two ears. Indeed, it has been appreciated since the early 1950s that the intelligibility of speech in the presence of another utterance is improved when the target and competing sentences originate from different locations in space. For example, Spieth et al. (1954) found that the intelligibility of two overlapping speech signals improved as the spatial separation between them was increased. For large spatial separations (between 90° and 180°) the number of utterances correctly identified improved by as much as 20%. More recent studies have considered speech intelligibility in the presence of multiple competing sentences, and they indicate that the intelligibility of the target is principally determined by the proximity of the competing speech to the target location (Hawley et al., 1999).

Computational approaches to ASA (for a review see Rosenthal and Okuno, 1998) have been described which are able to segregate speech from interfering noise, and these have been employed as front-ends to ASR systems with promising results (e.g., Barker et al., 2000a,b; Brown et al., 2001). However, the large majority of this work has only addressed monaural grouping mechanisms. A number of systems that incorporate binaural grouping cues have been described (Denbigh and Zhao, 1992; Bodden, 1993; Glotin et al., 1999; Okuno et al., 1999; Shamsoddini and Denbigh, 2001; Roman et al., 2002). However, only a few of these sound separation algorithms have been evaluated in reverberant conditions, presumably because of the difficulty of the task. Notable exceptions are the systems described by Denbigh and Zhao (1992), Shamsoddini and Denbigh (2001) and Bodden (1993), although these do not explicitly model the auditory mechanisms that are thought to allow robust localisation of sound sources in reverberant environments (the 'precedence effect'; see Section 2.2.1).

From a purely engineering perspective, ASR systems that are intended to operate in the presence of reverberation or background noise typically exploit microphone array processing and blind source separation (for example, see Koutras et al., 2001; Seltzer and Raj, 2001). These approaches are not informed by principles of human auditory function, and typically make stronger assumptions about the number of sound sources and their characteristics than computational ASA systems (see van der Kouwe et al., 2001). In cases where only a single noisy and/or reverberated speech channel is available, the usual approach is to employ noise-robust feature vectors such as RASTA (Hermansky and Morgan, 1994)

or modulation-filtered spectrograms (Kingsbury et al., 1998). Interestingly, such acoustic features are often motivated by principles of auditory processing, such as forward masking and limited spectral resolution.

Recently, progress has also been made in developing ASR systems that apply principles of auditory function at processing stages beyond the initial extraction of acoustic features. Cooke and his co-workers (Cooke et al., 2001) have interpreted the robustness of speech perception mechanisms in terms of their ability to deal with 'missing data', and have proposed an approach to ASR in which a hidden Markov model (HMM) classifier is adapted to cope with missing or unreliable features. The missing data paradigm is complementary to computational ASA; an auditory model can be used to decide which acoustic components belong to a target speech source, and only these 'reliable' features are passed to the recogniser (for example, see Brown et al., 2001; Palomäki et al., 2001).

In this study, we propose a perceptually inspired approach to computational ASA which is able to segregate a target speech signal from interfering sound sources on the basis of spatial location. The current study extends our earlier work (Palomäki et al., 2001) in a number of important respects. First, we describe an improved model of binaural perception which is able to identify the spatial location of acoustic sources more robustly in the presence of background noise and reverberation. Secondly, we describe a novel strategy for performing spectral energy normalisation of acoustic features within a missing data ASR system. Finally, we present a detailed evaluation of the model in a number of reverberant acoustic environments, and compare its performance with a conventional speech recognizer that uses noise-robust feature vectors. We show that the model provides an effective front-end for missing data recognition of speech in noisy and reverberant conditions, and outperforms the conventional approach in most cases.

2. Model

The model (Fig. 1) is divided into monaural and binaural pathways. The monaural pathway is responsible for peripheral auditory processing, and produces feature vectors for the speech recogniser. The binaural pathway is responsible for sound localisation and separation according to common azimuth. Acoustic input to the model is obtained by spatialising speech and noise signals using a model of small room acoustics and realistic head-related impulse responses (HRIRs) (see Section 3.1 for details).

2.1. Monaural pathway

The acoustic inputs to the left and right ears of the model, $x_{\text{left}}(n)$ and $x_{\text{right}}(n)$, are initially processed by a simulation of the auditory periphery. Cochlear frequency analysis is modelled by a bank of $N = 32$ gammatone filters (Patterson et al., 1988; see also Brown and Cooke, 1994), with

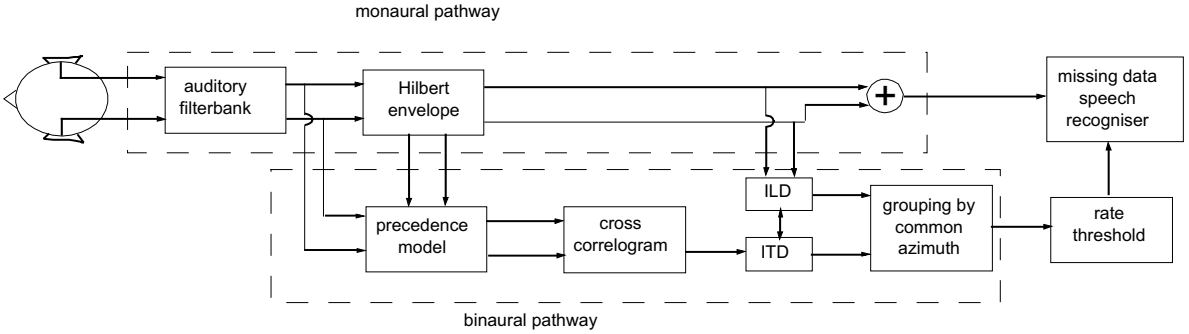

Fig. 1. Schematic diagram of the model.

centre frequencies equally spaced on the equivalent rectangular bandwidth (ERB) scale (Glasberg and Moore, 1990). We use fourth-order gammatone filters of the form

$$g(t) = t^3 \exp(-2\pi bt)\cos(2\pi f_0 t)u(t) \quad (1)$$

where f_0 is the centre frequency, $u(t)$ is the unit step function (i.e., $u(t) = 1$ for $t \geqslant 0$, 0 otherwise) and b is proportional to the ERB bandwidth. The lowest centre frequency was 50 Hz, and the highest was 8 kHz. Here, we use a digital implementation of the gammatone filter described by Cooke (1993). The output of each filter is half-wave rectified, giving a simple simulation of the auditory nerve response for each frequency channel which we denote by $a_{i,k}(n)$ for channel i of ear k, with $k \in \{\text{left,right}\}$.

The envelope of the auditory nerve response is required for two reasons; to obtain features for the speech recogniser (a vector of channel energies for each time frame), and to generate an inhibitory signal that is used in our simulation of precedence effect processing (see Section 2.2.1). Cooke (1993, Appendix C) shows that the cosine term in (1) can be replaced with a complex sinusoid, giving a (complex) filter whose output is a close approximation to the analytic signal of the input (see also Cohen, 1994). Hence, the instantaneous Hilbert envelope can be obtained directly from the complex gammatone filter coefficients by

$$\varepsilon(n) = \sqrt{\Re^2(n) + \Im^2(n)} \quad (2)$$

where $\Re(n)$ and $\Im(n)$ represent the output of the real and imaginary parts of the complex gammatone filter. To obtain features for the speech recognizer, the envelope $\varepsilon(n)$ of each channel is smoothed by a one-pole lowpass filter $H(z) = 1/(1 + az^{-1})$ with a time constant of 8 ms (i.e., $a = \exp(-1000/f_s \times 8)$ where $f_s = 20{,}000$ Hz is the sample rate). The smoothed envelope is sampled at 10 ms intervals and compressed by raising it to the power of 0.3. The resulting representation, which we denote by env(j), can be interpreted as an estimate of auditory nerve firing rate (we use the index j to denote frame number). When env(j) is plotted for each channel over time a 'rate map' is obtained, which may be regarded as an auditory spectrogram.

The binaural model produces an estimate of firing rate for the left and right ears, $\text{env}_L(j)$ and $\text{env}_R(j)$. In some experimental cases (see Section 3) recognition is based on the firing rate in one ear only, but in most cases the feature vectors passed to the recognizer consist of a linear combination of the features from the two ears:

$$\text{env}_{LR}(j) = \left[\frac{1}{2}\left(\text{env}_L(j)^{3.333} + \text{env}_R(j)^{3.333}\right)\right]^{0.3} \quad (3)$$

2.2. Binaural pathway

Human listeners primarily use interaural time difference (ITD) and interaural level difference (ILD) cues to localize sound sources in space (see Moore, 1997 for a review). Here, we describe a simple computational model of sound localisation which employs both cues, and incorporates aspects of precedence effect processing.

2.2.1. Modelling the precedence effect

The term 'precedence effect' refers to a group of psychophysical phenomena which are believed to underlie the ability of listeners to localise sound sources in reverberant environments (see Wallach et al., 1949; and more recent reviews by Zurek, 1987; Blauert, 1997; Litovsky et al., 1999). In such environments, direct sound is closely followed by multiple reflections from different directions; however, listeners usually report that the sound has originated from one direction only. The perceived location corresponds to the direction of the first wavefront; hence it appears that the directional cues in the first-arriving sound are given 'precedence' over cues contained in the later reflections.

The precedence effect is relevant to the study described here because we consider localisation of speech and noise sources in reverberant conditions. We note that an earlier version of our binaural CASA system did not include precedence effect processing, and was unable to accurately locate sound sources when reverberation was present (Palomäki et al., 2001). The current model rectifies this deficiency.

Several computational models of sound localisation have been proposed which incorporate aspects of precedence effect processing (for example, see Lindemann, 1986; MacPherson, 1991; Martin, 1997). These models focus on peripheral (rather than cognitive) factors and suggest that the precedence effect is underlain by an inhibitory mechanism that suppresses echoes. The closest in approach to the model we present here is the work of Martin (1997), who describes a computational implementation of Zurek's (1987) phenomenological model. In Martin's model, instantaneous information about the ITD of a sound source is suppressed by an inhibitory input which begins approximately 1 ms after the onset of an abrupt sound. Suppression is strong for a few milliseconds, and then recovers over a time scale of approximately 10 ms in accordance with Zurek's data.

Here, we adopt an approach in which acoustic onsets are emphasized in the simulated auditory nerve response prior to ITD analysis. The instantaneous envelope is computed from each gammatone filter as described in (2), and this is low-pass filtered in order to produce an inhibitory signal which is a smoothed and time-delayed version of the auditory nerve response. The low-pass filter has an impulse response of the form

$$h_{\mathrm{lp}}(n) = An\exp\left(\frac{-n}{\alpha}\right) \tag{4}$$

where the constant A is chosen to give unity gain at DC, and α is a time constant. The low-pass filter has a time constant of 15 ms, corresponding to $\alpha = 300$ samples at the sample rate used (20 kHz). When the inhibitory signal is subtracted from the auditory nerve response, sustained activity is suppressed but transients (which tend to contain reliable localisation information) are preserved. Specifically, the auditory nerve response is transformed by

$$r_{i,k}(n) = [a_{i,k}(n) - G(h_{\mathrm{lp}}(n) \bullet \varepsilon_{i,k}(n))]^{+} \tag{5}$$

where $a_{i,k}(n)$ is the auditory filter response for channel i of the kth ear, $\varepsilon_{i,k}(n)$ is the corresponding instantaneous envelope and G is a constant gain term that determines the strength of inhibition. The operator $[\,]^{+}$ indicates half-wave rectification, and the symbol $\bullet$ denotes convolution. We set $G = 1.0$ by inspection. Fig. 2 shows the output from each stage of processing in the precedence effect model, for a filter channel centered on 1 kHz that is responding to a transient speech sound.

2.2.2. Cross-correlogram

ITD is estimated by computing the cross-correlation between the output of the precedence processed auditory filter response at the two ears. Given the output of the precedence effect model for the left and right ear in channel i, $r_{i,\mathrm{L}}(n)$ and $r_{i,\mathrm{R}}(n)$, the cross correlation for delay τ and time frame j is

$$C_i(j,\tau) = \sum_{n=0}^{M-1} r_{i,\mathrm{L}}(jT - n) r_{i,\mathrm{R}}(jT - n - \tau) w(n) \tag{6}$$

where w is a window of width M time steps and T is the frame period (10 ms, or 200 samples). Currently, we use a rectangular window with $M = 600$, corresponding to a duration of 30 ms, and consider values of τ between -1 and 1 ms. For efficiency, the fast Fourier transform is used to evaluate (6) in the frequency domain. Computing $C_i(j,\tau)$ for each channel i ($1 \leqslant i \leqslant N$) gives a *cross-correlogram*, which is computed at 10 ms intervals of the time index j. The upper panel of Fig. 3A shows a cross-correlogram for one frame of a mixture of female and male speech, in which the voices originate from azimuths of $-20°$ and $+20°$, respectively.

Ideally, the cross-correlogram should exhibit a 'spine' at the delay τ corresponding to the ITD of a sound source. This feature can be emphasized by summing the channel cross-correlation functions, giving a *pooled cross-correlogram*, $P(j,\tau)$, which is shown in the lower panel of Fig. 3A:

$$P(j,\tau) = \sum_{i=1}^{N} C_i(j,\tau) \tag{7}$$

In free-field listening conditions, diffraction effects introduce a weak frequency-dependence to ITDs which is evident in the HRIR-filtered stimuli used here. As a result, the 'spine' can be unclear and (7) does not exhibit a clear peak at the ITD. We note that this complication does not arise in many cross-correlation models of ITD analysis (e.g.,

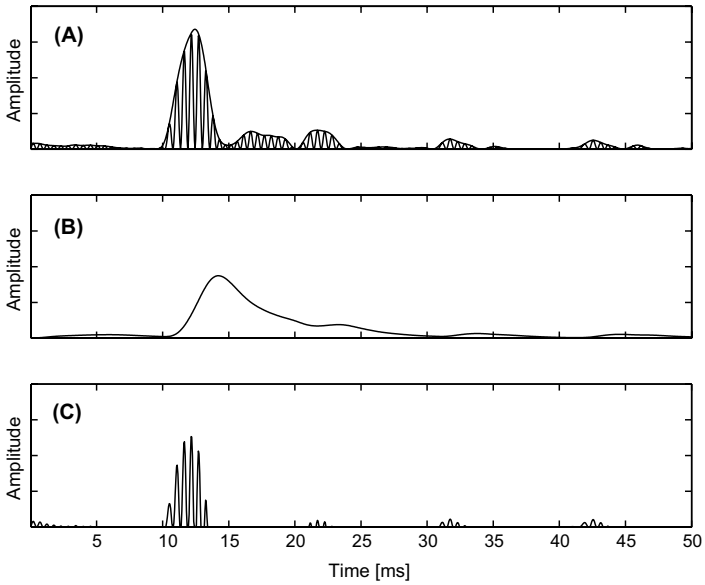

Fig. 2. Stages in precedence effect processing. (A) Output from one channel of the auditory model with centre frequency 1 kHz in response to speech, showing the rectified gammatone filter response and superimposed instantaneous envelope. (B) Inhibitory signal, obtained by lowpass filtering the envelope. (C) Inhibited gammatone filter response $r_{i,k}(n)$. Note that the filter response following a large peak is suppressed.

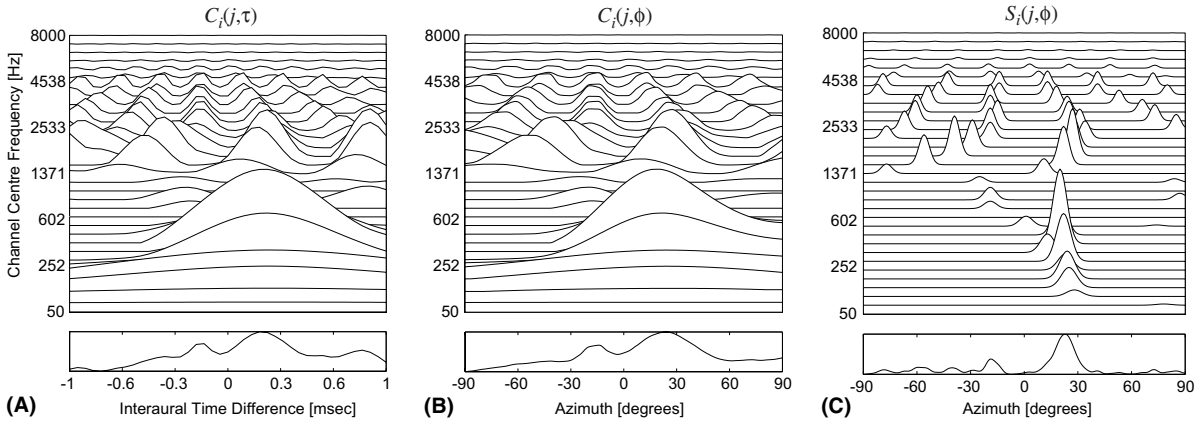

Fig. 3. Stages of binaural processing in the auditory model. Each panel shows the cross-correlogram (top) and corresponding pooled cross-correlogram (bottom) for one frame of a mixture of spatialised female speech and male speech. The two sound sources were located at $-20°$ and $+20°$ azimuth respectively. (A) Cross-correlogram $C_i(j,\tau)$, where τ indicates the interaural time difference (ITD) in ms. (B) Cross-correlogram $C_i(j,\phi)$ after warping of each channel from ITD to azimuth (ϕ). (C) Skeleton cross-correlogram $S_i(j,\phi)$. Note the improved sharpness of each peak in the pooled cross-correlogram.

Shackleton et al., 1992), because they only consider stimuli in which ITDs are synthetically generated and are therefore consistent across frequency. Here, we address this issue by warping each cross-correlation function (6) to an azimuthal axis, giving a modified cross-correlogram of the form $C_i(j,\phi)$ where ϕ is azimuth in degrees. The azimuth is quantised to a resolution of 1°, giving 181 points

between −90° and +90°. Warping is achieved by a table look-up, which relates the azimuth in degrees to its corresponding ITD in each channel of the auditory model. The functions relating azimuth to ITD were derived from cross-correlograms of random noise and are monotonic, being sigmoidal at low frequencies (where diffraction effects are greatest) and increasingly linear at high frequencies. A warped cross-correlogram and its corresponding pooled version, $P(j,\phi)$ are shown in Fig. 3B.

2.2.3. The skeleton cross-correlogram

A further stage of processing is motivated by the observation that the true position of peaks in the cross-correlogram can be obscured by the filtering characteristic of each frequency band. In particular, low frequency channels produce very broad peaks. To address this problem, we introduce the notion of a *skeleton* cross-correlation function. For each channel of the cross-correlogram, a skeleton function $S_i(j,\phi)$ is formed by superimposing Gaussian functions at azimuths corresponding to local maxima in the corresponding cross-correlation function, $C_i(j,\phi)$. First, each function $C_i(j,\phi)$ is reduced to a form $Q_i(j,\phi)$, which contains non-zero values only at its local maxima. Subsequently, $Q_i(j,\phi)$ is convolved with a Gaussian to give the skeleton function $S_i(j,\phi)$,

$$S_i(j,\phi) = Q_i(j,\phi) \bullet \exp\left(\frac{-\phi^2}{2\sigma_i^2}\right) \tag{8}$$

The standard deviations of the Gaussians, σ_i, vary linearly with frequency channel i, being 4.5 samples in the lowest frequency channel and 0.75 samples in the highest (these parameters were derived empirically using a small data set). This approach is similar in effect to applying lateral inhibition along the azimuthal axis, and causes a sharpening of the cross-correlation response (see Fig. 3C).

2.2.4. Interaural level difference (ILD)

ILD is only computed for frequency bands above 2800 Hz, since there is insufficient 'head shadow' at low frequencies to give an appreciable ILD (Blauert, 1997; Moore, 1997). The ILD in dB is computed as follows:

$$ild_i(j) = 10\log_{10}\left[\frac{\mathrm{eng}_{i,\mathrm{R}}(j)}{\mathrm{eng}_{i,\mathrm{L}}(j)}\right] \tag{9}$$

Here, $\mathrm{eng}_{i,k}(j)$ represents the energy in channel i of ear k at time frame j, obtained by raising the envelope $\mathrm{env}_{i,k}(j)$ to the power of 3.333 (in order to reverse the effect of compression) and then squaring to obtain the energy.

2.3. Grouping by common azimuth

Missing data ASR requires each acoustic feature to be labelled as 'reliable' or 'unreliable' (see Section 2.7). In practice, this information is provided to the recogniser in the form of a time–frequency *mask* (Cooke et al., 2001). Here, we use a binary mask in which a value of unity represents a reliable feature, and a value of zero represents an unreliable feature (we note that real-valued masks may also be employed; see Barker et al., 2000b). Typical masks are shown in Fig. 4.

In our system, the mask is estimated by a process that groups acoustic features according to common azimuth. We assume that sound sources have stationary locations; hence, source locations can be estimated from the mean pooled skeleton cross-correlogram $Z(\phi)$, computed over all K time frames of the input

$$Z(\phi) = \frac{1}{K}\sum_{j=1}^{K}\sum_{i=1}^{N} S_i(j,\phi) \tag{10}$$

The azimuths of the speech and noise sources, ϕ_s and ϕ_n, correspond to the two largest peaks in $Z(\phi)$. In order to allow the speech and noise sources to be discriminated, we assume that $\phi_\mathrm{s} > \phi_\mathrm{n}$ (i.e., that the speech lies to the right of the noise; we discuss how this assumption might be relaxed in Section 4). Values in the mask $m(i,j)$ are then set according to

$$m(i,j) = \begin{cases} 1 & \text{if } C_i(j,\phi_\mathrm{s}) > C_i(j,\phi_\mathrm{n}) \\ & \text{and } C_i(j,\phi_\mathrm{s}) > \Theta_\mathrm{c} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

The heuristic (11) sets the mask to unity (i.e., indicates a reliable region) if channel i is dominated by the speech source at time frame j, as estimated

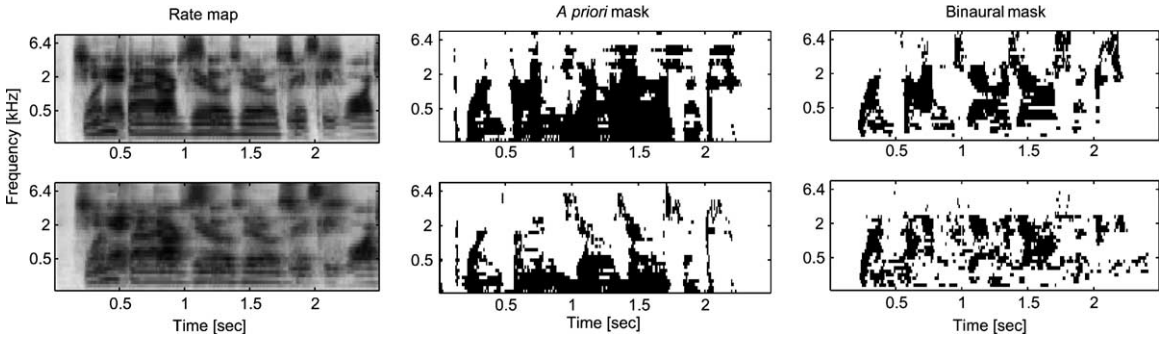

Fig. 4. Mask estimation examples for the utterance "one five zero zero six" in anechoic conditions (top) and for a T_{60} reverberation time of 0.3 s. (bottom), for which the angular separation is 40° and the SNR is 0 dB. From left to right, panels show the rate maps (firing rate of the auditory model), a priori masks and masks estimated from the binaural model.

from the relative height of the peaks in $C_i(j,\phi)$ at the azimuths of the speech and the noise. Additionally, a threshold value for $C_i(j,\phi_s)$ is applied, so that values below Θ_c are discarded. This ensures that time–frequency regions are discarded in which the energy of the speech source is low, but still above that of the noise. Here we set $\Theta_c = 10$ (model units) by inspection.

2.4. ILD constraint

The model checks the ILD in each frequency band for consistency with azimuth estimates from the cross-correlogram, in order to confirm whether each time–frequency region is dominated by the target speech source, by noise from the interfering source or by reflections. First, $ild_i(j)$ is calculated for each channel i at time frame j, as described in (9). The ILD estimate is then verified against the azimuth of the speech source, ϕ_s, derived from ITD cues. We consider that the ILD and ITD are consistent if

$$| ild_i(j) - \Omega_i(\phi_s) | < 0.5 \text{ dB} \tag{12}$$

where $\Omega_i(\phi_s)$ is an ILD template for channel i and azimuth ϕ_s. If the condition (12) does not hold, then the corresponding mask value $m(i,j)$ is set to zero.

The ILD template represents the 'ideal' ILD observed in each frequency channel for a sound source at a specified azimuth (see Roman et al., 2002; for a related approach). The templates are precomputed in each frequency band above 2800 Hz, for 5° increments of azimuthal angle in the frontal horizontal plane. Clean, unreverberated speech was used to generate the ILD templates, although any wideband sound will suffice; near-identical templates have been obtained using broadband noise rather than speech.

2.5. Rate threshold

In the final stage of mask estimation, we apply a threshold to the energy in each frequency channel. A moving average, $E_i(j)$, of the energy is computed for each channel i over a 200 ms window, with a frame shift of 100 ms (note that we use the average energy across the two ears of the model, $\text{eng}_{i,\text{LR}}(j)$, in this process). Then, if the condition

$$10\log_{10}\left[\frac{\text{eng}_{i,\text{LR}}(j)}{E_i(j)}\right] > \Theta_r \tag{13}$$

does not hold, the corresponding mask element $m(i,j)$ is set to zero. Here, Θ_r is a rate threshold which we set to -11 dB by inspection. The heuristic (13) is particularly effective in conditions where the signal-to-noise ratio is low. In such cases, azimuth estimation may be inaccurate (particularly when the angular separation between sources is small) and mask estimation is improved by rejecting time–frequency regions that have low energy, since these are likely to be dominated by the interfering noise.

2.6. A priori mask

We employ a baseline system in which mask estimation is based on a priori information about regions of uncorrupted speech (see also Cooke et al., 2001). This is achieved by measuring the difference (in each channel i at time frame j) between the energy for speech contaminated by noise and/or reverberation, $\mathrm{eng}^{\mathrm{noisy}}_{i,\mathrm{LR}}(j)$, and clean speech $\mathrm{eng}^{\mathrm{clean}}_{i,\mathrm{LR}}(j)$:

$$m_{\text{a priori}}(i,j) = \begin{cases} 1 & \text{if } 10\log_{10}\left[\dfrac{\mathrm{eng}^{\mathrm{noisy}}_{i,\mathrm{LR}}(j)}{\mathrm{eng}^{\mathrm{clean}}_{i,\mathrm{LR}}(j)}\right] < \Theta_{\mathrm{ap}} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

The threshold Θ_{ap} was tuned to give optimal performance for each experimental condition. The a priori masks derived from (14) serve two purposes; they allow the limits of the missing data approach to be tested using a near-optimal mask, and they allow us to test how close to this ideal performance we can achieve using a posteriori information only. Fig. 4 shows two masks generated using the a priori heuristic.

2.7. Missing data speech recogniser

The speech recogniser used in this study employs the missing data technique (Cooke et al., 2001), in which a hidden Markov model (HMM) system is adapted to deal with missing or unreliable data. The classification problem in speech recognition involves the assignment of an acoustic vector Y to a class W, such that the posterior probability $P(W|Y)$ is maximised. However, when a noise intrusion is present or when the speech is corrupted by environmental conditions such as reverberation, some components of Y are likely to be unreliable or missing. In these cases, the acoustic model $P(Y|W)$ cannot be computed as usual. The 'missing data' technique addresses this 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. As noted in Section 2.3, the recogniser is provided with a 'mask' which represents the time–frequency distribution of reliable and unreliable components (see Fig. 4).

In the simplest approach, the unreliable components are simply ignored, so that classification is based on the marginal distribution $P(Y_{\mathrm{r}}|W)$. However, when Y is an acoustic vector it is usually known that the uncertain components have bounded values, and this information can be exploited during classification using the so-called 'bounded marginalisation' method (Cooke et al., 2001). Here, we use bounded marginalisation in which Y is an estimate of auditory nerve firing rate, so the lower bound for Y_{u} is zero and the upper bound is the observed firing rate.

2.8. Spectral energy normalisation

In order to achieve robustness to the convolutional distortion caused by reverberation and HRIR filtering, a spectral normalisation method was developed. In conventional ASR systems (for example, see Kingsbury, 1998), acoustic feature vectors are often normalised by the spectral mean and variance in each frequency band. However, such an approach is not effective in the presence of non-stationary noise (such as an interfering speaker), because clean regions of the speech signal may be normalised by a spectral mean and variance that are computed when both speech and noise sources are present.

This problem can be addressed within the missing data framework by computing a normalisation factor from the reliable components Y_{r} only, as indicated by the time–frequency mask. 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 reliable features in that channel. More specifically, we compute a normalisation factor η_i for channel i as follows:

$$\eta_i = \frac{1}{L}\sum_{l\in\Gamma} \mathrm{env}_{i,\mathrm{LR}}(l) \tag{15}$$

Here, Γ is a set containing the indices of the L largest values of $\mathrm{env}_{i,\mathrm{LR}}(j)$ for which $m(i,j)$ is unity. It is important to emphasise here that $\mathrm{env}_{i,\mathrm{LR}}(j)$ contains only positive values. The rationale for (15) is

that a normalising factor should be computed from time–frequency regions which are labelled as reliable in the missing data mask and have high energy, since they are least likely to be corrupted by an interfering sound.

Generally, L is set to K/B, where K is the number of time frames in the input and B is a constant (we use $B = 15$). However, 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. If the number of reliable regions is zero, the normalisation factor η_i is determined by interpolation from adjacent channels. During training of the recognizer, the whole utterance was included in the search for the L largest values, since the training was performed with clean speech (i.e., all regions are reliable).

3. Evaluation

We have evaluated our system using a variety of reverberation and noise conditions, generated by a simulation of small room acoustics. In the first experimental case, the effect of reverberation on recognition performance was investigated. Secondly, we determined the extent to which recognition performance was influenced by the spatial separation between speech and noise sources. In the third experiment, the effect of the type of noise intrusion on recognition performance was assessed. Finally, we investigated the reliability of azimuth estimation by the model in small-room reverberation conditions for a number of spatial separations in the presence of interfering speech, both with and without precedence effect processing.

3.1. Producing acoustic input using a model of small room acoustics

Sound propagation from the acoustic source to the ear canal is simulated using conventional spatialisation techniques (for an overview, see Møller, 1992). Room reflections are estimated using the image model of small room acoustics (Allen and Berkley, 1979) and then the direction dependent filtering effects of the pinna, head and torso are modelled by convolving the direct sound and reflections with a head-related impulse response (HRIR) for each ear. The set of HRIRs used in this study were measured from the KEMAR artificial head by Gardner and Martin (1994). The dimensions of the simulated room were chosen to mimic a small office (length 6 m, width 4 m and height 3 m). The sound receiver ('listener') was positioned in the middle of the floor at a height of 2 m, and speech and noise were emitted from a distance of 1.5 m at different horizontal angles. The image model approximates the paths from a sound source to a receiver by treating each boundary of the room as a mirror, in which the source is reflected. Sound reflections therefore correspond to direct paths between the mirror-image sources and the sound receiver. The different reverberation times used in our experiments were created by varying the absorption characteristics of the room boundaries according to data for commonly used building materials (Hall, 1991). Typically, acoustic properties of surface materials vary a great deal across frequency and are therefore characterised by absorption coefficients for octave bands, which can be transformed to octave band reflection coefficients for the image model (see also Huopaniemi et al., 1997). Sound propagation is also influenced by air absorption, which introduces a low pass filtering effect that mostly depends on distance and air humidity.

To model the interaction between acoustic space and the listener's head and torso we define a *binaural room impulse response* (Møller, 1992). The total transmission from sound source v to each ear, calculated from R reflection paths, is

$$\begin{aligned} h^v_{\text{left}}(n) &= \sum_{p=0}^{R} h^v_{\text{s,p}}(n) \bullet h^v_{\text{a,p}}(n) \bullet h^v_{\text{left},\phi_\text{p},\theta_\text{p}}(n) \\ h^v_{\text{right}}(n) &= \sum_{p=0}^{R} h^v_{\text{s,p}}(n) \bullet h^v_{\text{a,p}}(n) \bullet h^v_{\text{right},\phi_\text{p},\theta_\text{p}}(n) \end{aligned} \tag{16}$$

Here the term $h^v_{\text{s,p}}(n)$ represents surface absorptions along reflection pathway p, estimated from the image model; $h^v_{\text{a,p}}(n)$ is an air propagation filter (assuming approximately 50% relative humidity) for pathway p; and $h^v_{\text{left},\phi_\text{p},\theta_\text{p}}(n)$ and $h^v_{\text{right},\phi_\text{p},\theta_\text{p}}(n)$ are the left and right HRIRs, where ϕ_p and θ_p

represent the azimuth and elevation at which reflection pathway p strikes the head. The air propagation filter is obtained as follows. Firstly, we estimate an air absorption filter for propagation through 1 m of air, and then raise it to the power of the distance assuming that air is a convolutive medium. It should be noted that circular wave attenuation of sound pressure level is estimated separately according to the $1/r$ law. The techniques applied for room surface modelling and air absorption are described in detail by Lokki (2002).

For simplicity and computational efficiency, the azimuth and elevation of each transmission path were quantised to fit the resolution of the KEMAR HRIR data (see Gardner and Martin, 1994). The elevation angle was rounded to the nearest 10 degrees in the interval $-40°$ to $90°$, and larger negative values were always rounded to -40. The azimuth resolution was $5°$ in the vicinity of the horizontal plane and decreased as the elevation increased to higher positive or negative angles.

Our experiments required acoustic mixtures in which speech and an interfering noise were presented from different spatial locations. These were generated by separately convolving speech $s(n)$ and noise $z(n)$ signals with a binaural impulse response (16) corresponding to the desired location, and summing the spatialised signals at each ear to give a binaural mixture:

$$\begin{aligned} x_{\text{left}}(n) &= [h^s_{\text{left}}(n) \bullet s(n)] + [h^z_{\text{left}}(n) \bullet z(n)] \\ x_{\text{right}}(n) &= [h^s_{\text{right}}(n) \bullet s(n)] + [h^z_{\text{right}}(n) \bullet z(n)] \end{aligned} \tag{17}$$

3.2. Recogniser architecture and corpus

The system was evaluated on a 240 utterance subset of male speakers from the TiDigits connected digits corpus (Leonard, 1984). The sample rate of the speech data was 20 kHz. Excitation patterns were obtained for the training section of the corpus, and were used to train 12 word-level HMMs (a silence model, 'oh', 'zero' and '1–9') each consisting of 8 no-skip, straight-through states with observations modelled by a 10 component diagonal Gaussian mixture. All models were trained on clean, unreverberated signals. The speech recognition accuracy of the system on the clean test set was 98.6%.

The rock music, female speech and male speech from Cooke's (1993) corpus of noise intrusions were used to test the model (Cooke designates these signals as n4, n7 and n8 respectively). The amplitude of each noise intrusion was scaled to give a range of signal-to-noise ratios (SNRs) from 0 to 200 dB.

Noise intrusions and test utterances were convolved with left ear and right ear binaural impulse responses to give spatial separations of $10°$, $20°$ and $40°$. The binaural impulse response incorporated a room impulse response, generated by the image model, to give T_{60} reverberation times of 0, 0.3 or 0.45 s (the T_{60} is the time required for the sound level to drop by 60 dB following sound offset). The spatialised noise and utterance signals were then summed for each ear, giving a binaural mixture.

We also trained the same HMM recognizer with mel-cepstral coefficients (MFCCs) to provide a baseline comparison. Feature vectors consisted of 13 MFCCs, together with their first and second order temporal derivatives. In the MFCC processing chain, signals were first transformed to the mel-spectral domain after which left and right ear signals were summed. Then the combined left and right ear spectrum was compressed logarithmically and transformed to the cepstral domain by taking the discrete cosine transform (DCT). Finally, cepstral mean subtraction (CMS) was applied to the cepstral representation. The accuracy of the MFCC baseline recognizer on the clean test set was 99.4%.

3.3. Experiment 1: effect of reverberation time

The first experiment evaluated the effect of reverberation on speech recognition performance, using T_{60} reverberation times of 0 (anechoic), 0.3 (mildly reverberant office) and 0.45 s ('live' office). Mixtures of male speech and a noise intrusion (another male speaker) were presented at SNRs of 0, 10, 20 and 200 dB in each reverberant condition. The spatial separation between sources was $40°$, with the target speech presented at $20°$ azimuth

and the interfering speech at −20° azimuth (i.e., the target speech was closer to the right ear of the 'listener').

Table 1 shows the speech recognition accuracy for each room condition. Results are given for the MFCC baseline system, and for missing data recognition in which masks have been estimated by the binaural processor (BINAURAL) or by the a priori heuristic (A PRIORI). For all room conditions at SNRs of 20 dB and below, the binaural missing data system achieves substantial improvements in recognition accuracy compared to the baseline system. Additionally, in the anechoic case the binaural processor yields masks which are very close to the a priori masks, giving near-ceiling performance at all SNRs except the lowest (0 dB).

For the 200 dB SNR case, the MFCC baseline system outperforms the binaural missing data recognizer in anechoic and reverberant conditions. However, for this condition there is little difference in performance between the baseline system and missing data recognition with a priori masks. This suggests that binaural mask estimation could be improved to a level that would match the performance of the MFCC system at high SNRs.

As expected, the performance of all three approaches degrades as the reverberation time increases, although the missing data approaches remain relatively robust compared to the baseline system. The largest drop in performance occurs between the 0 and 0.3 s conditions; further increase of the reverberation time to 0.45 s has a smaller effect. Again, we note that with a priori masks the recognition accuracy of the missing data approach remains almost at ceiling in the presence of reverberation, suggesting that mask estimation could be improved to give a higher level of performance.

3.4. Experiment 2: spatial separation

The second experiment investigated the effect of the spatial separation between speech and noise sources on recognition performance, using angular separations of 10°, 20° and 40°. In each condition, target speech and interfering male speech were presented symmetrically about the median plane at azimuth angles of (−5, 5), (−10, 10) and (−20, 20), and were scaled to give SNRs of 0, 10 and 20 dB. Recogniser performance was assessed using feature vectors for the favourable ear (i.e., the right ear, which was closest to the target speech), for the non-favourable (left) ear, and for the mean feature vectors from both ears. In all conditions, the T_{60} reverberation time was fixed at 0.3 s.

Table 2 shows recognition accuracy for each spatial separation, and for each of nine recogniser configurations (left, right and mean feature vectors for the MFCC baseline system, binaural mask estimation and a priori masks). Increasing the spatial separation between target speech and interfering speech improves the recognition performance with the binaural system, most notably in the 0 and 10

Table 1
Speech recognition accuracy (%) in the presence of interfering male speech for three rooms with T_{60} reverberation times of 0, 0.3 and 0.45 s

T_{60} (s)	Method	0 dB	10 dB	20 dB	200 dB
0	MFCC	5.9	50.4	78.2	99.7
	BINAURAL	92.9	97.2	97.6	98.2
	A PRIORI	96.3	97.2	97.6	98.2
0.3	MFCC	14.3	47.6	76.9	95.0
	BINAURAL	54.9	83.4	91.4	93.1
	A PRIORI	91.6	94.6	94.5	93.7
0.45	MFCC	13.0	47.1	75.4	94.3
	BINAURAL	53.8	80.1	90.9	92.3
	A PRIORI	91.2	94.8	94.9	93.9

Results are shown for SNRs of 0, 10, 20 and 200 dB, and for the MFCC-based recogniser, binaural missing data system (BINAURAL) and missing data recogniser with a priori masks (A PRIORI).

Table 2
Speech recognition accuracy (%) for angular separations of 10°, 20° and 40° azimuth between target speech and interfering speech sources, for SNRs of 0, 10 and 20 dB

Separation (°)	Method	0 dB	10 dB	20 dB
10	MFCC LEFT	18.2	49.7	78.2
	MFCC RIGHT	16.5	51.9	76.9
	MFCC MEAN	17.1	48.5	76.6
	BINAURAL LEFT	21.3	64.0	83.8
	BINAURAL RIGHT	18.9	62.1	81.5
	BINAURAL MEAN	21.4	63.4	84.0
	A PRIORI LEFT	90.2	94.5	95.3
	A PRIORI RIGHT	90.3	93.9	94.7
	A PRIORI MEAN	90.7	93.9	95.0
20	MFCC LEFT	11.1	46.0	74.5
	MFCC RIGHT	17.0	52.5	76.9
	MFCC MEAN	14.3	48.2	76.2
	BINARAL LEFT	36.0	70.7	86.4
	BINAURAL RIGHT	39.4	73.2	87.6
	BINAURAL MEAN	38.7	72.7	88.5
	A PRIORI LEFT	88.3	93.0	95.0
	A PRIORI RIGHT	91.7	94.1	95.1
	A PRIORI MEAN	90.2	93.5	95.6
40	MFCC LEFT	8.6	40.6	71.8
	MFCC RIGHT	23.1	56.2	80.9
	MFCC MEAN	14.3	47.6	76.9
	BINAURAL LEFT	51.4	81.3	89.6
	BINAURAL RIGHT	54.1	82.6	90.8
	BINAURAL MEAN	54.9	83.4	91.4
	A PRIORI LEFT	91.4	94.3	95.0
	A PRIORI RIGHT	91.4	94.2	94.6
	A PRIORI MEAN	91.6	94.6	94.5

For each system, results are shown for conditions in which recognition is based on the left ear features (LEFT), right ear features (RIGHT) or average of left and right features (MEAN).

dB SNR conditions. Performance with a priori masks is largely independent of the spatial separation between sources, since it is primarily determined by the SNR.

The performance of the binaural missing data system exceeds that of the MFCC system in all conditions. However, performance of the MFCC system approaches that of the binaural system with 10° angular separation at 0 dB SNR. In this challenging case, the binaural system failed to produce an accurate estimate of the azimuth of the target speech, often confusing it with the azimuth of the interfering noise (see Section 3.5).

We assessed the performance of each system on left ear, right ear and mean left–right feature vectors in order to investigate the benefit gained by using only the ear nearest to the target speech. Recognition accuracy for the binaural missing data system improves with angular separations of 20° or more when the excitation pattern for the ear nearest to the location of the target speech is used, rather than the opposite ear. The improvement is largest at the widest angular separation (40°), since the SNRs in the favourable and non-favourable ears differ most in this condition. The performance of the binaural system with averaged left–right ratemaps is either slightly below or above the performance in the right ear. A similar pattern is also seen for the other systems. No advantage was apparent for any system when the

angular separation between sources was small (10°), since the SNR in each ear will be similar.

3.5. Experiment 3: noise type

In the third experiment, the effect of intrusion type was investigated using three different noise intrusions from Cooke's (1993) corpus; male speech, female speech and rock music. For all conditions, the T_{60} reverberation time was 0.3 s, and the angular separation between speech and noise sources was 40° (corresponding to a noise azimuth of −20° and a target speech azimuth of 20°). Results were obtained for SNRs of 0, 10 and 20 dB.

Table 3 indicates that the performance of the binaural missing data system was best with the female speech intrusion and worst with rock music, across all SNRs. The MFCC baseline system shows a notably different pattern of results, performing worst with the male speech intrusion at all SNRs. These results suggest that the binaural system has an advantage when the spectrum of the noise intrusion significantly overlaps the spectrum of the target speech, as is the case when the intrusion is another male speaker.

3.6. Experiment 4: localisation performance and precedence effect model evaluation

In addition to speech recognition accuracy, we evaluated the accuracy of azimuth estimation by the binaural model. Furthermore, we investigated the utility of the precedence effect model by comparing results obtained with and without precedence processing. The bar charts in Fig. 5 show the percentage of hits at each azimuthal angle, quantised to 1°, for azimuth estimation during experiment 2. Each plot represents a total of 240 estimates of the azimuth of the target speech, one for each mixture in the test set. The mode (i.e., the azimuth angle which receives the greatest number of hits) and standard deviation are shown for each distribution.

For SNRs of 10 dB and above, location estimates from the model are distributed within 5° of the correct azimuth. However, decreasing the SNR causes the distribution to become broader, most notably when the SNR drops from 10 to 0 dB.

The effect of decreasing the angular separation between speech and noise sources results in more scattered estimates of the azimuth, and the model starts to confuse the location of the speech source with that of the noise. This is especially evident at 0 dB SNR. For 10° spatial separation and 0 dB SNR, the peaks in the pooled cross-correlogram are so close to each other that they tend to fuse, producing only a single location estimate which in many cases corresponds to the direction of the interfering utterance.

The left and right panels of Fig. 5 show comparisons between azimuth estimation with and without precedence effect processing. Precedence and

Table 3
Speech recognition accuracy (%) for male speech, rock music and female speech intrusions (see Table 1 for performance with male speech intrusion)

Noise type	Method	0 dB	10 dB	20 dB
Male speech	MFCC	14.3	47.6	76.9
	BINAURAL	54.9	83.4	91.4
	A PRIORI	91.6	94.6	94.5
Rock music	MFCC	9.5	50.3	86.7
	BINAURAL	32.7	78.8	91.9
	A PRIORI	88.5	91.6	93.2
Female speech	MFCC	16.5	47.7	80.2
	BINAURAL	53.9	84.3	92.8
	A PRIORI	90.5	93.5	93.7

The target speech and noise intrusion were mixed at SNRs of 0, 10 and 20 dB.

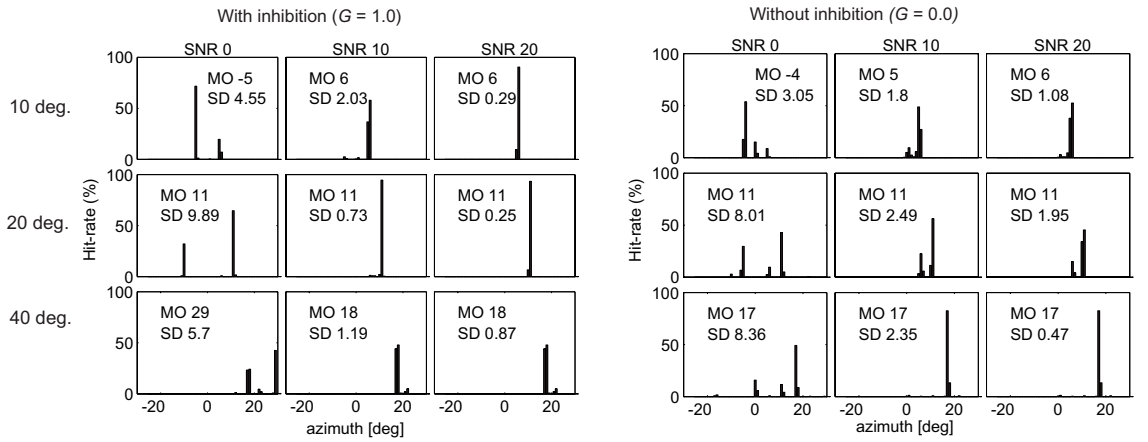

Fig. 5. Comparison of the distribution of target speech azimuth estimates, both with (left) and without (right) precedence effect processing. Each plot represents 240 data points, one for each utterance in experiment 2. Three angular separations of the speech and noise sources are shown (10°, 20° and 40°). The legend in each plot shows the mode (MO) and standard deviation (SD) of the distribution.

non-precedence test conditions were achieved by setting the inhibition gain G in (5) to 1.0 and 0.0 respectively. It is evident from Fig. 5 that the precedence model is particularly helpful in the 0 and 10 dB SNR conditions, in which it helps to produce azimuth estimates that are generally closer to the correct value.

4. General discussion

In this study we have described a binaural auditory preprocessor for missing data speech recognition. In a series of experiments, we evaluated the model with mixtures of speech and various types of spatially separated noise, at different SNRs and in different simulated acoustic environments.

Taken together, the results of our experiments suggest four main conclusions. Firstly, the binaural missing data system is much more robust than a conventional MFCC-based recogniser in the presence of an interfering sound source in small room acoustic conditions. Secondly, the performance of the binaural system depends on the angular separation between the speech and noise sources, giving substantial improvements in recognition accuracy at the largest separation (40°) but little improvement at the smallest separation (10°). Thirdly, the characteristics of the intrusive noise influence the performance of the binaural system; location cues appear to be particularly helpful when the spectra of the speech and noise substantially overlap. Finally, the performance of the binaural system is close to that obtained with a priori masks in anechoic conditions; performance in reverberation is lower, however, indicating that there is still room for improvement in the binaural mask estimation process.

On a related point, we also note that the MFCC-based baseline system performed slightly better than the binaural missing data recogniser when no intrusion was present. This suggests that the robustness of the binaural system to room reverberation requires further improvement. It is possible that an approach proposed by Palomäki et al. (2002) could be combined with the current system. They employed a reverberation masking algorithm to identify time–frequency regions that were not contaminated by echoes; these regions were retained in a missing data mask, whereas other spectral regions were discarded. Substantial improvements in the performance of their missing data recogniser over that of an MFCC based front-end were reported for reverberation times ranging from 0.7 to 2.7 s.

Computational models of the precedence effect have been proposed by a number of workers (e.g., Lindemann, 1986; MacPherson, 1991; Martin, 1997). Although some of these models are in good agreement with certain psychophysical findings such as the echo threshold, they have typically been tested in a rather limited context, and in some cases only with a single echo (Lindemann, 1986; Martin, 1997). In this study our intention was not to develop a comprehensive computational model of binaural precedence and its many complex subphenomena; rather, we required a simplified approach that embodied the main principles thought to underlie the precedence effect, and which improved localisation accuracy in a range of reverberant conditions. This goal was achieved, although much more work is needed on this aspect of the localisation model before its performance will approach that of human listeners.

In this study we needed to address the problem of convolutional distortion caused by room reverberation and HRIR filtering. Missing data recognition is primarily a technique for handling additive noise; the performance of previously described missing data recognisers (e.g., Barker et al., 2000a,b; Cooke et al., 2001) may be severely degraded by spectral variation in the transmission channel or even by a change of input gain. Conventional spectral normalisation strategies (such as that used with modulation filtered spectrograms; see Kingsbury, 1998) will not work within the missing data framework, since unreliable time–frequency regions will affect the way in which reliable regions are normalised. Our solution is to tightly integrate spectral normalisation and mask estimation, such that normalisation is based only on reliable (and relatively intense) speech regions as indicated by the mask estimation process. In a recent paper, we have combined this normalisation method with single channel missing data systems, and shown that the resulting recogniser is robust against convolutional noise and spectral distortion (Palomäki et al., 2004).

Azimuth estimation in this study was based only on ITD, which does not provide sufficient information to discriminate elevation or make front/back decisions. However, it is known that discrimination of elevation is poor when spatially separated noise is present, and that the intelligibility of speech in the presence of noise does not improve if the two sources are separated only by elevation (Hawley et al., 1999). In contrast, speech intelligibility improves substantially when the noise is located at a different azimuthal angle (Hawley et al., 1999; Spieth et al., 1954). Hence, in this study we concentrated our efforts on azimuth estimation, and avoided 'cone of confusion' errors (Blauert, 1997) that occur with ITD by restricting source locations to the frontal plane only.

We also note that ITD plays an important role in auditory scene analysis, in which it is used as a cue for linking acoustic events from the same sound source over time (Darwin and Hukin, 1999). It should be noted, however, that our model may be at odds with some psychophysical findings regarding the role of ITD in concurrent sound separation, since we employ ITD for both simultaneous (across-frequency) and sequential (across-time) grouping. For example, Hukin and Darwin (1995) have shown that listeners only exhibit a weak tendency to segregate a harmonic from a vowel, when that harmonic is given a different ITD to the remaining components of the vowel (see also Culling and Summerfield, 1995). Hence, it appears that across-frequency grouping is primarily mediated by other cues, such as harmonicity. Future work will address this issue by integrating harmonicity and common onset cues into our system; this might give rise to further performance gains, particularly since harmonicity is known to be a relatively robust cue for auditory grouping in the presence of reverberation (Darwin and Hukin, 2000; see also Shamsoddini and Denbigh, 2001).

Currently, the binaural system locates the speech and noise sources automatically; however, to ensure that the appropriate source is passed to the recogniser, we inform the system that the speech always lies to the right of the noise. This constraint could be relaxed by integrating our model with the multi-source decoder described by Barker et al. (2000a). Their system applies top–down constraints from speech models in order to identify acoustic fragments which have a high likelihood of resembling the training set. Hence it could be determined, prior to speech recognition per se, whether the acoustic features selected by a

binaural missing data mask correspond to a speech or non-speech sound source. We also currently constrain the sound sources to be stationary in space, so that an average source location can be computed over the whole input signal. In future work, we intend to adapt the model to deal with moving sound sources (see Roman and Wang, 2003). Again, top–down constraints from a multi-source decoder could help to track sources that move rapidly in space.

Finally, the experiments described here simulated an indoor acoustic environment in which speech and a noise intrusion were presented from spatially separated sources. This configuration is quite challenging for a computational model, and is not atypical of actual listening conditions. Clearly, however, speech perception by human listeners remains robust in environments containing many interfering noises that are distributed in auditory space. Future work on the binaural missing data system will focus on improving performance in multi-source environments of this kind.

Acknowledgments

KJP was mainly funded by the EC TMR SPHEAR project and partially supported by the Academy of Finland (project number 1277811) and a Finnish Tekniikan edistämissäätiö grant. GJB was supported by EPSRC grant GR/R47400/01. Parts of this work were undertaken while GJB was a visiting scientist at the Center for Cognitive Science, The Ohio State University. DLW was partially supported by an AFOSR grant and an NSF grant.

References

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Amer. 65 (4), 943–950.

Barker, J., Cooke, M.P., Ellis, D.P.W., 2000a. Decoding speech in the presence of other sound sources. Proc. ICSLP 4, 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 1, 373–376.

Blauert, J., 1997. Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition). MIT Press, Cambridge, MA.

Bodden, M., 1993. Modelling human sound-source localization and the cocktail-party-effect. Acta Acoust. 1, 43–55.

Bregman, A.S., 1990. Auditory Scene Analysis. MIT Press, Cambridge, MA.

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

Brown, G.J., Wang, D.L., Barker, J., 2001. A neural oscillator sound separator for missing data speech recognition. Proc. IJCNN 4, 2907–2912.

Cohen, L., 1994. Time–Frequency Analysis: Theory and Applications. Prentice Hall, Englewood Cliffs, NJ.

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 (3), 267–285.

Culling, J.F., Summerfield, Q., 1995. Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. J. Acoust. Soc. Amer. 98 (2), 785–797.

Darwin, C.J., Carlyon, R.P., 1995. Auditory grouping. In: The Handbook of Perception and Cognition. In: Moore, B.C.J. (Ed.), Hearing, Vol. 6. Academic, London, pp. 387–424.

Darwin, C.J., Hukin, R.W., 1999. Auditory objects of attention: the role of interaural time differences. J. Exp. Psychol. Hum. Percept. Perform. 25 (3), 617–629.

Darwin, C.J., Hukin, R.W., 2000. Effects of reverberation on spatial, prosodic and vocal-tract size cues to selective attention. J. Acoust. Soc. Amer. 108 (1), 335–342.

Denbigh, P.N., Zhao, J., 1992. Pitch extraction and separation of overlapping speech. Speech Comm. 11, 119–125.

Gardner, B., Martin, K.D., 1994. HRTF measurements of a KEMAR dummy-head microphone. Technical Report #280, MIT Media Lab. Available from: http://web.media.mit.edu/~kdm/hrtf.html.

Glasberg, B.R., Moore, B.C.J., 1990. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 47 (1–2), 103–138.

Glotin, H., Berthommier, F., Tessier, E., 1999. A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition. In: Proc. EUROSPEECH, 1999, pp. 2351–2354.

Hall, D.E., 1991. Musical Acoustics, second ed. Pacific Brooks Cole Publishing, Grove, CA.

Hawley, M.L., Litovsky, R.Y., Colburn, H.S., 1999. Speech intelligibility and localization in a multi-source environment. J. Acoust. Soc. Amer. 105 (6), 3436–3448.

Hermansky, H., 1998. Should recognizers have ears?. Speech Comm. 25 (1–3), 3–27.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589.

Huopaniemi, J., Savioja, L., Karjalainen, M., 1997. Modeling of reflections and air absorption in acoustical spaces: a

digital filter design approach. In: Proc. IEEE Workshop on Applications of Signal Processing to Acoustics and Audio, New Paltz, NY.

Hukin, R.W., Darwin, C.J., 1995. Effects of contralateral presentation and of interaural time differences in segregating a harmonic from a vowel. J. Acoust. Soc. Amer. 98 (3), 1380–1387.

Kingsbury, B.E.D., 1998. Perceptually inspired signal-processing strategies for robust speech recognition in reverberant environments. Ph.D. 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.

Koutras, A., Dermatas, E., Kokkinakis, G., 2001. Improving simultaneous speech recognition in real room environments using overdetermined blind source separation. Proc. EUROSPEECH 2, 1009–1012.

Leonard, R.G., 1984. A database for speaker-independent digit recognition. Proc. ICASSP 3, 111–114.

Lindemann, W., 1986. Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front. J. Acoust. Soc. Amer. 80 (6), 1623–1630.

Litovsky, R.Y., Colburn, S.H., Yost, W.A., Guzman, S.J., 1999. The precedence effect. J. Acoust. Soc. Amer. 106 (4), 1633–1654.

Lokki, T., 2002. Physically-based auralization. Ph.D. Thesis. Publications in Telecommunications Software and Multimedia, Helsinki University of Technology.

MacPherson, E.A., 1991. A computer model of binaural localization for stereo imaging measurement. J. Audio Eng. Soc. 39 (9), 604–622.

Martin, K.D., 1997. Echo suppression in a computational model of the precedence effect. In: Proc. IEEE Workshop on Applications of Signal Processing to Acoustics and Audio, New Paltz, NY.

Møller, H., 1992. Fundamentals of binaural technology. Appl. Acoust. 36, 171–218.

Moore, B.C.J., 1997. An Introduction to the Psychology of Hearing, fourth ed. Academic Press, New York.

Nabelek, A.K., Robinson, P.K., 1982. Monaural and binaural speech perception in reverberation for listeners of various ages. J. Acoust. Soc. Amer. 71 (5), 1242–1248.

Okuno, H.G., Nakatani, T., Kawabata, T., 1999. Listening to two simultaneous speeches. Speech Comm. 27, 299–310.

Palomäki, K.J., Brown, G.J., Barker, J., 2002. Missing data speech recognition in reverberant conditions. In: Proc. ICASSP, Orlando, 13th–17th May. pp. 65–68.

Palomäki, K.J., Brown, G.J., Barker, J., 2004. Techniques for handling convolutional distortion with 'missing data' automatic speech recognition. Speech Comm. 43, 123–142.

Palomäki, K.J., Brown, G.J., Wang, D.L., 2001. A binaural model for missing data speech recognition in noisy and reverberant conditions. In: Proc. Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis (CRAC), Aalborg.

Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., Rice, P., 1988. APU report 2341: an efficient auditory filterbank based on the gammatone function. Applied Psychology Unit, Cambridge.

Roman, N., Wang, D.L., 2003. Binaural tracking of multiple moving sources. In: Proc. ICASSP, pp. 149–152.

Roman, N., Wang, D.L., Brown, G.J., 2002. Location-based sound segregation. In: Proc. ICASSP, Orlando, 13th–17th May. pp. 1013–1016.

Rosenthal, D.F., Okuno, H.G., 1998. Computational Auditory Scene Analysis. Lawrence Erlbaum Associates, Mahwah, NJ.

Seltzer, M.L., Raj, B., 2001. Calibration of microphone arrays for improved speech recognition. Proc. EUROSPEECH 2, 1005–1008.

Shackleton, T.M., Meddis, R., Hewitt, M.J., 1992. Across frequency integration in a model of lateralization. J. Acoust. Soc. Amer. 91 (4), 2276–2279.

Shamsoddini, A., Denbigh, P.N., 2001. A sound segregation algorithm for reverberant conditions. Speech Comm. 33, 179–196.

Spieth, W., Curtis, J.F., Webster, J.C., 1954. Responding to one of two simultaneous messages. J. Acoust. Soc. Amer. 26 (3), 391–396.

van der Kouwe, A.J.W., Wang, D.L., Brown, G.J., 2001. A comparison of auditory and blind separation techniques for speech segregation. IEEE Trans. Speech Audio Process. 9, 189–195.

Wallach, H., Neumann, E.B., Rosenzweig, M.R., 1949. The precedence effect in sound localization. Amer. J. Psychol. 52, 315–336.

Watkins, A.J., 1991. Central, auditory mechanisms of perceptual compensation for spectral- envelope distortion. J. Acoust. Soc. Amer. 90 (6), 2942–2955.

Yost, W.A., 1997. The cocktail party problem: Forty years later. In: Gilkey, R.H., Anderson, T.R. (Eds.), Binaural Hearing in Real and Virtual Environments. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 329–348.

Zurek, P.M., 1987. The precedence effect. In: Yost, W.A., Gourevitch, G. (Eds.), Directional Hearing. Springer-Verlag, New York, pp. 85–105.