Recent advances in 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 recent advances in combining adaptive noise floor modelling and binaural localisation cues within this framework. The decoder is able to derive significant recognition performance benefits from both noise floor tracking and fragment location estimates. Using models trained on noise-free speech, the system achieves an average keyword recognition accuracy of 80.60% for the final test set on the PASCAL CHiME Challenge task.

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 is used in the PASCAL CHiME Speech Separation and Recognition Challenge [2]. The heterogeneous multi-source nature of the audio is readily apparent.

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 [4]. 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 [5].

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

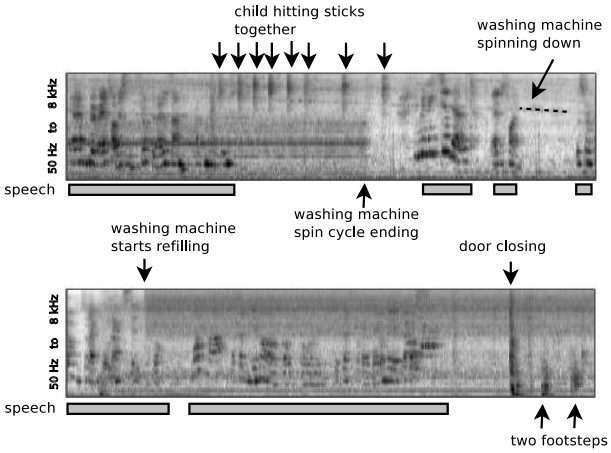


Figure 1: A time-frequency representation of a 20 second sample from the binaural CHiME domestic audio corpus used as noise background in the current study [3].

We recently proposed two extensions to the fragment decoding approach. The first extension employs an adaptive noise floor model to account for ambient, slowly varying noise floor [6]. The second one incorporates spatially motivated cues to bias the decoder towards accepting fragments that are believed to originate from a known target source location [7]. This paper presents our latest advances in combining the two extensions together with the fragment decoding approach.

An overview of the system is shown in Fig. 2. Section 2 reviews the basic fragment decoding framework. Adaptive noise floor modelling is presented in Section 3. Section 4 describes techniques of localising 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 5. Section 6 compares the recognition performance delivered by various ASR systems on the CHiME challenge. Section 7 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

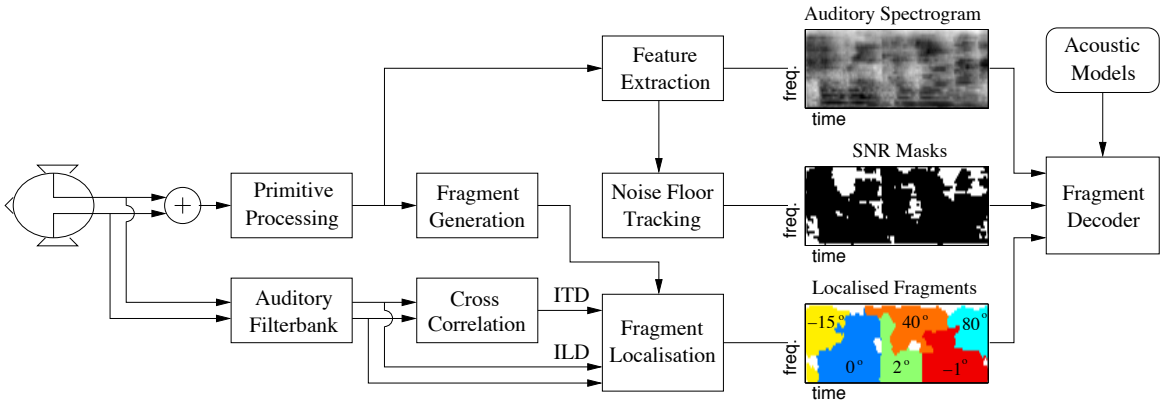


Figure 2: Overview of the proposed system. Localised spectro-temporal fragments are indicated using different colours.

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 [8].

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 segmentation 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 [5] 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.

2.1. Formulation

The fragment decoding framework is formalised as follows. Let $\mathbf{Y}$ be a sequence of noisy speech observations $\{\mathbf{y}_1, ..., \mathbf{y}_T\}$ where each $\mathbf{y}_t$ is a feature vector representing a spectral energy component at time t. The ASR task is to find the best word sequence given these observations, or equivalently to find the best underlying acoustic model state sequence $\mathbf{Q} = \{\mathbf{q}_1, ..., \mathbf{q}_T\}$:

$$\hat{\mathbf{Q}} = \underset{\mathbf{Q}}{\operatorname{argmax}} \; P(\mathbf{Q}|\mathbf{Y}) \tag{1}$$

The sequence of noise-free target speech vectors $\mathbf{X}$ and the foreground/background segmentation $\mathbf{S}$ are not directly observed but can be introduced by integrating over all possibilities,

$$\hat{\mathbf{Q}} = \underset{\mathbf{Q}}{\operatorname{argmax}} \sum_{\mathbf{S}} \left\{ \int_{\mathbf{X}} P(\mathbf{Q}, \mathbf{X}, \mathbf{S}|\mathbf{Y}) \mathrm{d}\mathbf{X} \right\} \tag{2}$$

Typically, the sum over $\mathbf{S}$ is intractable, so we instead select a single segmentation and state sequence that jointly maximise the integral,

$$\hat{\mathbf{Q}}, \hat{\mathbf{S}} = \underset{\mathbf{Q}, \mathbf{S}}{\operatorname{argmax}} \int_{\mathbf{X}} P(\mathbf{Q}, \mathbf{X}, \mathbf{S}|\mathbf{Y}) \mathrm{d}\mathbf{X} \tag{3}$$

$$= \underset{\mathbf{Q}, \mathbf{S}}{\operatorname{argmax}} \int_{\mathbf{X}} P(\mathbf{Q}|\mathbf{X}) P(\mathbf{X}|\mathbf{Y}, \mathbf{S}) P(\mathbf{S}|\mathbf{Y}) \mathrm{d}\mathbf{X} \tag{4}$$

The acoustic model represented by the integral in the above can be estimated by making a series of independence assumptions described in detail in [5]. A simple segmentation model, $P(\mathbf{S}|\mathbf{Y})$, assigns equal probability to any foreground/background segmentation that can be constructed from the set of fragments that have been identified by the front-end processing, i.e. the region covered by each of N fragments must be either allocated exclusively to the foreground or to the background – in this way 2^N segmentations can be generated. All other segmentations are assigned a probability of 0. The maximisation over state sequence $\mathbf{Q}$ and segmentation $\mathbf{S}$ can then be achieved via a Viterbi search over a lattice of segmentation and state sequence hypotheses as described in [5].

2.2. Fragment generation

The key to success in the fragment decoding system lies in the reliable identification of fragments of significant extent in frequency and/or time: the larger the fragments are, the more they constrain the segmentation model. Periodicity information is among the most robust cues for auditory grouping and it has been the major cue for fragment generation in previous fragment decoding systems (e.g. [9, 10]).

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.

The pitch-based grouping is implemented via the computation of an autocorrelogram [11, 12]. First, the signal is passed through a 32-channel Gammatone filterbank with centre frequency distributed between 50 Hz and 8000 Hz with equal spacing on an equivalent rectangular bandwidth (ERB) scale. The autocorrelogram is then formed from the short-time autocorrelation computed on the output of each Gammatone filter (using a 30 ms Hann window). For periodic sounds, the autocorrelogram exhibits symmetric tree-like structures whose stems are located on the delays that correspond to the pitch periods of sources in the acoustic mixture. These pitch-related structures are exploited to group spectral components at each time frame, from which local pitch estimates are computed. Simultaneous pitch tracks are formed by linking local pitch estimates across time,

and each pitch track is then used to recruit a spectro-temporal fragment [9]. Energy not accounted for by the pitch-based fragments is segmented into disjoint 'inharmonic fragments' using techniques also described in [9].

3. Adaptive Noise Floor Modelling

In many natural listening conditions the auditory scene can be approximately described as a slowly varying noise floor plus highly unpredictable acoustic 'events'. This work combines adaptive noise floor modelling and fragment decoding techniques to handle both the quasi-stationary and unpredictable components of the noise background. The adaptive noise floor model is used to estimate the degree to which energetic acoustic events are masked by the noise floor. A fragment decoding system then attempts to interpret the energetic regions that are not accounted for by the noise floor model. The combined technique will be termed adaptive noise floor speech fragment decoding (ANF-SDF).

3.1. Noise floor tracking

We employ an adaptive noise floor tracking algorithm [13, 6] similar to minimum tracking-based methods [14, 15, 16]. The tracker models a rolling buffer of noisy speech as a mixture of Gaussians using the expectation maximisation (EM) algorithm. The distribution that has the minimum mean value is used as the noise estimate. The mixture model is continuously updated with a half second increment, producing a fresh noise floor estimate for every half second.

3.2. Combing noise tracking and SFD

The output of noise floor tracking can be expressed as a spectro-temporal map holding local signal-to-noise ratio (SNR) estimates. Such local SNR maps have formed the basis of missing data mask estimation in many previous missing data ASR systems [17, 18, 19]. The local SNR estimate in dB is computed as:

$$SNR = 20\big(\log_{10}\big(10^{y} - 10^{\hat{n}}\big) - \hat{n}\big) \qquad (5)$$

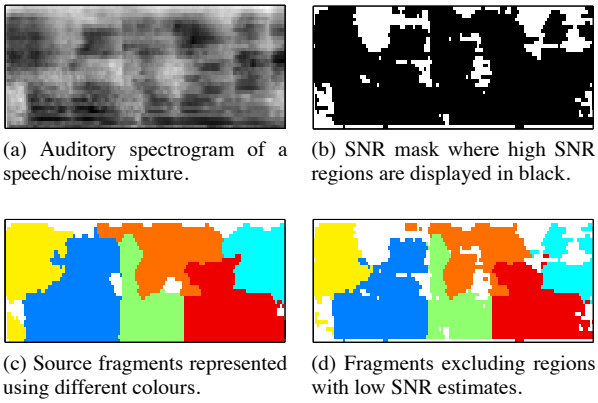
where y represents the log-compressed noisy observation and $\hat{n}$ is the log-compressed noise estimate. A soft missing data mask is obtained by applying a sigmoid function to the local SNR estimates:

$$\omega_{tf} = \frac{1}{1 + e^{-\alpha(SNR_{tf} - \beta)}} \qquad (6)$$

where α determines the slope of the sigmoid function and the centre β serves as the SNR threshold when computing soft masks. A higher SNR threshold will cause more T-F regions to be biased towards being interpreted as the noise background during decoding.

The T-F elements with values less than 0.5 are identified in the SNR-based soft mask. These low SNR regions are most likely to have originated from some noise sources, and they are interpreted as part of the background during fragment decoding, regardless of any segregation hypothesis. The fragments excluding these low SNR regions are treated by SFD as normal.

Fig. 3 illustrates this process. Fig. 3a is the auditory spectrogram of a speech/noise mixture. The missing data mask derived from local SNR estimates is shown in Fig. 3b, where regions with soft value > 0.5 are displayed in black. Source fragments are represented in Fig. 3c using different colours. Fig. 3d shows the fragments after the low SNR regions tracked by the adaptive noise floor model are forced into the background (represented in white). The process is akin to using the missing data mask in Fig. 3b to filter the fragments in Fig. 3c.



(a) Auditory spectrogram of a speech/noise mixture.

(b) SNR mask where high SNR regions are displayed in black.

(c) Source fragments represented using different colours.

(d) Fragments excluding regions with low SNR estimates.

Figure 3: Combining speech fragment decoding and adaptive noise floor modelling.

The ANF-SFD system differs from the standard SFD system because fragment decoding is only applied to regions that are not accounted for by the adaptive noise floor model, i.e. the noise floor is marked as being part of the background in all fragment labelled hypotheses. The standard SFD system would, by contrast, segment the regions dominated by the noise floor into fragments (often poorly because the noise floor tends to exhibit weak grouping cues) and may be prone to errors if any of these fragments happens to match the speech models.

4. Binaural fragment localisation

Binaural localisation cues are important for sound organisation [20, 21]. For example, if the target source was known to be directly ahead, then regions of energy coming from other directions could be labelled as part of the background. This section presents a binaural extension to the fragment decoding system.

4.1. Fragment-based localisation cues

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 [22]. Reliability can be increased, however, by integrating estimates over extended spectro-temporal regions. Indeed, [23] shows how, in a multisource scenario, fragments derived from periodicity cues can be localised with sufficient precision to benefit a simultaneous speaker tracking task.

ITD is estimated by computing a cross-correlation on the output of each auditory filter, based on Sayer and Cherry's implementation of the Jefress model [21]. 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 [23] this idea is generalised so that the summary is computed by integrating the cross-correlation functions over a spectro-temporal fragment.

A running cross-correlation is computed on the output

of each gammatone filter. At a given time step, the cross-correlation for each channel is computed iteratively with a decay window of 8 ms for temporal integration – long enough to produce robust correlations and short enough to approximately satisfy the assumption of stationarity over the correlation window. The cross correlation is normalised following the approach described in [24]. 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.

This work does not use ILD cues as they provide little discrimination power on the CHiME task [7].

4.2. Integrating binaural cues

We integrate binaural cues and acoustic models in a probabilistic framework via the segmentation model in (4). By assuming independence of fragments, the segmentation model can be approximated as :

$$P(\mathbf{S}|\mathbf{Y}) = \prod_{f \in \mathcal{F}_\mathbf{S}} P(f) \prod_{f \notin \mathcal{F}_\mathbf{S}} 1 - P(f) \qquad (7)$$

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 6.

5. Speech recognition task

The recognition system has been evaluated using the 2011 PASCAL CHiME Speech Separation and Recognition 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*.

All the recognisers employ word level HMMs with a topology that was standardised in the CHiME Challenge [2]. Our recognition systems are trained on the noise-free CHiME training set. The binaural training and test data is reduced to a single channel by averaging the left and right ear signals. Feature extraction is then applied to the single channel signals. A set of models is initially trained using all 17 000 utterances, then speaker dependent models are constructed by using 500 utterances from each speaker as adaptation data.

The baseline system employs 39-dimensional MFCC features (deltas and delta-deltas) and cepstral mean normalisation. SFD based systems require spectral features – missing features are localised in the spectral domain but not in the cepstral domain. The spectral features employed in the work were produced via a 32-channel Gammatone filterbank distributed in frequency between 50 Hz and 8000 Hz on the ERB scale, log-compressed and supplemented with deltas to form 64-dimensional feature vectors.

6. Analysis and experiments

We evaluate four fragment decoding systems. The first system, labelled as SFD, is the standard fragment decoding system. The second SFD system, ANF-SFD, combines SFD with adaptive noise floor modelling, as discussed in Section 3. ITD-SFD incorporates binaural localisation cues (ITD cues only) as discussed in Section 4. The final system, ANF-ITD-SFD, combines both extensions. All the SFD based systems employ fragments produced as discussed in Section 2.2.

Table 1 shows the keyword accuracies of baseline systems for the development set. As might be expected, the performance of the MFCC system degrades rapidly as SNR is reduced since little account is taken of the noise. The SFD in the baseline single-channel configuration produces more robust performance. For example, at 0 dB 77 % of the tokens are recognised correctly compared to only 49 % for the MFCC system.

Table 1: Keyword recognition accuracy rates (%) of baseline systems for the development set

	-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	70.25	72.58	77.25	82.17	85.75	86.58

6.1. Incorporating adaptive noise floor modelling

ANF-SFD is a fragment decoding system combined with adaptive noise floor modelling. The soft SNR-based masks were computed using (6): α was heuristically fixed to 0.1 and the SNR threshold β was tuned on the development set. Table 2 shows the keyword recognition accuracies with different values for β. The results of the standard SFD system is also included for comparison.

The combined ANF-SFD system exhibits improved performance over the standard SFD system at SNRs across various SNR conditions. The best results on the development set were obtained with the SNR threshold of -6 dB.

6.2. Incorporating localised fragment

The azimuthal angle of each source fragment was calculated from ITD estimated as described in Section 4.1. In the CHiME dataset the target speaker source is always located at 0 degree. Clearly, originating from 0 degrees does not imply that the source is the target speaker. However, originating from a direction other than 0 degrees should be logically taken as evidence that the fragment is not part of the speech source. Hence, the estimates could be used in the manner of a filter that rejects fragments from wider angles, or which reduces the probability of including these fragments as part of the foreground.

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 abscissa shows a rejection threshold. – i.e. fragments whose absolute angle is above this threshold are counted as part of 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.

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

Table 2: Keyword recognition accuracy rates (%) of the 'ANF-SFD' systems with various SNR threshold β for the development set. The results of the standard SFD system are also included for comparison.

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
SFD	70.25	72.58	77.25	82.17	85.75	86.58	79.10
$\beta = -12$ dB	70.92	74.17	78.25	82.83	87.58	87.17	80.15
$\beta = -9$ dB	71.33	74.67	78.33	82.42	87.75	87.67	80.36
$\beta = -6$ dB	**71.75**	**74.00**	**78.58**	**82.50**	**87.67**	**87.75**	**80.38**
$\beta = -3$ dB	71.75	73.50	78.17	82.42	87.33	87.25	80.07
$\beta = 0$ dB	71.42	72.83	77.00	82.58	87.33	87.42	79.76
$\beta = 3$ dB	70.92	71.83	76.50	82.58	87.25	87.67	79.46

Table 3: Keyword recognition accuracy rates (%) of the 'ITD-SFD' systems for the development set, with various $P(f)$ values for fragments inside (in) and outside an azimuth threshold (out), respectively. The results of the standard SFD system are also listed.

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
SFD	70.25	72.58	77.25	82.17	85.75	86.58	79.10
in = 0.5, out = 0.4	71.58	73.33	77.83	82.33	85.58	87.25	79.65
in = 0.5, out = 0.45	71.33	73.08	77.75	82.67	85.83	87.33	79.67
in = 0.52, out = 0.4	71.25	73.67	79.17	82.67	86.33	88.25	80.22
in = 0.52, out = 0.45	**71.75**	**73.33**	**78.75**	**83.17**	**86.25**	**88.25**	**80.25**
in = 0.55, out = 0.4	69.83	71.33	78.50	83.25	86.83	88.50	79.71
in = 0.55, out = 0.45	70.17	71.17	77.92	83.58	86.83	88.58	79.71

tion 4.2). The respective values of $P(f)$ for fragments inside an azimuth threshold and the remaining fragments were optimised on the development set, as shown in Table 3. The azimuth threshold was heuristically selected to be 18 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.

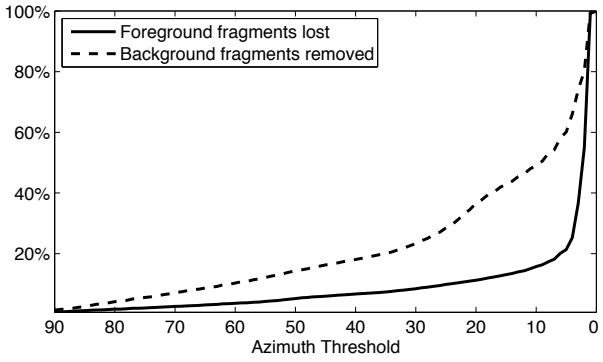


Figure 4: True rejection rate *vs*. false rejection rate of fragments using the absolute azimuth as a threshold.

The 'ITD-SFD' system produces improvement over the SFD baseline across all SNRs. 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. The best results were obtained with $P(f)$ set to 0.52 for fragments inside the azimuth threshold and 0.45 for the remaining fragments. With this setting fragments coming from the front are slightly biased towards foreground whereas fragment from lateral angles are biased towards being labelled as background.

6.3. Fusion of noise floor and localisation cues

The adaptive noise floor modelling and fragment localisation techniques can be combined together to further improve recognition accuracies. The noise model is first used to identify spectro-temporal regions that are likely to be part of the noise background. The remaining fragments are localised based on localisation cues extracted from binaural recordings at the fragment level.

Table 4 shows the best results of the combined system (ANF-ITD-SFD) for the development set. The parameters for the ANF-ITD-SFD system were optimised independently from the ANF-SFD system and the ITD-SFD system and they were then fixed for the final evaluation test set. The results for the final test set are shown in Table 5.

7. Conclusions

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 adaptive noise floor modelling and binaural localisation cues with acoustic models in a probabilistic framework to simultaneously separate and recognise speech. Essentially, the noise model is being allowed a first view of the data to estimate the degree to which energetic acoustic events are masked by the noise floor. A fragment decoding system then uses models of the target speech to interpret the energetic regions that are poorly predicted by the noise model. The binaural cues are integrated over each spectro-temporal fragment, which bias the decoder towards accepting fragments that are believed to originate from a known target source location.

In the current system the fragment separation, noise tracking and binaural fragment localisation are conducted indepen-

Table 4: Keyword recognition accuracy rates (%) of various ASR systems for the development set.

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
MFCC	31.08	36.75	49.08	64.00	73.83	83.08	56.33
SFD	70.25	72.58	77.25	82.17	85.75	86.58	79.10
ANF-SFD	71.75	74.00	78.58	82.50	87.67	87.75	80.38
ITD-SFD	71.75	73.33	78.75	83.17	86.25	88.25	80.25
ANF-ITD-SFD	**72.67**	**75.08**	**79.25**	**83.67**	**88.42**	**88.00**	**81.18**

Table 5: Keyword recognition accuracy rates (%) for the final test set.

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
MFCC	30.33	35.42	49.50	62.92	75.00	82.42	55.93
ANF-ITD-SFD	**72.25**	**73.17**	**81.75**	**84.08**	**85.50**	**86.83**	**80.60**

dently of each other. Options exist for closer coupling. For example, the ongoing noise floor estimate could be used to inform parameters of the pitch estimation and across frequency pitch grouping processes that are essential to the harmonic fragment generation. Working in the other direction, spectro-temporal regions that are clearly implicated in a fragment of an acoustic event, by pitch or location grouping cues, should not be contributing to the noise floor estimate.

8. References

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

[2] http://www.dcs.shef.ac.uk/spandh/chime/challenge.html.

[3] 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.

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

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

[6] N. Ma, J. Barker, H. Christensen, and P. Green, "Combining speech fragment decoding and adaptive noise floor modelling," *IEEE T. Audio. Speech.*, submitted.

[7] ——, "Binaural cues for fragment-based speech recognition in reverberant multisource environments," in *Proc. Interspeech'11*, submitted.

[8] 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.

[9] 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.

[10] 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.

[11] J. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, pp. 128–134, 1951.

[12] M. Slaney and R. Lyon, "A perceptual pitch detector," in *Proc. IEEE ICASSP'90*, Albequerque, 1990, pp. 357–360.

[13] N. Ma, J. Barker, H. Christensen, and P. Green, "Distant microphone speech recognition in a noisy indoor environment: combining soft missing data and speech fragment decoding," in *ISCA Workshop on Statistical And Perceptual Audition (SAPA'10)*, Makuhari, 2010.

[14] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE T. Speech. Audi. P.*, vol. 9, no. 5, pp. 504–512, 2001.

[15] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE T. Speech. Audi. P.*, vol. 11, no. 5, pp. 466–475, 2003.

[16] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Commun.*, vol. 48, no. 2, pp. 220 – 231, 2006.

[17] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP'00*, Beijing, 2000, pp. 373–376.

[18] P. Renevey and A. Drygajlo, "Detection of reliable features for speech recognition in noisy conditions using a statistical criterion," in *Proc. CRAC'01*, Aalborg, Denmark, 2001.

[19] C. Cerisara, S. Demange, and J. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Comput. Speech. Lang.*, vol. 21, pp. 443–457, 2007.

[20] C. Cherry, "Some experiments on the recognition of speech with one and two ears," *J. Acoust. Soc. Am.*, vol. 24, pp. 554–559, 1953.

[21] 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.

[22] 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.

[23] 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.

[24] C. Faller and J. Merimaa, "Sound localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, 2004.