MASK ESTIMATION BASED ON SOUND LOCALISATION FOR MISSING DATA SPEECH RECOGNITION

Sue Harding, Jon Barker and Guy J. Brown

Department of Computer Science, University of Sheffield
211 Portobello Street, Sheffield, S1 4DP, United Kingdom
{s.harding,j.barker,g.brown}@dcs.shef.ac.uk

ABSTRACT

This paper describes a perceptually motivated computational auditory scene analysis (CASA) system that combines sound separation according to spatial location with 'missing data' techniques for robust speech recognition in noise. Missing data time-frequency masks are produced using cross-correlation to estimate interaural time differenre (ITD) and hence spatial azimuth; this is used to determine which regions of the signal constitute reliable evidence of the target speech signal. Three experiments are performed that compare the effects of different reverberation surfaces, localisation methods and azimuth separations on recognition accuracy, together with the effects of two post-processing techniques (morphological operations and supervised learning) for improving mask estimation. Both post-processing techniques greatly improve performance; the best performance occurs using a learnt mapping.

1. INTRODUCTION

It is well known that speech recognition by human listeners is robust even in the presence of interfering sounds, such as the voice of another speaker. In contrast, error rates for automatic speech recognition are often more than an order of magnitude greater than those for human listeners, and are particularly large in the presence of background noise and room reverberation [1].

There is much evidence that a process of auditory scene analysis (ASA) contributes to the robustness of human speech recognition, in which listeners perceptually group sound components that are likely to have arisen from the same acoustic source [2]. An important perceptual grouping cue is spatial location; specifically, it appears that listeners can exploit a difference in location between target and masking sound sources. For example, the intelligibility of two overlapping speech signals improves as the spatial separation between them is increased [3].

Human listeners are able to localise sounds mainly using information about differences in sound intensity and time of arrival at the two ears; so-called interaural time differences (ITD) and interaural intensity differences (IID) [4]. Computational auditory scene analysis (CASA) systems which use these cues may offer an effective means for machine separation of sounds, as a precursor to automatic speech recognition. In particular, such perceptually-motivated systems may offer an approach to sound separation which makes fewer assumptions about the number of sound sources and their characteristics than blind statistical techniques (for a review, see [5]).

This work was funded by EPSRC grant GR/R47400/01.

In this paper, we describe a CASA system which combines sound separation according to ITD with 'missing data' techniques for robust speech recognition in noise [6]. In our approach, a binaural auditory model is used to construct a time-frequency mask which indicates whether each acoustic feature constitutes reliable evidence of the target speech signal or not. The acoustic features and corresponding mask are then passed to a missing data speech recogniser, which treats reliable and unreliable regions differently during decoding. Our particular concern is how localisation cues, and subsequent image processing operations, can be used to derive a near-optimal time-frequency mask.

The current paper extends our previous work ([7], [8]) in several respects. Firstly, we compare the effectiveness of a number of localisation algorithms. Secondly, we investigate post-processing of time-frequency masks using morphological image processing operations (erosion/dilation) and supervised learning techniques. Thirdly, we use a wider variety of test conditions that are more representative of real acoustic scenes, including randomised interfering speech utterances. Finally, we compensate for the effects of reverberation by training the recogniser on reverberated speech, rather than by correcting for spectral distortion using normalisation.

2. OVERVIEW OF METHODS

2.1. 'Missing data' speech recognition

The missing data recogniser was based on Hidden Markov models (HMMs) trained on clean spatialised and reverberated utterances from the TI digits corpus [9]. The Roomsim simulator[1] was used to produce impulse responses for a room of size 6 m x 4 m x 3 m with MIT data for a KEMAR head[2] in the centre of the room, 2 m above the ground, with a source at azimuth 0, 5, 10, 20 or 40 degrees at a radial distance of 1.5 m. The left and right ear impulse responses for the appropriate azimuth and reverberation surface were convolved with the monaural data to produce binaural reverberated data. All surfaces within the room were assumed to have the same reverberation characteristics. Two reverberation surfaces were used, 'acoustic plaster' and 'platform floor wooden', with the characteristics listed in table 1. Eight-state ten-mixture HMMs with delta coefficients were trained using 4228 clean speech utterances by 55 male speakers[3] at azimuth 0 for each of the two surfaces.

[1] http://media.paisley.ac.uk/~campbell/Roomsim/
[2] http://sound.media.mit.edu/KEMAR.html
[3] female speakers would be expected to perform equally well

Table 1. Reverberation surface characteristics, showing the estimated T60 reverberation time at standard frequencies for two surfaces, 'acoustic plaster' (AP) and 'platform floor wooden' (PFW). Times are in seconds.

Surface	Frequency (Hz)						
	125	250	500	1000	2000	4000	Mean
AP	1.02	0.49	0.17	0.14	0.11	0.11	0.34
PFW	0.22	0.31	0.49	0.48	0.65	0.89	0.51

Recognition was performed using a separate test set of 240 utterances, also at azimuth 0, each mixed at 0 or 20 dB with one of a set of 240 interfering male utterances, matched in length to the original 240 utterances; the interfering speech was reverberated and spatialised at one of the other azimuths listed above. The signal-to-noise ratio (SNR), i.e. test utterance to interfering utterance, was calculated from data spatialised at azimuth 0.

Each signal was processed to form an auditory spectrogram (figure 1), which constituted the acoustic features for the recogniser. These were produced using a 64-channel gammatone filterbank with centre frequencies ranging from 50 Hz to 8 kHz, an analysis window of 20 ms and frame shift of 10 ms on data sampled at 20 kHz. The mixed speech signal entering the ear furthest from the interfering speech was used for recognition.

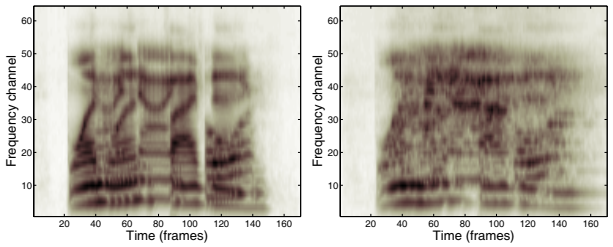

Fig. 1. Auditory spectrograms for the utterance 'one one five nine' at azimuth 0 mixed at SNR 0 dB with utterance 'nine nine four' at azimuth 40, both by male speakers: left, anechoic; right, reverberated, surface 'platform floor' (see table 1).

2.2. Localisation for the missing data mask

In addition to the mixed speech signal to be recognised, the missing data speech recogniser requires a time-frequency mask indicating regions of the signal that can be considered as reliably belonging to the source of interest (figure 2). Localisation cues were used to separate the two sources and to determine the mask. In this case, the reliable regions were considered to be those that appeared to originate from a source close to azimuth 0. For comparison, *a priori* masks were also produced: these are ideal masks created using *a priori* knowledge of the difference between the clean speech and the mixed speech signals (figure 2, left).

The azimuth of each source was determined from the cross-correlogram for each time frame produced by passing each of the two binaural inputs through an auditory filterbank, computing the cross-correlation between each frequency channel to estimate the ITD, warping the ITD to its corresponding azimuth and then emphasising the peaks by convolving each peak with a Gaussian to produce a 'skeleton' cross-correlogram (figure 3, left: see [7] for details). Peaks in the cross-correlogram indicated the possible location of a sound source.

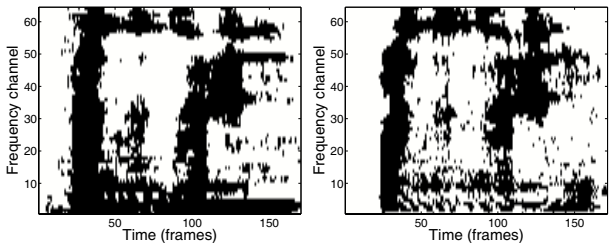

Fig. 2. Missing data masks for the reverberated data in figure 1: left, *a priori* mask; right, produced using localisation.

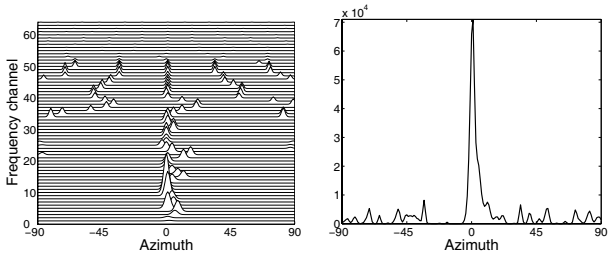

Fig. 3. Left, the skeleton cross-correlogram, for time frame 42, for the two utterances in figure 1; right, the summary cross-correlogram for the same time frame.

Two methods were used to determine the missing data mask from the cross-correlogram. The first (the 'summary' method) used a summary cross-correlogram, produced by summing over all channels in each frame, to identify the dominant azimuths associated with that frame (figure 3, right); the cross-correlogram channel energy at these azimuths was compared and the azimuth of the largest value was used as the azimuth of the dominant source in that channel and frame. The second method (the 'no summary' method) used the energy in each channel directly, selecting the azimuth of the largest peak as the azimuth of the dominant source. In each case, an element of the mask corresponding to a particular time frame and frequency channel was set to 1 if the dominant source was at azimuth 0 (to a given tolerance: in this case plus or minus one degree); otherwise the element was set to 0. Note that it is also possible to define 'soft' masks (as opposed to discrete masks) in which each element of the mask takes a real value between 0 and 1: this type of mask was used in experiment 3.

2.3. Post-processing the masks

Two techniques were applied to the localisation masks described above in order to improve recognition accuracy: (a) erosion and dilation to remove noise from and fill holes in the masks (figure 4, left); (b) using artificial neural networks (ANNs) to transform the masks to be more similar to *a priori* masks (figure 4, right).

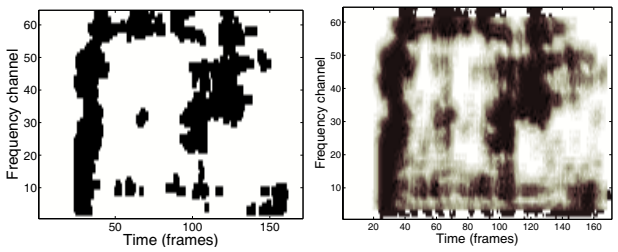

Fig. 4. Missing data masks produced by post-processing the localisation mask shown in figure 2: left, after erosion and dilation with a square of side 3 (see section 3.2); right, soft mask learned from ANN (see section 3.3)

3. EXPERIMENTS

3.1. Experiment 1 - baseline localisation masks

Experiment 1 investigated the effects of localisation method, reverberation surface and azimuth separation on recognition performance. Localisation masks were produced using the 'summary' and 'no summary' methods described in section 2.2. The recognition accuracy when using *a priori* masks (section 2.2) was also measured.

Figure 5 (circles) shows the basline recognition accuracy for the two localisation methods and the two reverberation surfaces, for SNR 0 dB. The baseline results were better for the 'no summary' method than for the 'summary method' in all cases except one (azimuth 40, acoustic plaster) in which the results differed by less than 0.5%; the difference between localisation methods was most pronounced for the smallest azimuth separation. In all conditions, accuracy was better for larger azimuth separation. The *a priori* mask results were around 96% for the 'acoustic plaster' surface and 91-92% for the 'platform floor' surface.

Recognition accuracy for SNR 20 dB was very similar for both localisation methods, being around 97-98% for the 'acoustic plaster' surface and around 94% for the 'platform floor' surface. The *a priori* mask results for these two surfaces were around 98% and 95% respectively.

3.2. Experiment 2 - post-processing using morphological operations

In experiment 2, the localisation masks produced by experiment 1 were post-processed using the morphological operations of erosion and dilation. Each of these operations treats the mask as an image and applies a structuring element, which defines the neighbouring pixels, to each pixel in turn. Erosion removes noise from the mask by setting the neighbouring pixels to zero if the input pixel is zero; dilation fills holes in the mask by setting the neighbouring pixels to one if the input pixel is one. Performing erosion followed by dilation removes small objects from the image while preserving the shape and size of larger objects. Each mask was processed using a structuring element consisting of a square of side 2 pixels (condition 'ed2') or of side 3 pixels (condition 'ed3').

Figure 5 (triangles) shows the recognition accuracy for these two conditions. In general, applying erosion and dilation was beneficial, especially for the smaller azimuth separations, and reduced

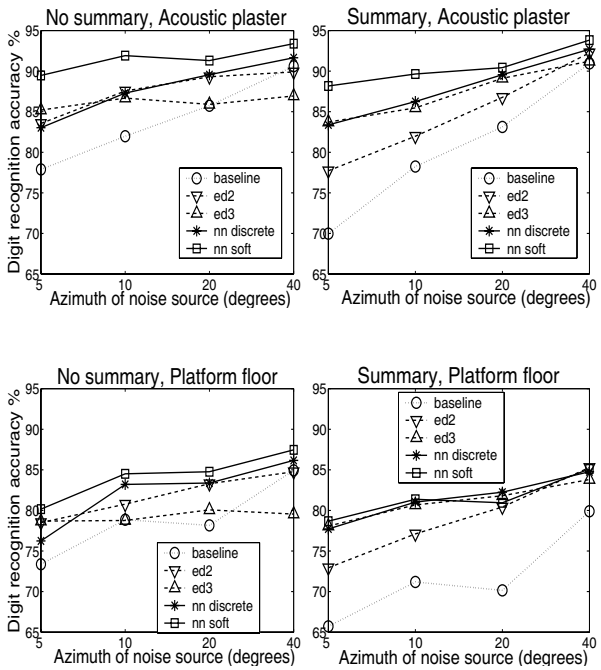

Fig. 5. Results of experiments 1 (baseline), 2 (ed2/ed3) and 3 (nn discrete/nn soft), using two localisation methods and two reverberation surfaces, for SNR 0 dB.

the differential effects of the localisation method: performance using the 'no summary' method with condition 'ed2' was similar to that for the 'summary' method with condition 'ed3'. This applied to both reverberation surfaces.

Recognition accuracy for SNR 20 dB was within 1% of that for experiment 1.

3.3. Experiment 3 - post processing using a learnt mapping

Rather than post process the localisation masks using an *ad hoc* morphological operation it may be possible to learn a better mapping directly from the data. To investigate this idea, artificial neural networks (ANNs) were trained to estimate the *a priori* mask from the unprocessed localisation masks. Each point in the localisation mask, l_{tf}, was formed into a feature vector by supplementing it with the points in a square context window extending n frames backward and forward in time, and n channels up and down in frequency. The ANN was trained to map this input onto the corresponding value of the *a priori* mask, a_{tf}. The input context size, n, was set to either 3, 4, or 5. For each value of n, ANNs were trained with either 20, 60, or 120 hidden units. Thus a total of nine different ANN topologies were considered. In all cases linear output nodes were employed.

Training data consisted of 24 utterance pairs mixed at 0 dB in the 'acoustic plaster' reverberation condition. The pairs were mixed with separations of 5, 10, 20 and 40 degrees (making 96 pairs in total). 500 training examples were constructed from each of the 96 mixtures by randomly sampling points from the localisation mask to form a 48,000 example 'azimuth-independent' training set. The random sampling was designed to avoid points at

the mask edges to ensure that the time-frequency context window could be accommodated.[4] Where the localisation map and its context were uniformly 0, the target was set to 0 regardless of the actual value in the *a priori* mask. This was found to be necessary to prevent undue influence of 'noisy' regions in the *a priori* mask where the target is only just above the masking threshold. The ANN was trained using 100 iterations of scaled conjugate gradient descent. Performance was monitored during training using a small cross validation set to safeguard against over-fitting.

Discrete mask estimates were constructed by selecting zero or one depending on whether the network outputs were less than or greater than 0.5. Continuously-valued (soft) masks were also constructed in which the ANN outputs were used directly, except that values greater than one were set to one, and negative values were set to zero so as to preserve a probabilistic interpretation.

ANNs were trained for each of the 9 topologies described above. The localisation masks for the 240 utterance test set mixed in the 'acoustic plaster' condition were processed by each network. For each topology the percentage decrease in L1 distance between the localisation mask and the *a priori* mask was measured. The decreases were relatively small, i.e. around 10–15%. The ANN with a context size of 3 and with 60 hidden units performed as well as any of the larger networks and was therefore selected for use in the recognition experiments.

The above procedure was used to train ANNs for post processing localisation masks produced using both the 'summary' and the 'no summary' methods The resulting ANNs were applied to the test data in both the 'acoustic plaster' condition (used in ANN training) and the 'platform floor' condition (unseen during training). The resulting discrete and soft masks were tested in conjunction with the recognition system described earlier. In the case of soft masks, mask values are treated as a probability that the data is reliable, and the missing data acoustic model probability calculation becomes an interpolation between the 'missing' and 'present' interpretations (see [10] for details).

Figure 5 (solid lines) presents results for both the discrete and soft masks. The discrete masks performed as well as, or better than any of the discrete masks generated by the morphological operations. Using the ANN outputs to form soft masks produced a further increase in recognition performance. Largest gains were observed for the 'acoustic plaster' condition on which the ANNs were trained; however, the system also performed well in the more severe, unseen 'platform floor' condition, i.e. the ANN was able to generalise from one reverberation condition to another.

4. CONCLUSIONS

Recognition accuracy for the baseline results (i.e. without any post-processing) was better for the simpler 'no summary' localisation method than for the 'summary' method, but there was little difference between the two methods after post-processing using morphological operations (erosion/dilation); these operations produced a large improvement in performance, and tended to reduce the effect of azimuth separation. An even greater increase in performance was seen for the second post-processing technique using a learnt mapping. This technique provides a more principled approach than the rather *ad hoc* hand-tuning used in experiment 2 and in [7], and generalised well to a second reverberation surface and across different azimuth separations. It is likely, however, that the greater improvement seen for the soft masks over the discrete masks was due largely to the transformation of a hard decision (which may be wrong) into a softer decision [10]. It may be possible to improve performance further by creating masks learnt directly from cross-correlograms or from soft masks that make use of more of the underlying information, rather than from discrete masks from which information has already been removed.

In these experiments, we have assumed that the target source location is at azimuth zero. This is a safe assumption for many applications and simplifies the early stages of determining the masks, but is not a strict requirement: we assume only that the position of the target source is known, whereas the distracting sources may have unknown locations. Furthermore, although only one distracting source was used in these experiments, no assumptions were made about the number of sources present and the techniques described are expected to work well for multiple distractors. Future work will deal with targets at unknown locations as well as multiple distracting sources and moving sources, and will also test the generality of this approach, using a wider variety of test conditions. We will also investigate whether improved localisation masks can be obtained by combining information from interaural level and time differences.

[4] Note, when processing the test data, points close to the edge of the localisation masks which do not have sufficient context were copied directly to the *a priori* mask estimate unaltered.

5. REFERENCES

[1] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, pp. 1–15, 1997.

[2] A.S. Bregman, *Auditory Scene Analysis: the perceptual organization of sound*, MIT Press, Cambridge, MA, 1990.

[3] W. Spieth, J.F. Curtis, and J.C. Webster, "Responding to one of two simultaneous messages," *Journal of the Acoustical Society of America*, vol. 26, pp. 391–396, 1954.

[4] B.C.J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, London, 5th edition, 2003.

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley & Sons, 2001.

[6] M.P. Cooke, P.D. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[7] 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 Communication*, In press.

[8] N. Roman, D.L. Wang, and G.J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[9] R.G. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP*, 1984, vol. 3, pp. 111–114.

[10] 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.