Binaural Feature Selection for Missing Data Speech Recognition

Sue Harding, Jon Barker and Guy J. Brown

Department of Computer Science
University of Sheffield, UK
{s.harding,j.barker,g.brown}@dcs.shef.ac.uk

Abstract

The 'missing data' approach for robust speech recognition uses masks indicating which regions of an acoustic mixture provide reliable evidence of the target to be recognised. Binaural cues for spatial location were used to determine missing data masks for signals consisting of utterances from three concurrent male speakers in reverberant conditions, by deriving probability distributions from estimates of interaural time and level differences (ITD and ILD) for the mixed signals. In such a system, a decision must be made about whether the acoustic features used for decoding are selected from the left or right ear, or a combination of the two. Here, features were selected from the "better ear" (as determined by a simple heuristic) within whole time frames, or within individual time-frequency elements. A combination of left and right ear features gave better recognition performance than using either ear alone, and the best results were obtained when selecting features within individual time-frequency elements.

1. Introduction

Robust performance in multi-source acoustic environments, such as meeting rooms, remains a challenging problem for automatic speech recognition (ASR). The solution to this problem may lie in an approach to ASR which is more strongly motivated by mechanisms of human hearing. In particular, computational models of *auditory scene analysis* (ASA) – the process by which listeners extract a perceptual description of a single source from an acoustic mixture – may offer effective front-end processing for robust ASR in adverse conditions.

One of the cues that listeners exploit in ASA is spatial location; specifically, listeners tend to perceptually segregate acoustic events that arise from different locations in space [1]. For example, Spieth et al. [2] have shown that the intelligiblity of two overlapping speech signals increases as the spatial separation between them is increased. Motivated by such observations, a number of workers have described computational systems for sound separation that exploit binaural cues to source direction [3, 4, 5]. Typically, such systems consist of four stages. In the first stage, audio input is acquired from a pair of spatially separated microphones or a dummy head (such as the KEMAR). Secondly, each audio input is processed by a bank of bandpass filters, which splits the input into different frequency channels. In the third stage, running estimates of the interaural time difference (ITD) and interaural level difference (ILD) are obtained for each frequency channel. Finally, directional filtering is achieved by selectively weighting time-frequency regions whose ITD and ILD correspond to the location of the target sound source.

The 'missing data' approach to ASR [6] provides an effective framework for linking such binaural sound separation approaches with speech recognition. In this approach, the binaural model is used to derive a time-frequency mask which indicates whether each acoustic feature constitutes reliable evidence of the target speech signal or not. The mask and the acoustic features are then passed to a modified hidden Markov model (HMM) decoder that treats reliable and unreliable features differently during decoding [4, 7].

An issue that arises in such systems is whether acoustic features should be selected from the left or right ear; both are available, but only one set of features is required for decoding. For example, consider the case where the target source is located at zero degrees azimuth (i.e., straight ahead). If a single interferer is present on one side of the head (or multiple interferers are located on the same side of the head) then the acoustic features from the ear furthest from the interference should be used, since these are likely to be the least corrupted. When maskers occur on both sides of the head, the choice of which features to use for decoding is not so obvious.

Again, a solution to this problem is suggested by perceptual studies. Devore and Shinn-Cunningham [8] have investigated the origin of the binaural advantage for human listeners in reverberant and multi-source environments, and conclude that listeners *dynamically* select the "better ear" according to short-term estimates of the target-to-masker ratio. We adopt a similar approach here, within the 'missing data' ASR framework. In particular, we consider whether acoustic features from the "better ear" should be selected on a frame-by-frame or channel-by-channel basis.

2. System description

2.1. The 'missing data' speech recogniser

The 'missing data' recogniser uses HMMs trained on spectro-temporal acoustic features. During recognition, it takes two types of input: firstly, the signal to be recognised, also in the form of spectro-temporal acoustic features (figure 1); secondly, a 'missing data mask' which indicates which portions of the signal can reliably be assumed to belong to the source of interest. The mask may be discrete, with each element set to 0, meaning that the target is masked, or to 1, meaning that the target is dominant; alternatively the mask may be 'soft', with each element having a real value between 0 and 1 representing the probability that the target is dominant [9]. Figure 2 shows an example of a discrete mask determined using *a priori* knowledge of the target and masker, in which an element is set to 1 if the ratio of energy in the mixed signal to that in the clean signal is less than a threshold and set to 0 otherwise, and a soft localisation mask, determined using spatial localisation cues from binaural data, as described below.

Utterances from the TIDigits corpus [10] were used for training and testing the recogniser. Reverberation and spatiali-

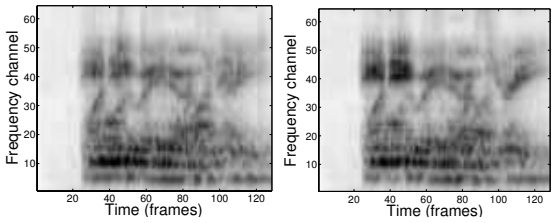

Figure 1: Auditory spectrograms for the left and right ears for the utterance 'one two eight oh' at azimuth 0 mixed at SNR 0 dB with utterance 'eight eight four three' at azimuth 30, and 'four two one eight' at azimuth -30, all by male speakers, with reverberation surface 'acoustic plaster'.

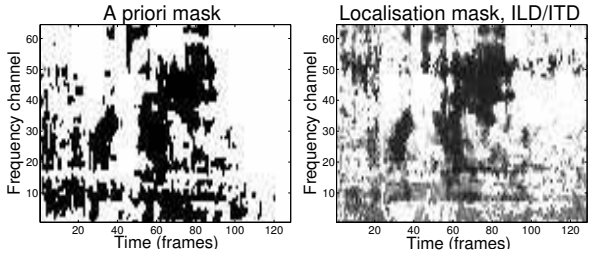

Figure 2: Missing data masks for the mixed utterances in Fig. 1: left, discrete *a priori* mask (calculated from utterances at azimuth 0); right, soft mask produced using localisation cues as described below. Lighter areas have lower probability; darker areas higher probability.

sation were applied to these utterances using the Roomsim simulator[1] with a simulated room of size 6 m x 4 m x 3 m. The receiver was a KEMAR head (data from [11]) in the centre of the room, 2m above the ground, and the source was at azimuth 0, 5, 7.5, 10, 15, 20, 30 or 40 degrees at a radial distance of 1.5 m from the receiver. All surfaces of the room were assumed to have identical reverberation characteristics, defined as surface 'acoustic plaster', with mean estimated T60 reverberation times of 0.34 seconds. Impulse responses were determined for each of the azimuths listed above, and convolved with the monaural utterances to produce binaural reverberated and spatialised data.

Each binaural signal was passed through a 64-channel gammatone filterbank with centre frequencies from 50 Hz to 8 kHz, an analysis window of 20 ms and a frame shift of 10 ms. Interframe differences (delta features) were concatenated with the output of the filterbank to create the acoustic feature vectors for the recogniser.

A set of 4228 clean reverberated utterances by 55 male speakers, spatialised at 0 degrees azimuth, were used to train the recogniser, which consisted of eight-state ten-mixture HMMs.

2.2. Missing data mask estimation

A further set of training data was used to determine probability distributions that were used to create soft missing data masks. This training set consisted of 120 pairs of utterances, matched for length, with one utterance at 0 degrees azimuth (corresponding to the location of the target utterance) and another at 5, 10, 20 or 40 degrees, or at -5, -10, -20 or -40 degrees. After reverberation and spatialisation, the binaural signals were mixed at signal-to-noise ratios (SNR) of 0, 10 or 20 dB (the SNR was calculated from data spatialised at azimuth 0 degrees).

Interaural time and level differences (ITD and ILD) were determined for each of these mixed utterances by passing each of the binaural inputs through the gammatone filterbank described above and then cross-correlating each pair of frequency channels for each frame. The largest peak in each channel was used to estimate the ITD. The ILD was calculated by summing the energy in each channel and finding the ratio of the energy in each ear.

Two histograms were produced from the ITD and ILD values found for the training data, by assigning each to a bin (of size 0.1 for ILD and 0.01 for ITD). The first histogram counted observations of combinations of ILD and ITD produced by both the target and the masker, and the second counted only those observations produced by the target. Observations were assigned to target or masker by examining each element of the corresponding *a priori* mask. The ratio of the second (target) histogram to the first histogram (for both target and masker) provided the probability that any combination of ILD and ITD was produced by a source at azimuth 0, i.e. the target source (see [5] for a related approach). Since there was a wide variation in the distributions for different channels (figure 3), separate probability distributions were produced for each frequency channel. A gradual progression in the shape of the distribution can be seen as the dominance of each of the two cues varies.

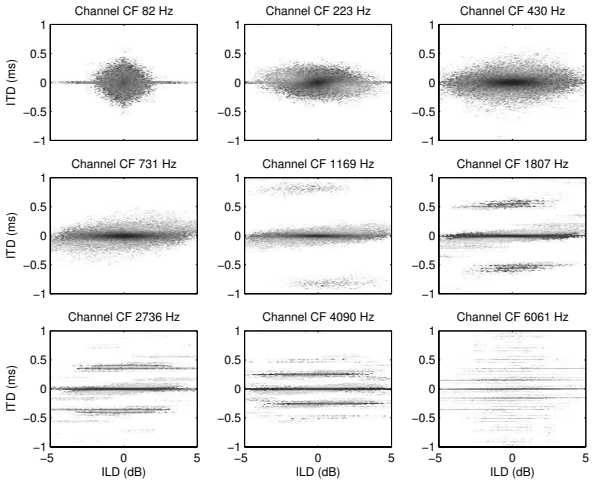

Figure 3: Examples of ILD/ITD probability distributions for a source at azimuth 0 degrees, for a selection of 64 frequency channels equally spaced on the ERB scale, with the centre frequencies shown. Lighter areas have lower probability; darker areas higher probability.

A threshold was also applied to the histogram for the target plus masker to reduce the effect of time-frequency elements for which insufficient training data was present, as such elements would otherwise have resulted in an excessively high probability when only a few elements were present but were allocated to the target. Elements with histogram values below a threshold of 10 were treated as if no data were present, which smoothed the distributions.

Missing data masks were determined by calculating the ITD and ILD for each time-frequency element of a test utterance and using the probability distribution as a look-up table

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

for that combination of ILD and ITD to find the probability that the element was dominated by a source at azimuth 0. An example of a mask created in this way is shown in figure 2.

The masks were passed to the missing data recogniser along with the mixed signal to be recognised, and the recognition accuracy was measured for a set of test utterances described below.

3. Experiments

The aim of the experiments was to investigate the effect on recognition performance of varying the acoustic features used for decoding, by selecting or combining features for one or both ears. A source at azimuth zero would be expected to produce similar signals in each ear, while a source at some other azimuth would be expected to produce a more intense signal in the ear closest to the source. When a target source at azimuth 0 is added to a masking source at some other azimuth, the most advantageous ear would be expected to be the one furthest from the masker and, due to the symmetrical placement of the target and the asymmetry of the masker's position, the signal entering this ear would be quieter than that entering the other ear. When two maskers are present, the most advantageous ear may vary depending on the positions of the maskers as well as instantaneous changes in the masking utterances. The experiments tested the effect of using the signals entering either ear as well as creating composite signals by combining the right and left signals according to which was likely to be the most advantageous.

The test data consisted of 240 utterances by male speakers (different from those used in the training sets), each mixed with two other utterances also by male speakers. The target utterance was always at azimuth 0; the first of the two masking utterances was at one of 7 azimuths (5, 7.5, 10, 15, 20, 30 or 40 degrees) and the second was either at azimuth -10 or +10 (asymmetrical maskers) or at one of the azimuths -5, -7.5, -10, -15, -20, -30 or -40 such that the two maskers were symmetrically placed on either side of the head (symmetrical maskers). The masking utterances were mixed at 0 dB SNR (calculated from utterances spatialised at azimuth 0) and the maskers were then mixed with the target utterance at SNR 0 dB (also measured at azimuth 0).

Composite signals were created by combining features from the left and right ears in one of two ways, based on the assumption that it was advantageous to select the less intense of the two features, since this was least likely to be corrupted by noise. First, pairs of time frames from the left and right ears were compared, and the frame with the lowest overall energy was selected, to create 'per frame' composite features. Second, each time-frequency element from the left ear was compared with the corresponding element from the right ear and the element with the lowest energy was selected, to create 'per element' composite features (figure 4). In each case, delta features were copied from the appropriate ear.

Recognition was performed on the two asymmetrical and one symmetrical test sets using each of the two types of composite signal. The recogniser also used the missing data masks produced using localisation cues as described above. Results for each of the three masker configurations and 7 azimuth separations are shown in figure 5, with mean performance over all azimuth separations shown in figure 6.

Results using the original signals for the left and right ears were similar when the maskers were symmetrically placed on different sides of the head, as would be expected. When the maskers were asymmetrically arranged on different sides of the head (i.e. the second masker was at azimuth -10), using the right

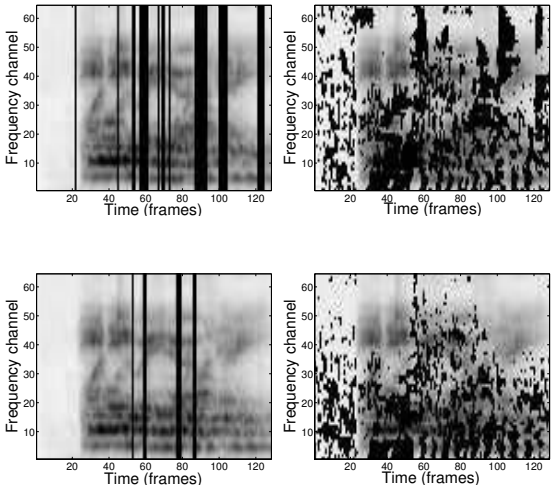

Figure 4: Effect on spectrograms of combining features from left and right ears: top, for the mixed utterances in figure 1 with maskers at 30 and -30 degrees azimuth (symmetrical); bottom, for the same utterances with maskers at 30 and 10 degrees (asymmetrical). The dark areas show the frames or elements of the left ear that were changed by selecting the quieter ear. Left, effect of selecting the quieter ear per frame; right, effect of selecting the quieter ear per element.

ear gave better performance for the lower azimuth separations: there appeared to be an advantage in using the ear furthest from any masker in these more difficult listening conditions. When the maskers were on the same side of the head (i.e. the second masker was at azimuth 10), there was a clear advantage in using the ear on the other side of the head for recognition.

Using a composite signal produced better mean performance than using either ear alone, in all three masker configurations, as shown in figure 6. The performance when combining the signals on a per element basis was better than when combining them per frame. The improvement when using composite masks rather than a single ear was greatest for the smaller azimuth separations (5).

4. Discussion

These results showed that, when multiple masking sources were present, the ear furthest from any masking source gave the better performance. However, using features from only one ear did not perform as well as dynamically selecting features from either ear based on an estimate of the "best ear". The improvement resulting from the use of composite signals was greatest when the maskers were close to the target.

The results for the 'per frame' composite signals are broadly consistent with the perceptual results in [8], in which it was suggested that listeners may be able to switch between ears at each instant to select the ear with the highest target-to-masker ratio. Our results indicate that an even greater advantage can be obtained by selecting the better ear within small time-frequency regions. It is an interesting question whether human listeners also obtain optimal performance by selecting features from different ears within the same critical band.

The method used here discards information from the left or right signals. An alternative would be to perform recognition

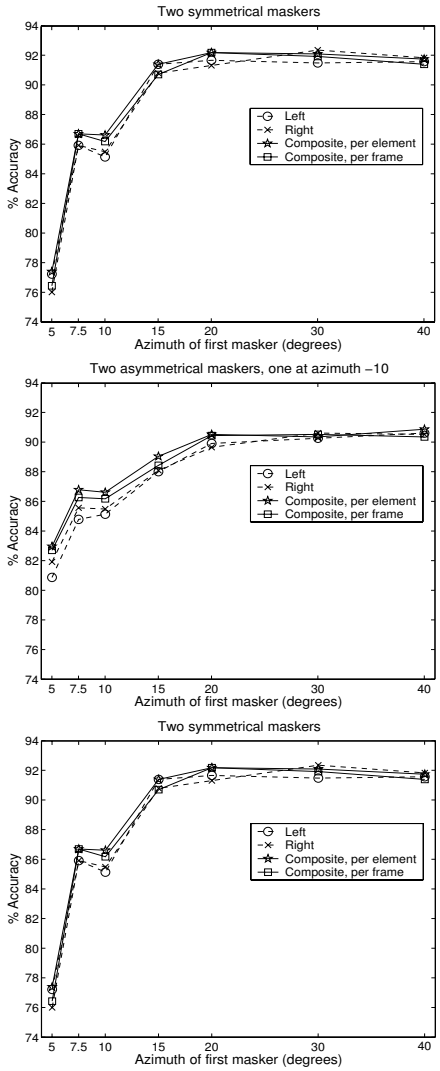

Figure 5: Recognition results: top, using symmetrical maskers; middle, using asymmetrical maskers with the second masker at azimuth -10 degrees; bottom, using asymmetrical maskers with the second masker at azimuth 10 degrees.

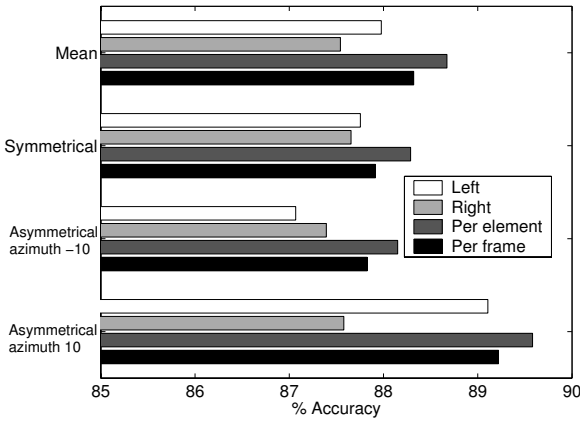

Figure 6: Mean recognition results over all azimuths for each masker configuration and type of input signal, and over all three configurations.

on the signals from both ears and combine the resulting likelihoods, although it remains to be seen whether the additional information from the less advantageous ear would be beneficial. Unless such a method is adopted, it is necessary to make an informed decision over which parts of the binaural input signal should be used for recognition as the selection of one ear over another may result in less than optimal recognition performance unless the configuration of target and masking sources is known in advance.

The method described here has proved successful and is attractive in its simplicity. Further work will investigate whether it can be applied in more general conditions, with the receiver, target and room aligned asymmetrically.

5. Acknowledgements

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

6. References

[1] A. S. Bregman, *Auditory scene analysis*. Cambridge, MA: MIT Press, 1990.

[2] 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.

[3] M. Bodden, "Modelling human sound-source localization and the cocktail party effect," *Acta Acustica*, vol. 1, pp. 43–55, 1993.

[4] 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*, vol. 43, no. 4, pp. 361–378, 2004.

[5] 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.

[6] M. P. Cooke, P. 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] S. Harding, J. Barker, and G. J. Brown, "Mask estimation based on sound localization for missing data speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, March 2005.

[8] S. Devore and B. Shinn-Cunningham, "Perceptual consequences of including reverberation in spatial auditory displays," in *Proceedings of ICAD*, Boston, MA, July 2003, pp. 75–78.

[9] J. P. Barker, L. Josifovski, M. P. Cooke, and P. D. Green, "Soft decisions in missing data techniques for robu st automatic speech recognition," in *Proc. ICSLP*, 2000, pp. 373–376.

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

[11] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.