Binaural cues for fragment-based speech recognition in reverberant multisource environments

Ning Ma, Jon Barker, Heidi Christensen, Phil Green

Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
{n.ma, j.barker, h.christensen, p.green}@dcs.shef.ac.uk

Abstract

This paper addresses the problem of speech recognition using distant binaural microphones in reverberant multisource noise conditions. Our scheme employs a two stage fragment decoding approach: first spectro-temporal acoustic source fragments are identified using signal level cues, and second, a hypothesis-driven stage simultaneously searches for the most probable speech/background fragment labelling and the corresponding acoustic model state sequence. The paper reports the first successful attempt to use binaural localisation cues within this framework. By integrating binaural cues and acoustic models in a consistent probabilistic framework, the decoder is able to derive significant recognition performance benefits from fragment location estimates despite their inherent unreliability.

1. Introduction

Automatic speech recognition (ASR) technology is finally starting to become commonplace. However, in most applications the expectation is that the user is employing a close-talking microphone. For ASR technology to become truly ubiquitous it needs to be freed from this constraint and designed to work reliably with *distant* microphones.

The scarcity of distant microphone ASR applications is not a lack of demand, but rather because recognition in these conditions is a difficult and largely unsolved problem [1]. There are two sources of variability that make it more challenging than close-talking ASR. First, there exists an increased *channel variability*. The speech signal arriving at the microphone is reverberated by a room response, which in turn is dependent on a host of details that may be changing over time in significant and unpredictable ways. Second, there will generally be substantial additive noise because the microphones will unselectively capture signals from all sound sources in the environment. Further, most 'everyday' environments will contain an unknown number of sound sources whose activity level – and possibly location – is changing over time. Fig. 1 displays a time-frequency (T-F) representation of audio recorded in a family home that gives some indication of the complexity of a typical acoustic scene.

Our approach to distant microphone ASR is inspired by the human ability to attend to individual components of complex acoustic mixtures, even when only presented with a single acoustic channel [3]. We model this ability using a two-stage approach: first, an 'auditory' front-end exploits the continuity of signal characteristics to identify robust spectro-temporal source *fragments*, i.e. regions in the spectro-temporal domain in which the energy is dominated by a single acoustic source. Second, a statistical back-end, through a process termed *fragment decoding*, selects sound source fragments based on the extent to which they match models of the target source [4].

This work was supported by EPSRC grant CHiME, EP/G039046/1.

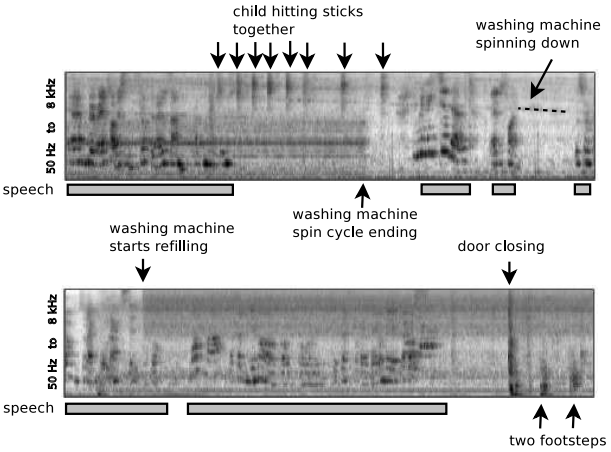

Figure 1: A time-frequency representation of a 20 second sample from the binaural CHiME domestic audio corpus used as noise background in the current study [2].

This paper reports an original *binaural* extension to the fragment decoding approach which incorporates spatially motivated cues to bias the decoder towards accepting fragments that are believed to originate from a known target source location. Section 2 reviews the basic fragment decoding framework. Section 3 describes the audition-inspired techniques that are used to isolate and localise the source fragments that act as input to the decoding process. The reverberant binaural speech-in-noise data used for evaluation is described in Section 4. Section 5 examines the performance of the fragment processing front-end and compares the recognition performance delivered by various fragment localisation strategies. Section 6 discusses future directions and concludes this paper.

2. The fragment decoding framework

The energy in a speech signal is not evenly spread across time and frequency but instead is highly concentrated in local T-F regions (e.g. formant resonances). Typically, even when the noise background has higher energy than the speech on average, in these local regions the speech energy will be many decibels greater than the noise. This view of masking leads naturally to the 'missing data' approach to robust ASR [5].

The difficulty with the missing data ASR approach is that the foreground/background *segmentation* is obviously not provided *a priori*. In some situations a good candidate segmenta-

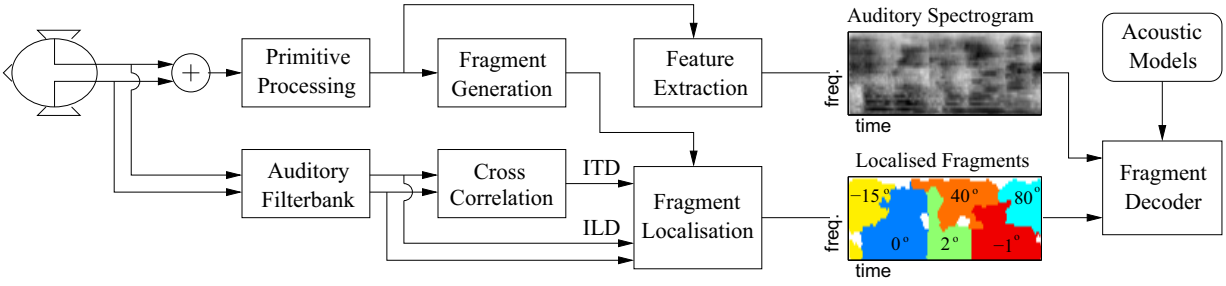

Figure 2: Overview of the proposed system. Localised spectro-temporal fragments are indicated using different colours.

tion can be estimated using a simple model of the noise, but this is not generally possible when the noise is itself highly unpredictable. The fragment decoding framework [4] acknowledges that the segmentation is not directly observed, and instead employs a segmentation model that represents a distribution of possible segmentations estimated from the noisy data. In particular this distribution only allows segmentations that are consistent with a set of local spectro-temporal sound source fragments.

Given the noisy observation $\mathbf{Y}$, the SFD framework couples the searches for the acoustic model state sequence $\mathbf{Q}$ and the segmentation hypothesis $\mathbf{S}$ that together are most probable:

$$\hat{\mathbf{Q}}, \hat{\mathbf{S}} = \arg\max_{\mathbf{Q},\mathbf{S}} P(\mathbf{Q}, \mathbf{S}|\mathbf{Y}) \quad (1)$$

$$= \arg\max_{\mathbf{Q},\mathbf{S}} P(\mathbf{Q}|\mathbf{S}, \mathbf{Y}) P(\mathbf{S}|\mathbf{Y}) \quad (2)$$

$P(\mathbf{Q}|\mathbf{S}, \mathbf{Y})$ is equivalent to missing data decoding given hypothesis $\mathbf{S}$, and $P(\mathbf{S}|\mathbf{Y})$ is the segmentation model. The previous studies employ a simple segmentation model which assigns equal probability to any foreground/background segmentation that can be constructed from a set of fragments. In this study $P(\mathbf{S}|\mathbf{Y})$ is inferred by binaural localisation cues.

3. Binaural cues for fragment decoding

This study presents a novel extention to the fragment decoding system by incorporating binaural information. An overview of the system is shown in Fig. 2. In summary, spectro-temporal fragments are first isolated from single channel mixtures based on periodicity. They are then localised based on localisation cues extracted from binaural recordings. This approximate location information is used to bias the fragment decoder towards selecting fragments that come from the same direction of the target source while rejecting the others.

3.1. Single channel fragment generation

Periodicity is among the most robust cues for auditory grouping and can also provide some resistance to reverberation [6]. It has been the major cue for fragment generation in previous fragment decoding systems (e.g. [7, 8]). The strategy for fragment generation is to exploit the distinctness and continuity of signal-level properties of the individual sound sources. Frequency channels dominated by the same periodic or quasi-periodic source will have a common fundamental frequency ($F0$), hence it can be used as evidence to label channels as belonging to the same fragment. Further, by tracking the $F0$ trajectory of sound sources it is possible to extend cross-frequency grouping through time. Ma et al. [7] discuss details of the $F0$-based grouping via the use of an autocorrelogram. Energy not accounted for by the $F0$-based fragments is segmented into disjoint 'inharmonic fragments' [7].

3.2. Binaural fragment localisation

Localisation estimates can be made by measuring the time and level difference of the signal arriving at the two ears, known as the interaural time difference (ITD) and the interaural level difference (ILD), respectively. If the *direction of origin* of the energy dominating each T-F element could be estimated, then this cue could be used to segment the representation. Unfortunately, binaural cues cannot be measured reliably within single frequency filter channels due to phase ambiguity and room reverberation [9]. Reliability can be increased, however, by integrating estimates over extended spectro-temporal regions [10].

ITD is estimated by computing a cross-correlation on the output of each auditory filter [11]. When estimating the location of a single source, the standard approach is to sum the cross-correlation functions across frequency – to form a so-called *summary cross-correlogram* – and then to find the delay of the largest peak. In [10] this idea is generalised so that the summary is computed by integrating the cross-correlation functions over a spectro-temporal fragment.

To address the problem of low frequency bands having very broad peaks, we skeletonise the cross-correlogram by replacing the largest peak in each channel by a Gaussian, instead of replacing all the local peaks as suggested in [12].

In a similar way to ITD estimation, ILD is computed using energy integrated over a fragment, before being converted into decibels (dB). As low frequencies tend to produce ambiguous level differences because diffraction reduces the effect of head shadow [13], only the frequency bands above 1600 Hz are used for computing fragment ILD.

3.3. Integrating binaural cues

We integrate binaural cues and acoustic models in a probabilistic framework via the segmentation model in (2). By assuming independence of fragments, the segmentation model can be approximated as :

$$P(\mathbf{S}|\mathbf{Y}) = \prod_{f \in \mathcal{F}_\mathbf{S}} P(f) \prod_{f \notin \mathcal{F}_\mathbf{S}} 1 - P(f) \quad (3)$$

where $\mathcal{F}_\mathbf{S}$ is the subset of fragments labelled as the foreground under hypothesis $\mathbf{S}$, and $P(f)$ is the probability of fragment f belonging to the target source. Once a fragment has been localised, its estimated location can be used to inform $P(f)$. This probability becomes smaller for fragments that do not come from the same direction of the target source, and larger if they do. More details will be given in Section 5.

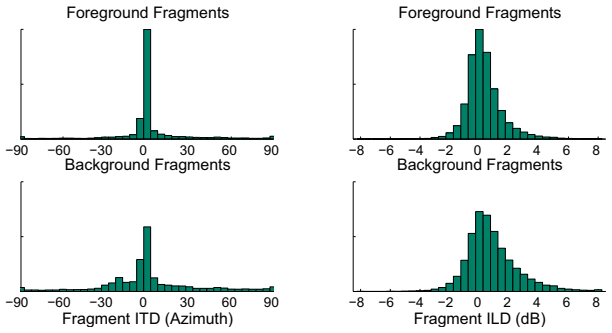

Figure 3: Histogram of ITD and ILD estimates for fragments in the foreground (top) or fragments in the background (bottom). Fragments are weighted by their area. SNR = 0 dB.

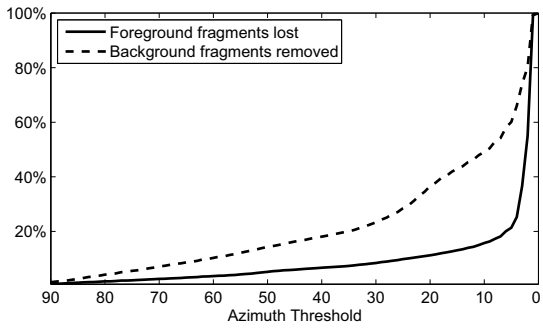

Figure 4: True rejection rate *vs.* false rejection rate of fragments using the absolute azimuth as a threshold.

4. Speech recognition task

The recognition system has been evaluated using the CHiME Challenge data set [2], sampled at 48 kHz. The task entails the recognition of Grid command utterances that have been mixed into binaural recordings made in a noisy domestic environment after convolution with carefully measured room impulse responses. The target speech is positioned directly in front of the manikin. The SNRs have been controlled by selecting temporal positions within the CHiME recordings that would result in the required SNR when the sources are mixed at their naturally occurring levels. Note, this means that *the noise backgrounds are necessarily different in each SNR condition*.

Our recognition systems are trained on the noise-free Grid corpus training set. We have assumed a matched-reverberation condition, i.e. the training data is convolved with a BRIR recorded at the same position as that used in the construction of the test set. The binaural training and test data is then reduced to a single channel by averaging the left and right ear signals. Feature extraction is then applied to the single channel signals.

5. Analysis and experiments

5.1. Fragment localisation

A fragment analysis was applied to the 0 dB test set. An 'oracle' foreground/background labelling for each fragment was determined with access to the premixed target speech and noise backgrounds. Fragments are labelled foreground if more than half of their elements have a local SNR above 0 dB[1]. Having labelled fragments as either 'more foreground' or 'more background' we can then look at the distribution of location estimates for each class and see whether it is discriminative.

ITD: Fig. 3 (left) shows the distribution of fragment azimuth estimates for the foreground and background fragment classes. For the target speech fragments there is a very large peak at 0 degrees and the vast majority of fragments are localised as being between -10 and +10 degrees. For the background there is still a peak at 0 degrees but now a significant number of fragments are estimated to originate from angles away from the centre. Clearly, originating from 0 degrees does not imply that the source is the target speaker but, logically, originating from a direction other than 0 degrees should be taken as evidence that the fragment is not part of the speech source.

Fig. 4 illustrates the potential for using azimuth estimates as a filter that rejects fragments from wide angles by assigning them to the background. The dashed curve shows the increasing proportion of noise fragments that would be correctly rejected as the threshold is decreased, while the solid curve shows the proportion of speech fragments that would be falsely rejected. With a 20 degree threshold around 40% of noise fragments can be rejected at a cost losing only 10% of speech fragments.

ILD: The ILD estimates, however, exhibit a very similar distribution for the foreground and background classes (Fig. 3 right). A possible explanation is that the noise fragments may not be 'pure', i.e. a fragment may contain energy from multiple sources. Further, even if fragments are pure, each T-F element will contain energy contributions from multiple sources so the *average* energy across a fragment may be similar for the left/right channels. This is less a problem for ITD estimation since only the largest peak is taken into account for each T-F element, i.e. ITD computation can accommodate the incoherence of the fragment, whereas the ILD computation does not.

ITD + ILD: Fig. 5 shows the 2D distribution of ITD/ILD pairs. It is clear that the foreground class shows a dense centre while the background class is more spread out. It is also noticeable that the background distribution is slightly tilted to the bottom-left, demonstrating the correlation between ITD and ILD.

5.2. Recognition experiments

Table 1 shows the keyword recognition accuracies for each ASR system tested. 'MFCC' represents a conventional baseline system trained using the 'standard' 13 cepstral coefficients plus deltas and accelerations, with cepstral mean normalisation applied. As might be expected, since little account is taken of the noise, performance degrades rapidly as SNR is reduced.

The SFD in the baseline single-channel configuration produces more robust performance. At -6 dB, 72 % of the tokens are recognised correctly compared to only 31 % for the MFCC system. To measure the ceiling performance we used the 'oracle' foreground/background labelling (Section 5.1) for each fragment. A decoding was then performed using the one segmentation formed by this explicit labelling. The result is shown in the row 'Oracle Lab'. As can be seen, the SFD baseline is 2-6% below this 'oracle' decoding across all SNRs.

The 'SFD+ITD' system employs the azimuth estimates to inform the probability $P(f)$ in the segmentation model (Sec-

[1] Ideally fragments would be 'pure' and hence all their points would carry the same label, but of course, the fragment analysis is imperfect

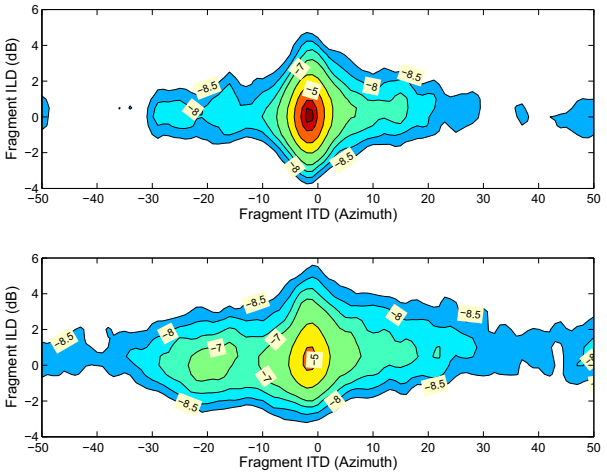

Figure 5: 2D histogram of ITD and ILD estimate pairs for fragments in the foreground (top) or fragments in the background (bottom). The histogram is smoothed with a 2D Gaussian kernel and converted into log probability distributions.

Table 1: Keyword recognition accuracy rates (%)

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
MFCC	31.08	36.75	49.08	64.00	73.83	83.08
SFD	72.33	72.25	78.83	82.00	87.17	87.92
Oracle Lab	76.00	78.00	83.42	87.25	89.17	90.67
SFD + ITD	**72.67**	**73.58**	**80.25**	**84.00**	**88.08**	**89.67**

tion 3.3). We set $P(f)$ to 0.55 for fragments inside an azimuth threshold, and $P(f) = 0.4$ for others, i.e. fragments coming from the front are slightly biased towards foreground whereas fragment from lateral angles are biased towards being labelled as background. These values were optimised in the 0 dB SNR condition and then fixed across all SNRs. The azimuth threshold was selected to be 20 degrees, which according to Fig. 4 rejects a high proportion of background for little loss of speech data. Since smaller fragments tend to have less reliable location estimates, for fragments with less than 8 T-F elements we set the $P(f)$ to 0.5, i.e. they are not biased towards either foreground or background.

The 'SFD+ITD' system produces improvement over the SFD baseline across all SNRs (the difference is significant at the 0.01 level for SNRs above 0 dB). By penalising fragments that do not come from the direction of the target source while favouring those that do, the fragment decoder is able to make use of a better segmentation model than the simple one which assigns equal probability to any foreground/background segmentation constructed from the fragments.

As suggested in Section 5.1, ILD does not provide much discrimination power, and our experiments show that it does not improve the recognition accuracy.

6. Discussion and conclusions

We have demonstrated a fragment-based approach for incorporating binaural localisation cues in a probabilistic framework for distant speech recognition. One important factor for the success of this approach is that the binaural cues are integrated over the spectro-temporal region defined by a source fragment. As demonstrated in the past, in reverberant and multisource environments binaural cues for individual time-frequency element can be too ambiguous to be useful [9, 10].

Our experiments show that the ITD cue is a more powerful cue than ILD for the particular task used in this study. Even so, there is still some scope for improving localisation estimates and employing a more sophisticated segmentation model. Future work is needed to fully exploit localisation cues within this framework.

If the quality of the fragments is the system's bottleneck then it may be possible to achieve more significant gains by using localisation cues *within the fragment generation* stage itself. Models which group T-F elements through time and frequency using the joint statistics of both pitch and location estimates need to be explored.

To conclude, this paper has presented a fragment based recognition system that addresses the problem of distant microphone speech recognition in reverberant multisource conditions. The system combines binaural localisation cues with periodicity cues and acoustic models in a probabilistic framework to simultaneously separate and recognise speech. The simple masking model on which it is based allows it to operate in a wide range of noise background by exploiting cues for separation rather than relying on details of the noise itself.

7. References

[1] M. Wöelfel and J. McDonough, *Distant speech recognition*. Wiley, 2009.

[2] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments," in *Proc. Interspeech'10*, 2010.

[3] A. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[4] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Commun.*, vol. 45, pp. 5–25, 2005.

[5] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and uncertain acoustic data," *Speech Commun.*, vol. 34, no. 3, pp. 267–285, 2001.

[6] C. Darwin and R. Hukin, "Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention," *J. Acoust. Soc. Am.*, vol. 108, no. 1, pp. 335–342, 2000.

[7] N. Ma, P. Green, J. Barker, and A. Coy, "Exploiting correlogram structure for robust speech recognition with multiple speech sources," *Speech Commun.*, vol. 49, no. 12, pp. 874–891, 2007.

[8] J. Barker, N. Ma, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Comput. Speech. Lang.*, vol. 24, no. 1, pp. 94–111, 2010.

[9] M. Stern, G. Brown, and D. Wang, "Binaural sound localization," in *Computational auditory scene analysis: principles, algorithms, and applications*, D. Wang and G. Brown, Eds. IEEE Press/Wiley-Interscience, 2008, ch. 5, pp. 147–186.

[10] H. Christensen, N. Ma, S. Wrigley, and J. Barker, "Integrating pitch and localisation cues at a speech fragment level," in *Proc. Interspeech'07*, Antwerp, 2007, pp. 2769–2772.

[11] B. M. Sayers and E. C. Cherry, "Mechanism of binaural fusion in the hearing of speech," *J. Acoust. Soc. Am.*, vol. 29, no. 9, pp. 973–987, 1957.

[12] K. Palomäki, G. Brown, and D. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Commun.*, vol. 43, pp. 361–378, 2004.

[13] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1997.