Mask Estimation for Missing Data Speech Recognition Based on Statistics of Binaural Interaction

Sue Harding, *Member, IEEE*, Jon Barker, and Guy J. Brown

***Abstract*—This paper describes a perceptually motivated computational auditory scene analysis (CASA) system that combines sound separation according to spatial location with the "missing data" approach for robust speech recognition in noise. Missing data time–frequency masks are created using probability distributions based on estimates of interaural time and level differences (ITD and ILD) for mixed utterances in reverberated conditions; these masks indicate which regions of the spectrum constitute reliable evidence of the target speech signal. A number of experiments compare the relative efficacy of the binaural cues when used individually and in combination. We also investigate the ability of the system to generalize to acoustic conditions not encountered during training. Performance on a continuous digit recognition task using this method is found to be good, even in a particularly challenging environment with three concurrent male talkers.**

***Index Terms*—Automatic speech recognition, binaural, computational auditory scene analysis (CASA), interaural level differences (ILD), interaural time differences (ITD), missing data, reverberation.**

I. Introduction

DESPITE much progress in recent years, robust automatic speech recognition (ASR) in noisy and reverberant environments remains a challenging problem. This is most apparent when one considers the relative performance of ASR systems and human listeners on the same speech recognition task. Word error rates for ASR systems can be an order of magnitude greater than those for human listeners, and the differences are largest when the speech is contaminated by background noise or room reverberation [1]. What aspects of the auditory system give rise to this advantage, and can the underlying mechanisms be incorporated into our computational systems in order to improve their robustness?

One obvious characteristic of human listeners is that they have two ears, whereas ASR systems typically take their input from a single audio channel. Binaural hearing underlies a number of important auditory functions (see [2] for a review). First, human listeners are able to localize sounds in space principally by measuring differences between the time of arrival and sound level at the two ears. These cues are referred to as interaural time differences (ITDs) and interaural level differences (ILDs). Second, binaural mechanisms suppress echoes and, therefore, counteract the effects of reverberation [3]. Finally, binaural hearing contributes to the ability of listeners to attend to a target sound source in the presence of other interfering sources. Evidence for this has come from psychophysical studies, which show that the intelligiblity of two overlapping speech signals increases as the spatial separation between them increases [4]. A number of processes appear to be involved in this respect. For example, listeners may simply attend to the ear in which the signal-to-noise ratio (SNR) is most favorable. In addition, more complex mechanisms may be involved, in which binaural comparisons are used to cancel interfering sound sources [5] or actively group acoustic energy which originates from the same location in space [6].

The notion of auditory grouping is a key component of Bregman's theory of auditory scene analysis (ASA), an influential account of the processes by which listeners segregate a target sound source from an acoustic mixture [7]. Bregman's work has stimulated interest in the development of computational auditory scene analysis (CASA) systems, which attempt to mimic the sound separation abilities of human listeners. A number of these systems have exploited binaural cues in order to separate a desired talker from spatially separated interferers [8]–[12].

In this paper, we describe a CASA system which exploits spatial location cues in order to improve the robustness of ASR in multisource, reverberant environments. Our approach is implemented within the "missing data" framework for ASR [13], and consists of two processing stages. In the first stage, acoustic features (spectral energies) and binaural cues (ITD and ILD) are derived from an auditory model. The binaural cues are used to estimate a time–frequency mask, in which each element indicates whether the corresponding acoustic feature constitutes reliable evidence of the target speech signal or not. In the second stage, the acoustic features and the time–frequency mask are passed to a missing data ASR system, which treats reliable and unreliable features differently during decoding.

The approach described here extends our previous work in a number of respects. First, our previous systems [12], [14], used heuristic rules to estimate time–frequency masks. Here, we adopt a more principled approach in which masks are estimated from probability density functions for binaural cues,

Manuscript received January 31, 2005; revised July 22, 2005. This work was supported by EPSRC under Grant GR/R47400/01. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Bhiksha Raj.

The authors are with the Department of Computer Science, University of Sheffield, Sheffield S14DP, U.K. (e-mail: s.harding@dcs.shef.ac.uk; j.barker@dcs.shef.ac.uk; g.brown@dcs.shef.ac.uk).

Digital Object Identifier 10.1109/TSA.2005.860354

which are obtained from a corpus of sound mixtures. In this respect, our current work is representative of the current trend in CASA for applying statistical rather than heuristic methods (for example, see [11], [15], [16]). Second, we consider the problem of a multisource environment with realistic room reverberation. This represents a much greater challenge than the anechoic or very mildly reverberant conditions used in some of our previous studies [11], [12]. Finally, whereas our previous approach used ITD only, here we investigate the relative effectiveness of ITD, ILD, and a joint ITD–ILD space. The latter approach is related to the binaural speech segregation system of Roman *et al.* [11]. Their system determines the relative strength of a target and interferer (and, hence, estimates a binary mask) by measuring the deviation of observed ITD and ILD in each frequency band of a binaural auditory model. Specifically, a supervised learning method is used for different spatial configurations and frequency bands within an ITD–ILD feature space. Given an observation x in the ITD–ILD feature space for a frequency band, two hypotheses are tested; whether the target is dominant (H_1) and whether the interferer is dominant (H_2). Using estimates of the bivariate densities $p(x|H_1)$ and $p(x|H_2)$, classification is then performed using a maximum *a posteriori* probability (MAP) decision rule.

Although our approach is similar to that of Roman *et al.*, there are important differences. Here, we estimate probability distributions for ITD, ILD, and ITD–ILD directly from training data, rather than using a parametric method. We also assume that the target is located at a known azimuth, which simplifies subsequent processing. Most importantly, we have evaluated our system in reverberant conditions. Reverberation remains a substantial problem for both ASR and CASA systems. In the case of ASR, it is well known that recognition accuracy falls as the T_{60} reverberation time increases and the ratio of direct sound to reverberation decreases [17]. Similarly, the performance of CASA systems is degraded by reverberation (e.g., [12]), to the extent that many CASA systems are only evaluated on anechoic mixtures. Here, we evaluate our combined CASA and ASR system using the same speech recognition task as Roman *et al.*, and obtain accuracies in reverberant conditions which are similar to those obtained by Roman *et al.* for anechoic mixtures.

Our approach also differs from speech recognition systems that use multiple microphones (for example, see [18]). Such systems typically perform spatial filtering using adaptive beamforming, in order to preserve the signal emanating from a target direction while suppressing noise and reverberation that originate from other directions. Here, we do not use spatial information to derive filtered acoustic features; rather, spatial cues are used to select acoustic features that are representative of the target speech signal. Similarly, we use spectral acoustic features rather than those that are intended to confer robustness against noise and reverberation, such as relative spectral perceptual linear prediction (RASTA-PLP) [19] or mean-normalized mel-frequency cepstral coefficients (MFCC) [20]. Our previous work suggests that such "noise robust" features do not perform well in conditions where both interfering sounds and reverberation are present [12]. We also note that our approach is a purely data-driven one, as opposed to model-based approaches such as parallel model combination [21] and hidden Markov model decomposition [22].

TABLE I
REVERBERATION TIMES (SECONDS) FOR TWO SURFACES

Surface	Frequency (Hz)						
	125	250	500	1000	2000	4000	Mean
Acoustic plaster	1.02	0.49	0.17	0.14	0.11	0.11	0.34
Platform floor wooden	0.22	0.31	0.49	0.48	0.65	0.89	0.51

The remainder of the paper is organized as follows. Section II explains the methods used (signal spatialization and reverberation, missing data ASR, and mask estimation using binaural cues). Section III describes a number of experiments, which examine the effects of cue selection and other training parameters on ASR performance. Section IV concludes the paper with a general discussion.

II. GENERAL METHODS

A. Signal Spatialization and Reverberation

Input data for the CASA system consisted of utterances from the TI digits corpus [23], to which reverberation and spatialization were applied. It was assumed that the source of interest was at azimuth 0°, although another azimuth could have been used. Signals consisting of two concurrent utterances by male speakers were used to test the system; similar acoustic mixtures were used as training data when generating the probability distributions used to create the missing data masks. Each utterance consisted of a speaker saying from one to seven digits, with each digit selected from the list "one two three four five six seven eight nine zero oh." The mixed utterances consisted of a target utterance at azimuth 0 mixed with another masking utterance at a different azimuth.

In order to add reverberation and spatial location to the original monaural utterances, impulse responses were created using the Roomsim simulator[1] with a simulated room of size $6 \times 4 \times 3$ m. The receiver was a KEMAR head[2] in the center of the room, 2 m above the ground, and the source was at azimuth 0°, 5°, 7.5°, 10°, 15°, 20°, 30°, or 40° at a radial distance of 1.5 m from the receiver. All surfaces of the room were assumed to have identical reverberation characteristics. Two reverberation surfaces were used, "acoustic plaster" and "platform floor wooden, " with mean estimated T_{60} reverberation times of 0.34 and 0.51 s, respectively (reverberation times at standard frequencies are shown in Table I): note that the latter surface was used only for generating training data, not test data. The room impulse responses were convolved with monaural signals to produce binaural reverberated data.

B. Auditory Model

Each signal was processed using a 64-channel gammatone filterbank with center frequencies spaced between 50 Hz and

[1]http://media.paisley.ac.uk/~campbell/Roomsim/

[2]http://sound.media.mit.edu/KEMAR.html

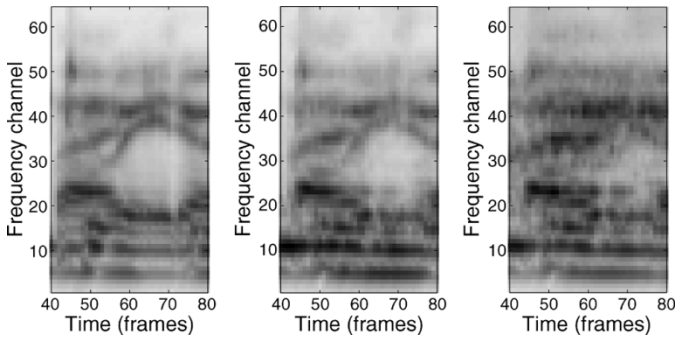

Fig. 1. Segments of auditory spectrograms for the utterance "one two eight oh" at azimuth 0 mixed at SNR 0 dB with utterance "five three" at azimuth 40, both by male speakers. Left, anechoic; middle, reverberated, surface "acoustic plaster." Right, reverberated, surface "platform floor wooden."

8 kHz on the equivalent rectagular bandwidth (ERB) scale [24]. An analysis window of 20 ms and frame shift of 10 ms was applied to data sampled at 20 kHz to create an auditory spectrogram for each signal (Fig. 1). Auditory spectrogram frames concatenated with interframe differences (delta features) provided the 128-dimensional acoustic feature vectors for the recognizer.

In order to find the ITD and ILD for the mixed utterances, a cross-correlogram was created by passing each of the binaural inputs through the auditory filterbank described above and computing the cross-correlation between each frequency channel for each time frame. The biggest peak in each channel was selected and the estimate of ITD was improved by fitting a quadratic curve to the peak. The ILD was calculated by summing the energy over each frequency channel and converting the ratio of the energy for the right and left ears to decibels. Further details can be found in [12].

C. The "Missing Data" Speech Recognizer

The missing data recognizer uses hidden Markov models (HMMs) trained on spectrotemporal acoustic features. During testing, the recognizer is supplied with features for an acoustic mixture and a time–frequency mask, that indicates which parts of the input constitute reliable evidence for the target source. The missing data mask may be discrete, i.e., each time–frequency element is either 0 (meaning the target is masked), or 1 (meaning the target is dominant); alternatively, "soft" masks may be used [25] in which each element takes a real value between 0 and 1 indicating the probability that the element is dominated by the target.

The HMM emission distributions are based on Gaussian mixture models (GMMs), with each component having a diagonal covariance matrix. During recognition, the soft missing data technique adapts the usual state-likelihood computation to compute state scores, $p'(\mathbf{x}|q)$, which take account of the unreliable data as follows:

$$p'(\mathbf{x}|q) = \sum_{\lambda} w_\lambda p'(\mathbf{x}|q,\lambda) \tag{1}$$

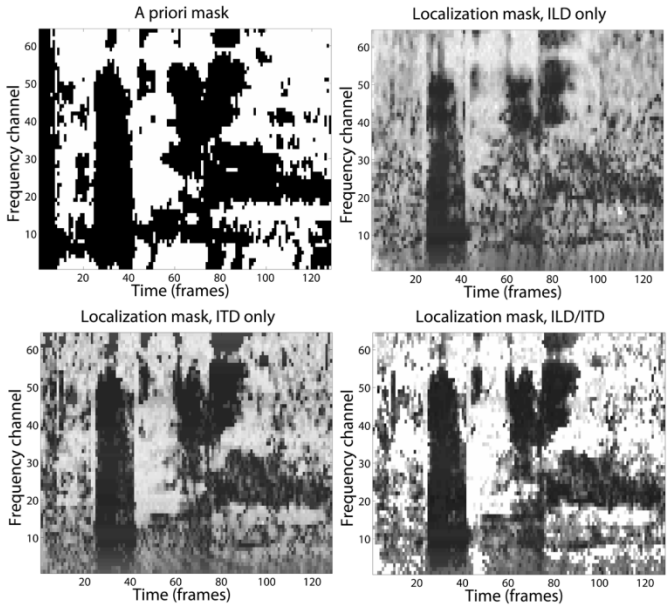

Fig. 2. Missing data masks for the mixed utterances in Fig. 1, reverberation surface "acoustic plaster." Top left: *a priori* mask. Top right: localization mask produced using ILD only. Bottom left: localization mask produced using ITD only. Bottom right, localization mask produced using ILD and ITD combined. Lighter areas have lower probability; darker areas higher probability.

where w_λ is the mixture weight for GMM component λ, and $p'(\mathbf{x}|q)$ is a score for each mixture component, given by forming a product over the spectral features of the form

$$p'(\mathbf{x}|q,\lambda) = \prod_i \left(\mathbf{m}_i p(\mathbf{x}_i|q,\lambda) + \frac{1-\mathbf{m}_i}{\mathbf{x}_i} \int_{-\infty}^{\mathbf{x}_i} p(\mathbf{x}'_i|q,\lambda) d\mathbf{x}'_i \right). \tag{2}$$

Here, $\mathbf{m}_i$ is the soft mask value for the ith feature and $p(\mathbf{x}_i|q,\lambda)$ are the univariate Gaussian distributions

$$p(\mathbf{x}_i|q,\lambda) = \frac{1}{\boldsymbol{\sigma}_{\lambda i}\sqrt{2\pi}} \exp^{-\frac{1}{2}\frac{(\mathbf{x}_i - \boldsymbol{\mu}_{\lambda i})2}{(\boldsymbol{\sigma}_{\lambda i})2}} \tag{3}$$

where σ_λ and μ_λ represent the standard deviation and mean, respectively, of mixture component λ.

Equation (2) can be interpreted as an interpolation between a present data term and a missing data term, where the missing data term is formed by averaging the likelihoods for all possible values that the masked speech could have had. The mask value $\mathbf{m}_i$ biases the interpolation toward either the present or missing data term. This implementation of soft missing data is identical to that used in [25]; though clearly, the current work uses a different technique for computing the soft mask.

It is possible to produce an "ideal" mask using *a priori* knowledge of the source and masker (Fig. 2, top left); such *a priori* masks can be used to provide an expected upper limit for recognition performance for the missing data approach. The main challenge for this approach is how to determine a mask from a mixed signal without prior knowledge; in this paper we use localization cues for this purpose.

A set of eight-state ten-mixture HMMs with delta coefficients were used for recognition, and were trained using clean reverberated speech (i.e., without any background noise or other speakers added). A set of 4228 clean speech utterances

by 55 male speakers, spatialized at azimuth 0° and with reverberation applied for surface "acoustic plaster," were processed as described above to create the acoustic features and used to train the recognizer.

D. Missing Data Mask Estimation Using ILD and ITD Probability Distributions

Soft missing data masks were determined from probability distributions which indicated the probability that a given ILD, ITD, or combination of ILD and ITD, found for a mixed test utterance, was produced by a target source at azimuth 0.

In order to create the probability distributions, ILD and ITD were identified for each time–frequency element of a set of mixed training utterances (described in Section II-A above). Probability distributions were produced for a range of training data, which was always selected from a set of 120 pairs of utterances, matched for length, for reverberation surface "acoustic plaster" or "platform floor wooden." One utterance was at azimuth 0° and the other at 5°, 10°, 20°, or 40°, or at −5°, −10°, −20°, or −40°, and the utterances were mixed at SNR 0, 10, and 20 dB. The choice of reverberation surface and SNR used in each training set depended on the experiment, but all of the eight azimuth separations listed above were always included.

After each ILD and ITD was determined for each time–frequency element of a training utterance, it was assigned to a bin. The bin sizes were selected according to the range of values observed for ILD and ITD in the training data and were set to 0.1 for ILD and 0.01 for ITD. These values were found during preliminary investigations to provide sufficient resolution for the probability distributions.

Two histograms were produced from the binned ILD and ITD values. The first histogram, H_a, counted the number of observations of each ILD/ITD combination in all of the training data (i.e., including observations produced by both the target source and the masking source); the second, H_t, counted the number of observations of each ILD/ITD combination produced by the target source alone. Observations likely to have been produced by the target source were identified using an *a priori* mask (Section II-C) created for each mixed training utterance: only those time–frequency elements of the *a priori* mask that were identified as belonging to the target source were included in histogram H_t.

The probability distribution was modeled using a Bayesian approach. Given an observation $o = (\mathrm{ILD}, \mathrm{ITD})$ for a time–frequency element of a mixed utterance, the probability that the observation was produced by a target source at azimuth zero is given by

$$p(\mathrm{Target}|o) = \frac{p(o|\mathrm{Target})p(\mathrm{Target})}{p(o)} \tag{4}$$

where

$$p(o|\mathrm{Target}) = \frac{H_t(o)}{\sum H_t} \tag{5}$$

$$p(o) = \frac{H_a(o)}{\sum H_a} \tag{6}$$

$$P(\mathrm{Target}) = \frac{\sum H_t}{\sum H_a}. \tag{7}$$

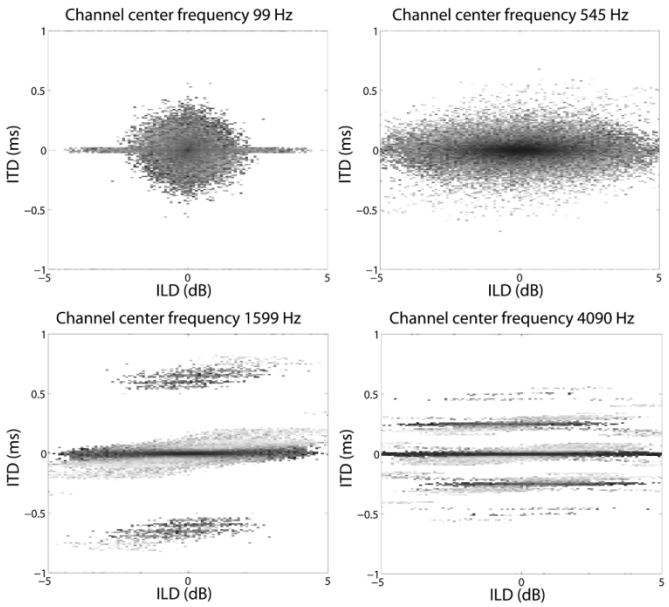

Fig. 3. Examples of ILD/ITD probability distributions for a source at azimuth 0°, for channels with center frequency 99 Hz (top left), 545 Hz (top right), 1599 Hz (bottom left), and 4090 Hz (bottom right). Lighter areas have lower probability; darker areas higher probability.

Then, (4) simplifies to

$$p(\mathrm{Target}|o) = \frac{H_t(o)}{H_a(o)} \tag{8}$$

which can be found from the histograms obtained from the training data as described previously.

Observations were counted separately for each frequency channel due to the large variation in the probability distributions for different channels.

A further step was performed to reduce the effect of insufficient training data for certain observations. Any time–frequency elements for which the number of observations $H_a(o)$ was less than a threshold value were treated as if no data were present. Using a threshold reduced the chance of an unrealistically high probability occurring when only a few elements were present but were allocated to the target; for example, a single target element would result in a probability of 1. When the denominator $H_a(o)$ was zero, the numerator $H_t(o)$ was also zero, and the corresponding probability was set to zero. This smoothed the progression from regions of low probability (where fewer observations occurred) to high probability (where larger numbers of observations occurred) (Fig. 3). The choice of threshold was determined heuristically, as described in Section III-B, in order to find a balance between including excessively high probabilities due to lack of training data, and excluding probabilities resulting from small amounts of training data that would nevertheless have been valid.

A missing data mask was created for each mixed test signal by identifying the ILD and ITD for each time–frequency element, and using the probability distribution as a look-up table for these two cues to determine the probability that each element was dominated by the target at azimuth 0.

Fig. 3 shows examples of probability distributions obtained in this way for four of the 64 frequency channels, for training data

TABLE II
SUMMARY OF EXPERIMENTAL CONDITIONS

Experiment	Probability distribution		Training data		Test data
	Binaural cues	Histogram threshold	Surface	SNR (dB)	No. of maskers
1	ILD, ITD or combined ILD/ITD	10	'acoustic plaster'	0, 10 and 20 (combined)	1
2	Combined ILD/ITD	0, 5, 10 or 50	'acoustic plaster'	0, 10 and 20 (combined)	1
3	Combined ILD/ITD	10	'acoustic plaster' or 'platform floor wooden'	0, 10 and 20 (combined)	1
4	Combined ILD/ITD	3	'acoustic plaster'	0, 10 or 20 (individually)	1
5	Combined ILD/ITD	10	'acoustic plaster'	0, 10 and 20 (combined)	2

with a combination of SNRs and histogram threshold 10. The examples are for a probability distribution based on both ILD and ITD cues, but distributions were also produced for each cue independently.

Examples of missing data masks created from probability distributions determined from ILD alone, ITD alone, and ILD and ITD combined are shown in Fig. 2.

E. Evaluation

The accuracy of the masks and the localization process was evaluated by measuring the recognition accuracy for a set of mixed test utterances for which missing data masks had been produced using each probability distribution. Only one reverberation surface, "acoustic plaster," was used for the test set. The test utterances consisted of 240 target utterances, different from those in the training set, with reverberation added, spatialized at azimuth 0° and mixed at an SNR of 0, 5, 10, 15, or 20 dB with one of a set of 240 masking male utterances, matched in length to the original 240 utterances. The masking speech was reverberated and spatialized at azimuth 5°, 7.5°, 10°, 15°, 20°, 30°, or 40°. The resulting mixture was processed to form an auditory spectrogram as for the training data. The SNR was calculated from data spatialized at azimuth 0°; the mixed speech signal entering the ear furthest from the masking speech was used for recognition.

A number of experiments were performed to investigate the importance of each of the two localization cues (ILD and ITD) together with the effect of different properties (such as reverberation surface and SNR) of the training data used to create the probability distribution. A further experiment investigated whether the method was successful in more difficult test conditions, using two masking utterances. These experiments are summarized in Table II.

III. EXPERIMENTS

A. Experiment 1—Effect of Cue Selection

Experiment 1 investigated the effects of selecting a single localization cue (ITD or ILD) and of combining both cues together. Probability distributions were created for each of these three conditions, using training data for reverberation surface "acoustic plaster" and for all three SNRs (0, 10, and 20 dB) combined together. When creating the probability distributions, the histogram H_a threshold value was set to 10 (see also Experiment 2).

Fig. 4 shows the recognition accuracy for the 240 mixed test utterances for each of the three conditions, together with baseline results for *a priori* masks. A recognizer was also trained and tested using MFCCs derived from the same training and testing data. For this approach, the data was processed using an analysis window of 25 ms and frame shift of 10 ms, with 23 frequency channels, 12 cepstral coefficients plus an energy term, delta and acceleration coefficients, and energy and cepstral mean normalization. The recognizer did not perform well under noisy conditions as can be seen from the MFCC results included in Fig. 4.

A separate plot is shown for each of the five SNRs used in the test data. The different effect of the cues was most pronounced for lower test SNR, but, overall, using both ILD and ITD produced better results than when only a single cue was used. Using ILD alone produced the worst performance in all cases, showing that this cue was not sufficient to define the probability distribution, especially for smaller azimuth separation of the two sources. When both cues were used, performance levelled out above azimuth separation of 10°. This experiment also confirmed that the approach generalized successfully to SNRs and azimuth separations which were not included in the training data.

B. Experiment 2—Effect of Histogram Threshold

During the production of the probability distribution, a threshold was applied to the histogram elements making up the denominator $H_a(o)$ to avoid attaching significance to ILD or ITD values for which there was insufficient training data. Any ILD/ITD combinations that had a denominator below the threshold were treated as if no data were present (see Section II-D). Increasing the threshold prevents combinations with few examples from having an excessively high probability, but setting the threshold too high may remove useful information from the probability distribution and, hence, from the localization masks. Experiment 2 tested the effect of varying the threshold on the performance for probability distributions using combined ILD/ITD cues and training data as in Experiment 1, to ensure that the most suitable threshold value was being used. Threshold values of 5, 10, and 50 were compared, together with no threshold (denoted threshold 0). For this experiment, only the test data with SNR 0 dB was used.

Fig. 5 shows the performance for each of the four threshold values; the results for threshold 10 (also shown in Fig. 4, top left) are emphasised. As the threshold increased, the performance improved for azimuth separation below 15°, but decreased for high azimuth separation, especially for the highest threshold (50). At low azimuth separation, ILD and ITD estimation is more difficult and, therefore, the probabilities are less accurate, especially when low threshold values allow more elements with little training data to be included in the probability distribution. At higher azimuth separation, ILD and ITD estimates are more

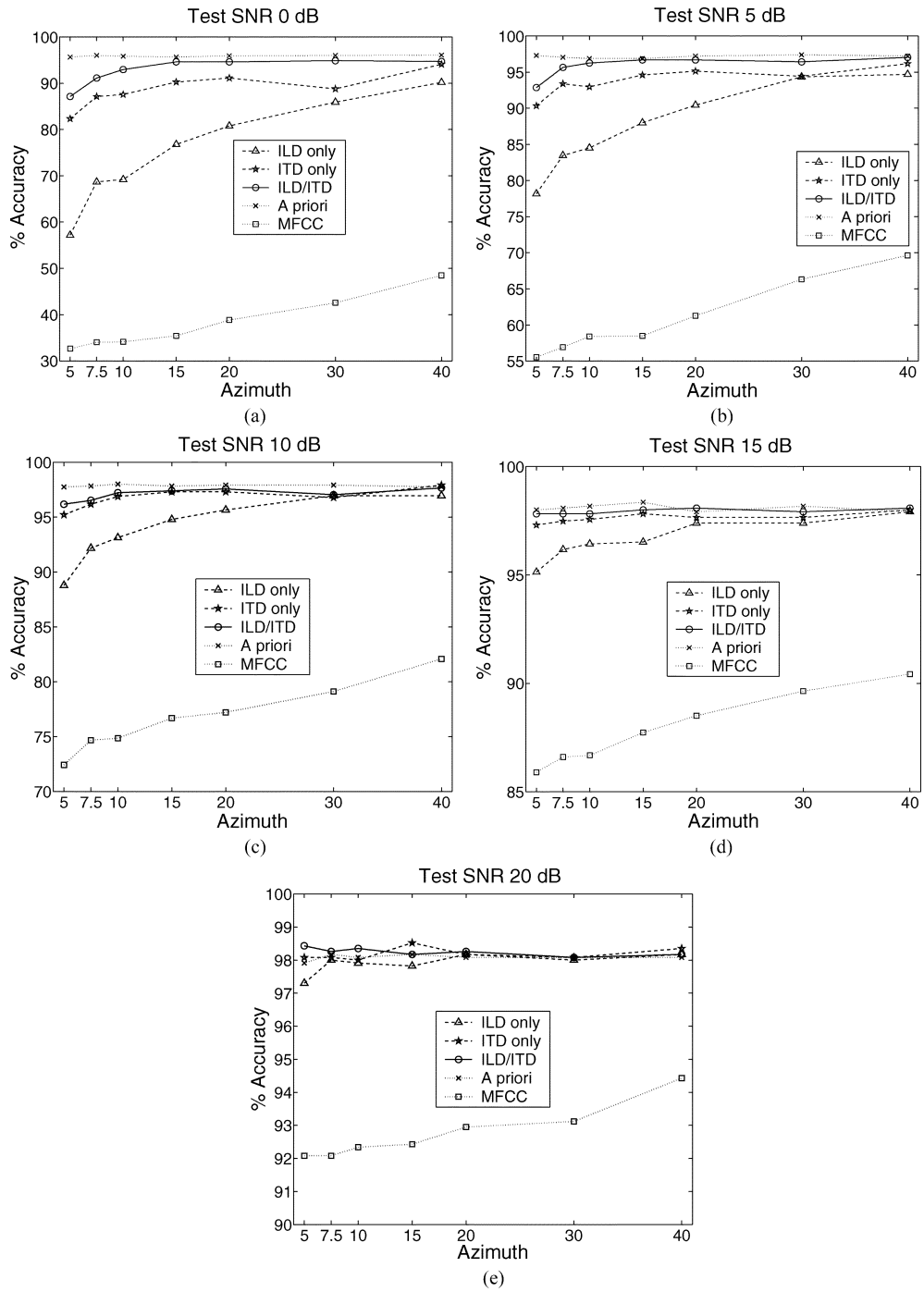

Fig. 4. Results of Experiment 1, showing the effect of varying localization cues (ILD and ITD). Training data: surface "acoustic plaster," SNR 0, 10, and 20 dB (combined together). Histogram threshold: 10. Test data: surface "acoustic plaster," SNR 0, 5, 10, 15, or 20 dB. One plot is shown per test data SNR. (Note that the ordinate scale varies.)

reliable, but higher threshold values will remove some of the reliable probabilities and cause poorer masks to be produced, as can be seen for threshold 50. These issues are discussed further in Section IV.

Although the shape of the curve differed for threshold 10 and 50 due to the factors mentioned above, the mean performance was similar (92.85% and 92.89%, respectively). However, performance decreased for threshold 5 (mean 92.23%) and deteriorated when no threshold was used (mean 85.23%), especially for azimuth separations of 15° and 30° for which there was no training data. In general, a lack of training data is likely to result in unreasonably high probability values and hence unreliable masks.

A threshold of 10 was considered to be the most suitable value, since the mean performance was high and there was no deterioration in performance with increasing azimuth.

C. Experiment 3—Effect of Training Data Surface

Experiment 1 showed that the method generalized well over azimuth separations and SNRs that were not in the training data

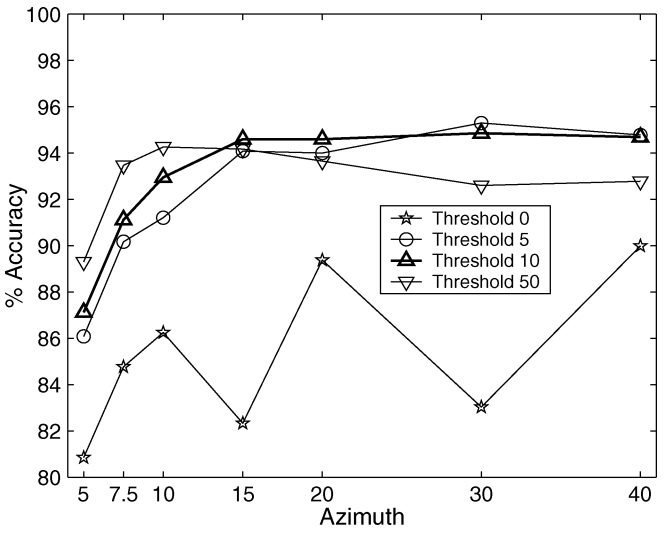

Fig. 5. Results of Experiment 2, showing the effect of varying histogram threshold for combined ILD/ITD cues. Training data: surface "acoustic plaster," SNR 0, 10, and 20 dB (combined together). Histogram threshold: 0, 5, 10, or 50. Test data: surface "acoustic plaster," SNR 0 dB.

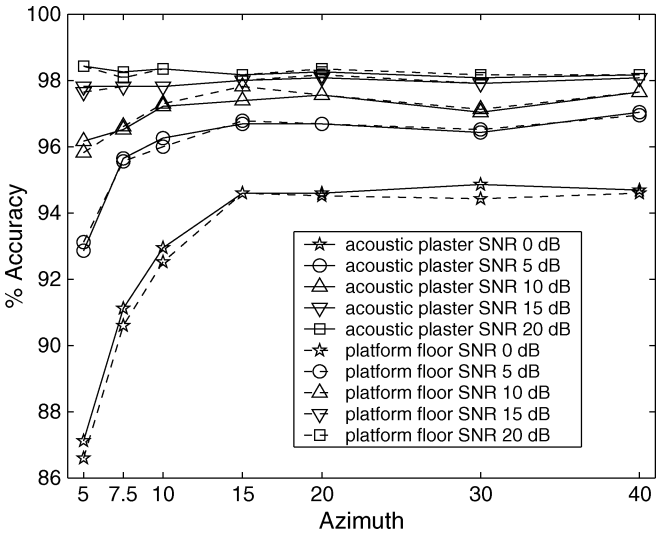

Fig. 6. Results of Experiment 3, showing the effect of varying the training data reverberation surface for combined ILD/ITD cues. Training data: surface "acoustic plaster" or "platform floor wooden," SNR 0, 10 and 20 dB (combined together for each surface). Histogram threshold: 10. Test data: surface "acoustic plaster," SNR 0, 5, 10, 15, or 20 dB.

set. Experiments 3 and 4 tested the importance of a match or mismatch between other parameters of the training and testing data. In Experiment 3, a training data set was used for a reverberation surface ("platform floor wooden") that did not match the test data; this was compared with the results for the combined ILD and ITD cues from Experiment 1.

Fig. 6 shows the performance for the two training sets, "acoustic plaster" (previously shown in Fig. 4) and "platform floor wooden." The reverberation surface used for the training data had little effect on the results, with a maximum difference between the surfaces of less than 1% for each condition (SNR and azimuth separation) in the test data.

D. Experiment 4—Effect of Training Data SNR

Experiment 4 examined whether using training data with an SNR that matched that of the test data might improve the performance. In the previous experiments, training data for each of three SNRs (0, 10, and 20 dB) was combined to produce the histograms used to create the probability distributions. For Experiment 4, a separate probability distribution was produced for each SNR and the performance compared for each training condition. The threshold was reduced from 10 to 3 when creating these three distributions to allow for the reduced quantity of training data compared with the distribution for the combined SNRs.

Fig. 7 shows the results plotted separately for each of the five test SNRs. For the lower test SNRs (0, 5 and 10 dB), there was an increase in performance of up to 3% for the smaller azimuth separations (5° and 7.5 °) when the training SNR was also low (0 dB). However, this was offset by a small decrease in performance of around 0.5% for the higher azimuth separations. The inverse occurred when a higher training SNR (10 or 20 dB) was used: performance decreased by up to 3% or 5%, respectively, for smaller azimuth separations, but there was a corresponding small increase in performance for higher azimuth separations. For the higher test SNRs (15 and 20 dB), the training data SNR had little effect. Overall, there was a small mean improvement when using a training SNR of 0 dB, suggesting that it is more important to use training data for the more difficult conditions.

E. Experiment 5—Using Multiple Masking Sources

The method used in these experiments can be applied to data that has more than one masking source. In Experiment 5, additional test data was produced for which two masking sources (both male speakers) were present, to check that good performance could be obtained in more difficult test conditions.

The first masker was at 5°, 7.5°, 10°, 15°, 20°, 30°, or 40° azimuth; the second was at $-10°$ or $+10°$. The two maskers were mixed at SNR 0 dB, and then this mixture was combined with the target at SNR 0 dB. Missing data masks were created using the ILD/ITD probability distribution for training data with surface "acoustic plaster" and a combination of SNR values (0, 10, and 20 dB). The signal entering the ear furthest from the first masker was used for recognition.

Fig. 8 shows the results of this experiment, together with the results for *a priori* masks for the one masker and two masker conditions. Performance was still good considering the difficulty of the task, although reduced by 3%–5% (compared with the single masker case) when both maskers were on the same side of the head, and by 4%–7% when maskers were on opposite sides of the head. Performance using the *a priori* masks was also reduced for the two maskers compared with a single masker.

IV. Discussion

We have shown, via the experiments summarized in Table I, that a target speech signal can be recognized in the presence of spatially separated maskers in a reverberant environment, using the statistics of binaural cues to estimate missing data masks.

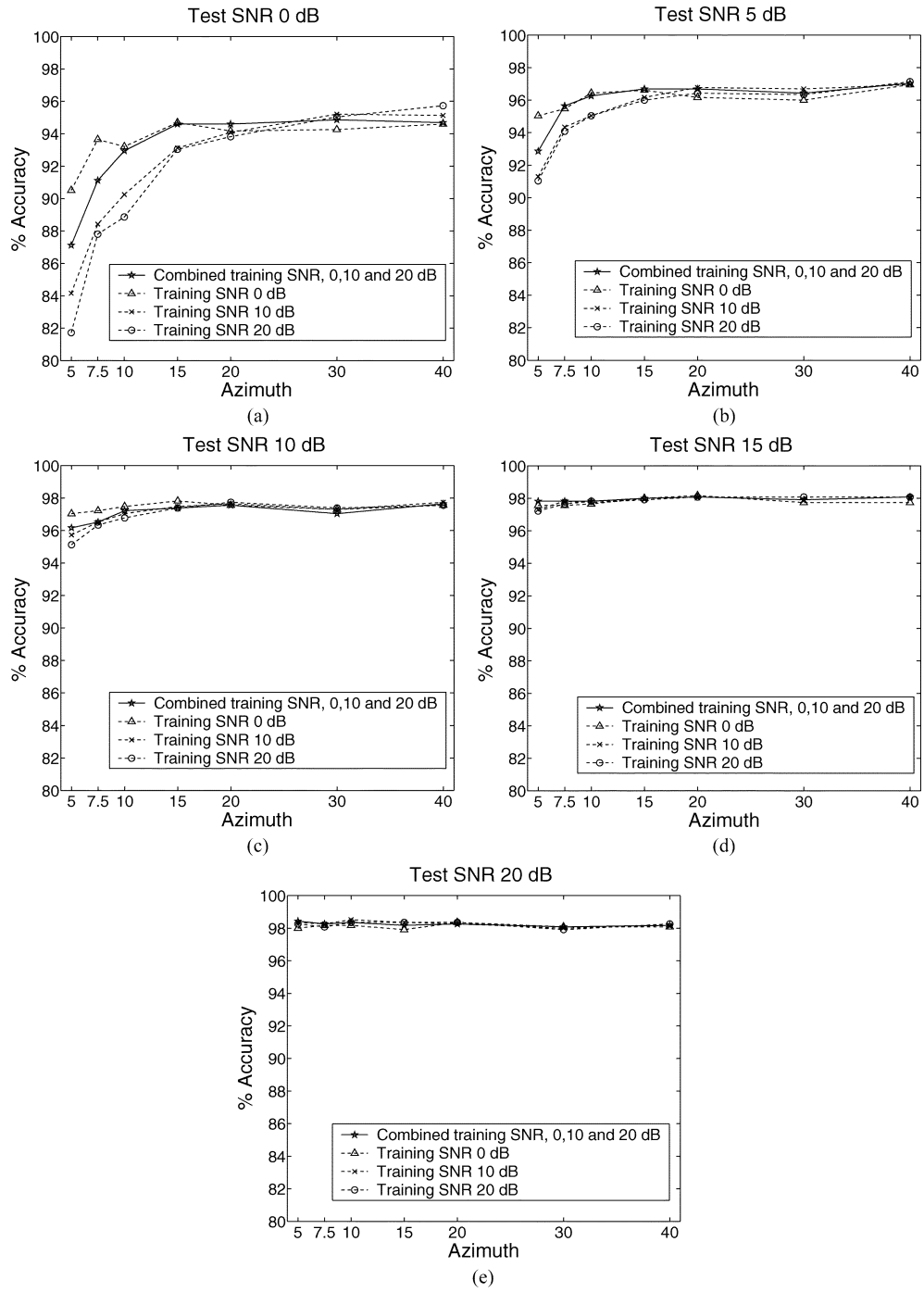

Fig. 7. Results of Experiment 4, showing the effect of varying the signal-to-noise ratio of the training data for combined ILD/ITD cues. Training data: surface "acoustic plaster," SNR 0, 10, or 20 dB (each used separately). Histogram threshold: 3. Test data: surface "acoustic plaster," SNR 0, 5, 10, 15, or 20 dB. One plot is shown per test data SNR.

The system generalized well to acoustic conditions that were not encountered during training.

Experiment 1 showed that the binaural source separation problem was most effectively addressed within a joint ITD–ILD space, rather than by using ITD or ILD alone. This finding is compatible with the theoretical analysis and simulations given by Roman *et al.* [11]. Additionally, we found that ILD alone was a far less effective cue than ITD alone in reverberant conditions. This result is consistent with predictions from psychophysics. For example, Ihlefeld and Shinn-Cunningham [26] show that reverberation decreases the mean magnitude of the ILD, making it an unreliable indicator of source azimuth. However, they suggest that the variation of the short-term ILD over time may still provide some indication of lateral position in reverberant conditions. Information about the temporal variation of binaural cues is not explicitly used in our current system, and will be investigated in future work.

As might be expected, in all experiments, recognition performance was lower when the test data were at smaller azimuth separation or had lower SNR. Both these conditions reduced the

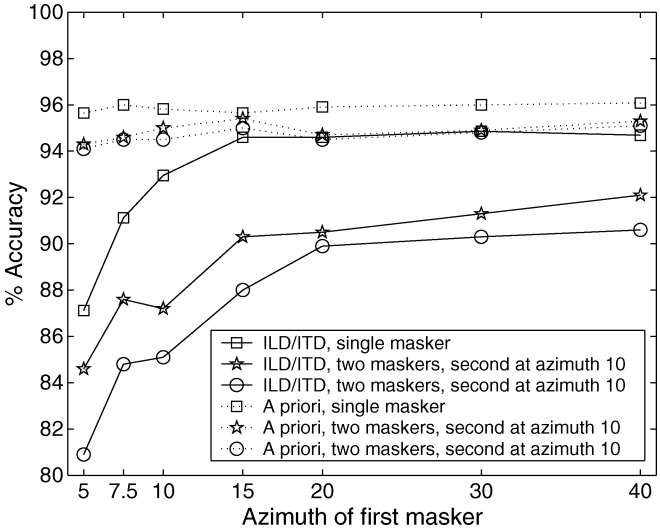

Fig. 8. Results of Experiment 5, using two masking sources, for combined ILD/ITD cues. Training data: surface "acoustic plaster," SNR 0, 10 and 20 dB (combined together). Histogram threshold: 10. Test data: surface "acoustic plaster," SNR 0 dB.

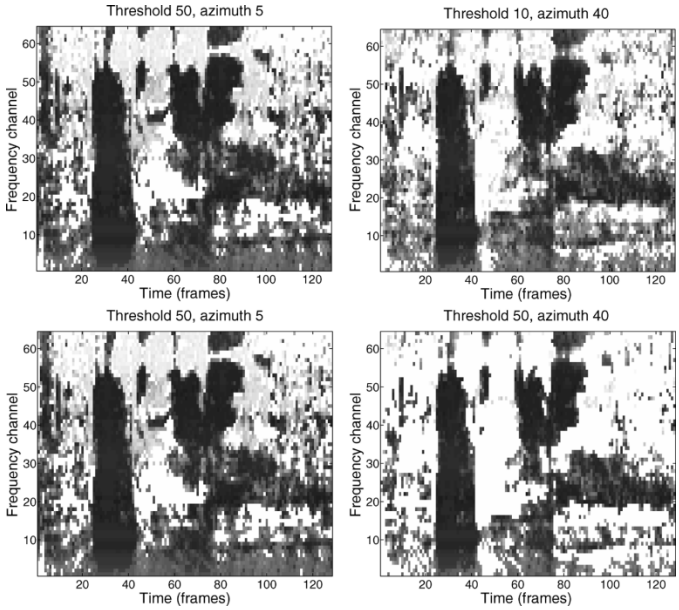

Fig. 9. Missing data masks for the mixed utterances in Fig. 1, reverberation surface "acoustic plaster," showing the effect of histogram threshold for two azimuth separations.

accuracy of the missing data masks, but for different reasons. When the azimuth separation between the sources is small, it is harder to obtain a reliable estimate of the ILD and ITD for the target and masker. At the smallest angular separation used (5°), the ILD is small and the ITD (measured as 0.045 ms from the KEMAR head-related impulse response) lies close to the smallest time lag detectable by our cross-correlation algorithm (which is limited by the sample period of 0.05 ms). When the SNR is lower, more elements will be dominated by the masker, and, therefore, the regions of the mask assigned to the target will be smaller.

The masks are affected in a similar way by the accuracy of the probability distributions, which is influenced by the training data used. Although Experiment 4 showed that the choice of reveberation surface used for training has little effect, the selection of training data for lower SNR is important, as discussed in Section III-D. Performance is also rather sensitive to the histogram threshold, as illustrated in Experiment 2. As the threshold increases, performance tends to increase for lower azimuth separation, but there is a corresponding decrease for higher azimuth separation. When the threshold is low, some of the probabilities in the distribution (and, hence, in the masks) may be excessively high; when the threshold is high, more of the probabilities will be zero. A low threshold results in more unreliable data in the mask, especially for the more difficult smaller azimuth separations, but the easier test conditions are less affected since the more reliable (higher probability) data also gets through. In contrast, a high threshold prevents some of the reliable data being included in the probability distribution and masks, but also reduces the quantity of unreliable data, which improves the performance for smaller azimuth separations but reduces performance for larger azimuth separations (Fig. 9). Using more training data, especially for the more difficult test conditions, would be expected to reduce the sensitivity of the system to the histogram threshold.

The method worked well in more challenging conditions, when multiple maskers were present (Experiment 5). The reduced performance for the *a priori* masks was probably due to the reduced variance in the combined maskers compared with a single masker, and a similar effect would be expected in the masks produced using localization cues. In the case of maskers on opposite sides of the head, two additional factors are likely to have been responsible for the reduced performance: first, the signal used for recognition was close to the second of the two maskers, so the ear advantage was reduced or removed; second, interactions between the two maskers would be expected to complicate the pattern of ILDs and ITDs and produce more confusions between the target and maskers.

Further work is required to separate the effects mentioned above, and to investigate whether extending the training data (for example, using more training utterances, additional reverberation surfaces and multiple maskers) improves performance under all test conditions. We will also use other reverberation surfaces for testing, and train and test the system with targets at other azimuths. It would also be interesting to compare the performance of this system with those of humans under similar conditions.

REFERENCES

[1] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, pp. 1–15, 1997.
[2] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. London, U.K.: Academic, 2003.
[3] P. M. Zurek, *The Precedence Effect Directional Hearing*, W. A. Yost and G. Gourevitch, Eds. New York: Springer-Verlag, 1987, pp. 85–105.
[4] W. Spieth, J. F. Curtis, and J. C. Webster, "Responding to one of two simultaneous messages," *J. Acoust. Soc. Amer.*, vol. 26, pp. 391–396, 1954.
[5] N. I. Durlach, "Equalization and cancellation theory of binaural masking level differences," *J. Acoust. Soc. Amer.*, vol. 35, no. 8, pp. 1206–1218, 1963.

[6] C. J. Darwin and R. W. Hukin, "Auditory objects of attention: the role of interaural time differences," *J. Experimental Psychology (Human Perception and Performance)*, vol. 25, no. 3, pp. 616–629, 1999.
[7] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.
[8] R. F. Lyon, "A computational model of binaural localization and separation," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1983, pp. 1148–1151.
[9] M. Bodden, "Modeling human sound-source localization and the cocktail party effect," *Acta Acust.*, vol. 1, pp. 43–55, 1993.
[10] A. Shamsoddini and P. N. Denbigh, "A sound segregation algorithm for reverberant conditions," *Speech Commun.*, vol. 33, no. 3, pp. 179–196, 2001.
[11] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, 2003.
[12] K. J. Palomäki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Commun.*, vol. 43, no. 4, pp. 361–378, 2004.
[13] M. P. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267–285, 2001.
[14] S. Harding, J. Barker, and G. J. Brown, "Mask estimation based on sound localisation for missing data speech recognition," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Philadelphia, PA, Mar. 2005, pp. I-537–I-540.
[15] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2000, vol. 13, pp. 793–799.
[16] J. Nix, M. Kleinschmidt, and V. Hohmann, "Computational auditory scene analysis by using statistics of high-dimensional speech dynamics and sound source direction," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1441–1444.
[17] B. W. Gillespie and L. E. Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Orlando, FL, May 2002, pp. 557–560.
[18] M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Commun.*, vol. 25, pp. 75–95, 1998.
[19] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 578–589, 1994.
[20] F. H. Liu, R. M. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," in *Proc. 6th ARPA Workshop on Human Language Technology*, Priceton, NJ, Mar. 1993, pp. 69–74.
[21] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 352–359, Sep. 1996.
[22] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Albuquerque, NM, Apr. 1990, pp. 845–848.
[23] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP*, vol. 3, 1984, pp. 111–114.
[24] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, no. 1–2, pp. 103–138, 1990.
[25] J. P. Barker, L. Josifovski, M. P. Cooke, and P. D. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP*, 2000, pp. 373–376.
[26] A. Ihlefeld and B. G. Shinn-Cunningham, "Effect of source location and listener location on ILD cues in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 115, p. 2598, 2004.

Sue Harding (M'05) received the B.Sc. in mathematics and computer science from the University of York, York, U.K., in 1980, the M.Sc. degree in radio astronomy from University of Manchester, Manchester, U.K., in 1982, and the Ph.D. degree in communication and neuroscience from Keele University, Keele, U.K., in 2003.

From 1982 to 1988, she worked as Pprogrammer and Systems Analyst with Simon Engineering, Stockport, U.K., and then as a Computing Officer in the Department of Computer Science, Keele University. Since 2003, she has worked as a Research Associate in the Department of Computer Science, University of Sheffield. Her research interests include auditory scene analysis and models of speech perception.

Jon Barker received the B.A. degree in electrical and information science from Cambridge University, Cambridge, U.K., in 1991 and the Ph.D degree in computer science from the University of Sheffield, Sheffield, U.K, in 1998.

He was a Researcher at ICP, Grenoble, France, and has been a Visiting Research Scientist at IDIAP, Martigny, Switzerland, and ICSI, Berkeley, CA. Since 2002, he has been a Lecturer in Computer Science at the University of Sheffield. His research interests include audio and audio-visual speech perception, robust automatic speech recognition, and audio-visual speech processing.

Guy J. Brown received the B.Sc. degree in applied science from Sheffield Hallam University, Sheffield, U.K., in 1988 and the Ph.D. degree in computer science and the M.Ed. degree from the University of Sheffield in 1992 and 1997, respectively.

He has been a Visiting Research Scientist at LIMSI-CNRS, Paris, France, ATR, Kyoto, Japan, The Ohio State University, Columbus, and Helsinki University of Technology, Espoo, Finland. He is currently a Senior Lecturer in Computer Science at the University of Sheffield. He has a long-established interest in computational auditory modeling, and also has research interests in automatic speech recognition and music technology. He has authored and coauthored more than 80 papers in books, journals, and conference proceedings.