A hearing-inspired approach for distant-microphone speech recognition in the presence of multiple sources

Ning Ma *, Jon Barker, Heidi Christensen, Phil Green

Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK

Received 11 January 2012; received in revised form 6 September 2012; accepted 9 September 2012
Available online 16 September 2012

Abstract

This paper addresses the problem of speech recognition in reverberant multisource noise conditions using distant binaural microphones. Our scheme employs a two-stage fragment decoding approach inspired by Bregman’s account of auditory scene analysis, in which innate primitive grouping ‘rules’ are balanced by the role of learnt schema-driven processes. First, the acoustic mixture is split into local time-frequency fragments of individual sound sources using signal-level primitive grouping cues. Second, statistical models are employed to select fragments belonging to the sound source of interest, and the hypothesis-driven stage simultaneously searches for the most probable speech/background segmentation 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. By integrating signal-level grouping cues with acoustic models of the target sound source in a probabilistic framework, the system is able to simultaneously separate and recognise the sound of interest from the mixture, and derive significant recognition performance benefits from different grouping cue estimates despite their inherent unreliability in noisy conditions. Finally, the paper will show that missing data imputation can be applied via fragment decoding to allow reconstruction of a clean spectrogram that can be further processed and used as input to conventional ASR systems. The best performing system achieves an average keyword recognition accuracy of 85.83% on the PASCAL CHiME Challenge task.

© 2012 Elsevier Ltd. All rights reserved.

Keywords: Distant-microphone speech recognition; Auditory scene analysis; Binaural localisation; Noise robustness

1. Introduction

Over many decades, research on automatic speech recognition (ASR) has made steady progress, and ASR technology is finally starting to become commonplace. In many applications, however, the expectation is that the user is employing a close-talking microphone – either embedded in a handheld communication device or attached to a wearable headset. For ASR technology to become truly ubiquitous, it needs to be freed from this constraint and designed to work reliably with distant microphones, allowing users to move freely and communicate spontaneously.

* This paper has been recommended for acceptance by Shrikanth Narayanan.

* Corresponding author. Present address: MRC Institute of Hearing Research, University Park, Nottingham NG7 2RD, UK.

Tel.: +44 115 9518508x371; fax: +44 115 9518503.

E-mail addresses: n.ma@ihr.mrc.ac.uk (N. Ma), j.barker@dcs.shef.ac.uk (J. Barker), h.christensen@dcs.shef.ac.uk (H. Christensen), p.green@dcs.shef.ac.uk (P. Green).
The scarcity of distant microphone ASR applications is not due to a lack of demand, but rather because recognition in these conditions is a difficult and largely unsolved problem (Baker et al., 2009; Wöllfle and McDonough, 2009). 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 factors 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. Most ‘everyday’ environments will contain an unknown number of sound sources whose activity level and location may be changing over time.

There exists an extremely diverse set of techniques for noise-robust speech recognition. For example, one of the most popular methods of reducing the effect of background noise is spectral subtraction (Boll, 1979; Martin, 2001; Cohen, 2003; Rangachari and Loizou, 2006), which assumes that it is possible to estimate the spectral characteristics of the noise. Such noise estimation can also be used to determine the noise masking pattern for missing data techniques (Cooke et al., 2001; Raj et al., 2004; Barker et al., 2001), measure feature uncertainty in uncertainty decoding (Droppo et al., 2002; Deng et al., 2005; Liao and Gales, 2008), or to adapt the speech model via a model combination technique (Varga and Moore, 1990; Gales and Young, 1996; Roweis, 2000; Frey et al., 2001). These techniques clearly depend on the quality of noise estimation and suffer in situations where an accurate noise model is difficult to estimate (e.g. when the interfering noise signal is unpredictable or has highly dynamic characteristics). When multiple microphone inputs are available (e.g. when using a microphone array), directional interferers can also be suppressed by incorporating spatial information via beamforming (Seltzer et al., 2004). These methods, however, often suffer from the negative effects of reverberation on the speech signal and cannot reliably separate the interferers if they are in close proximity to the speech source. Blind source separation (Takahashi et al., 2009) is also a popular method for separating a set of mixed signals with little information about the source signals, but it often underperforms when there are more sound sources than the number of sensors.

Human listeners, in contrast, are able to attend to individual components of complex acoustic mixtures, even when only presented with a single acoustic channel. This ability is called auditory scene analysis (ASA), and a popular account by Bregman (1990) describes the problem as a two-stage process in which innate primitive grouping ‘rules’ are balanced by the role of learnt schema-driven processes. This paper will review a probabilistic framework for auditory scenes analysis based on this account (Barker et al., 2005), and present recent advances which incorporate localisation cues and noise floor estimation within the framework. The hearing-inspired approach will be applied to the problem of distant microphone speech recognition in the context of the PASCAL CHiME Speech Separation and Recognition Challenge (Barker et al., submitted for publication).

An overview of the system is shown in Fig. 1, which models the ASA 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 (the primitive grouping stage). Second, a statistical back-end, through a process termed fragment decoding (Barker et al., 2005), selects source fragments based on the extent to which they match models of the target sound source.
(the schema-driven stage). The fragment decoding framework will be reviewed in Section 2. Two extensions to the fragment decoding approach are combined within the probabilistic framework. First, an adaptive noise floor model, presented in Section 3, is employed to accommodate slowly varying noise floor in the domestic noise background, and the remaining unpredictable acoustic ‘events’ (e.g. speech, human movement and mechanical sounds) are accounted for via source fragments (Ma et al., 2012). Second, spatially motivated cues, described in Section 4, are incorporated to bias the decoder towards accepting fragments that are believed to originate from a known target source location. The binaural localisation cues, such as interaural time difference (ITD) and interaural level difference (ILD), are integrated over spectro-temporal regions defined by source fragments. Such a process provides more reliable location estimates as reverberation tends to render instantaneous ITD/ILD estimates highly unreliable (Ma et al., 2011). This paper will show that the noise floor tracking component and the binaural localisation component are complementary and can be combined within the framework to further improve its robustness. Section 5 presents a data imputation technique that enables the fragment decoding framework to reconstruct the clean spectrogram that can then be processed further and used by a conventional ASR system. The reverberant binaural speech-in-noise data used for evaluation is described in Section 6, together with the recognition performance delivered by various ASR systems on the CHiME challenge. Finally, Section 7 concludes this paper and discusses future directions.

2. The fragment decoding framework

In the log-spectral domain, a noisy speech mixture can be segmented into (i) foreground regions dominated by the target speech which will have the statistics of noise-free speech models, and (ii) background regions where the speech is masked by noise energy, but in which the unobserved speech energy is known to have a value less than the observed masking energy. These two assumptions are supported by the well-known log-max approximation (Varga and Moore, 1990) – the log-spectral energy of the mixture is well approximated by the maximum of the individual log-spectra energies. Given statistical models of the noise-free speech and a foreground/background segmentation in the T–F domain, Cooke et al. (2001) show that the probability of any speech model sequence can be computed using missing data ASR techniques.

The difficulty with the missing data 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 speech fragment decoding (SFD) framework (Barker et al., 2005) 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: the spectro-temporal region dominated by a single sound fragment may be part of the target speech source (assigned to the foreground) or part of the background, but it cannot be split between the two. For example, Fig. 2 shows how two fragments identified from 80 T–F elements would lead to four permissible foreground/background segmentations.
Let $Y$ be a sequence of noisy speech observations $\{y_1, \ldots, y_t\}$ where each $y_t$ 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 $Q = \{q_1, \ldots, q_T\}$:

$$\hat{Q} = \arg\max_Q P(\{Q|Y\})$$ (1)

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

$$\hat{Q} = \arg\max_Q \sum_S \left\{ \int_X P(Q, X, S|Y)dX \right\}$$ (2)

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

$$\hat{Q}, \hat{S} = \arg\max_{Q, S} \int_X P(Q, X, S|Y)dX$$ (3)

$$= \arg\max_{Q, S} \int_X P(Q|X)P(X|Y, S)dX P(S|Y)$$ (4)

$$= \arg\max_{Q, S} \int_X P(X|Q) \frac{P(X|Y, S)}{P(X)} dXP(Q) P(S|Y)$$ (5)

The acoustic model represented by the integral in the above can be estimated by making a series of independence assumptions described in Barker et al. (2005). It should be pointed out that the segmentation model $P(S|Y)$ may use a different set of features (e.g. location estimates) extracted from the observed noisy signal than those used in the acoustic model. This is indicated in the graphical model representation of the framework, shown in Fig. 3, by using a different variable $Y'$ for the observation in the segmentation model. For the sake of simplicity this paper will still use $Y$ when discussing the segmentation model.

Under a foreground/background segmentation hypothesis $S$, the state likelihood can be evaluated using bounded marginalisation techniques (Cooke et al., 2001). The maximisation over both state sequence $Q$ and segmentation $S$ can be achieved via a Viterbi search over a lattice of segmentation and state sequence hypotheses (Barker et al., 2005).

2.2. Segmentation model

The segmentation model $P(S|Y)$ constrains the segmentation search space, similar to the way the language model constrains the state sequence search space. A simple segmentation model assigns equal probability to any foreground/background segmentation that can be constructed from the set of fragments that have been identified by front-end grouping 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 (see Fig. 2). All other segmentations are assigned a probability of zero.
The simple segmentation model assigns equal probability to any foreground/background segmentation that can be constructed from the set of fragments. Next we will discuss two additional components that can be combined to impose a better form for the segmentation model.

3. Adaptive noise floor tracking

In the CHiME corpus, the auditory scene can be approximately described as a slowly varying noise floor plus highly unpredictable acoustic events. The regions dominated by the noise floor tend to exhibit weak grouping cues and therefore it is often difficult to segment these regions into fragments. In order to handle both the quasi-stationary and unpredictable components of the noise background, we introduce an adaptive noise floor model to estimate the degree to which energetic acoustic events are masked by the noise floor. 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 labelling hypotheses. The combined technique will be termed adaptive noise floor speech fragment decoding (ANF-SDF).

3.1. Noise floor tracking

This work employs an adaptive noise floor tracking algorithm (Ma et al., 2012) similar to minimum tracking-based methods which are popular in the field of speech enhancement (Martin, 2001; Cohen, 2003; Rangachari and Loizou, 2006). Instead of tracking the minimum by averaging the previous noisy spectra over a finite window, the tracker adopted in this paper models the noisy spectra as a mixture of Gaussians and the mixture component that has the lowest RMS mean vector value is considered the noise floor estimate. Thus it avoids the smoothing issue commonly addressed in minimum statistics approaches. Such a GMM-based approach is often used for segmenting moving regions from the background in image processing (Stauffer and Grimson, 2000). It should be noted that although the GMM-based approach is adopted in this work, the system can also use other noise floor tracking techniques.

Let \( Y^w = \{y^1, \ldots, y^L\} \) represent a sequence of log-compressed noisy spectra in a window \( w \) of \( L \) frames. A GMM with diagonal covariance was fitted to the rolling window of noisy speech, using the expectation maximisation (EM) algorithm. Since adjacent spectral dimensions are correlated, a well-separated subset of frequency channels, \( \hat{Y}^w = \{\hat{y}^1, \ldots, \hat{y}^L\} \), was chosen from the full frequency band, so that features are nearly independent. In this work the subset of 5 frequency channels were equally spaced on the equivalent rectangular bandwidth (ERB) scale between 50 Hz and 8000 Hz.\(^1\) The maximum likelihood estimate (MLE) of the parameters for \( \hat{Y}^w \) is

\[
\hat{\theta} = \arg\max_\theta p(\hat{Y}^w | \theta)
\]

Let \( P(\hat{\theta}_k | \hat{y}) \) be the posterior probability of mixture component \( k \) for a sub-band feature vector \( \hat{y} \). The full-band mean of mixture component \( k \) for the noisy observation \( Y^w \) in window \( w \) can then be approximated as:

\[
\hat{\mu}_k^w = \frac{\sum_{t=1}^L P(\hat{\theta}_k | \hat{y}^t) y^t}{\sum_{t=1}^L P(\hat{\theta}_k | \hat{y}^t)}
\]

The noise floor estimate \( \hat{n} \) is assumed to be the full-band mixture component mean that has the lowest RMS value,

\[
\hat{n} = \hat{\mu}_\hat{k}^w
\]

where \( \hat{k} = \arg\min_k \left| \hat{\mu}_k^w \right| \).

3.2. Combining noise floor model and fragment decoding

The output of noise floor estimation can be represented as a spectro-temporal map holding local signal-to-noise ratio (SNR) estimates. Such a local SNR map have formed the basis of missing data mask estimation in many previous

---

\(^1\) To be more precise, their characteristic frequencies are located at 118 Hz, 439 Hz, 1057 Hz, 2247 Hz, and 4538 Hz respectively.
missing data ASR systems (Barker et al., 2000, e.g.; Renevey and Drygajlo, 2001; Cerisara et al., 2007). The SNR map, together with the set of fragments that have been identified, can impose a new form for the segmentation model $P(S|Y)$ in Eq. (5). First, the spectro-temporal regions whose SNR estimates are less than a threshold are identified. These low SNR regions are most likely to have originated from some slowly varying noise. They are then excluded from all the fragments and are interpreted as part of the background during fragment decoding. A higher SNR threshold will cause more regions to be interpreted as the noise background.

Fig. 4 illustrates this process. Fig. 4a is the auditory spectrogram of a speech/noise mixture. The low SNR regions tracked by the noise floor are shown as white in Fig. 4b whereas the regions with high SNR estimates are shown in black. Source fragments are represented in Fig. 4c using different colours. Fig. 4d 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 an SNR mask shown in Fig. 4b to mask the fragments in Fig. 4c. The segmentation model will interpret the low SNR regions as part of background during decoding and assign equal probability to any foreground/background segmentation that can be constructed from the remaining fragments.

The low SNR regions are not totally discarded during decoding. Instead the decoder employs soft decisions via the use of a soft missing data mask, each value of which is in the range of $[0,1]$ expresses a degree of confidence of being part of foreground, similar to the soft missing data decoding technique (Barker et al., 2000). The soft value will be used by the decoder as the prior probability of the T–F element being either foreground or background. For example, a value close to 0 indicates it is very likely to be part of the background, a value close to 1 means it is very likely to be foreground, and 0.5 would mean it will be considered equally likely to be either foreground or background.

In (Ma et al., 2012) the SNR map was converted to a soft mask by compressing the SNR estimates via a sigmoid function. The sigmoid function centre $\beta$ serves as the SNR threshold – any T–F elements whose SNR estimate are less than $\beta$ will have a soft mask value less than 0.5, thus they will be biased toward being interpreted as part of the noise background. It is also possible to statistically learn $\beta$ from data. However, since the SNR estimates may be erroneous in the first place, the soft mask values computed from these estimates become less effective. To reduce propagation of SNR estimation errors, a constant probability value is adopted for all the T–F elements whose SNR estimate are less than an SNR threshold $\beta$. Such a parameter can be empirically optimised using a development dataset and is more effective in reducing word errors than a sigmoid function.
4. Binaural localisation cues

The segmentation model discussed in Section 2.2 considers a uniform distribution for any foreground/background segmentation that can be constructed from a set of fragments. This simple model makes no assumption about the segmentations but often there is information available about which segmentation is more likely than others. For example, some fragments are more speech-like and segmentations that assign them to the speech foreground should be favoured over those that assign them to the noise background (Ma and Green, 2008); segmentations that include fragments with incompatible pitch information are less likely to be correct; localisation cues are also important for sound organisation (Cherry, 1953) and can be used to inform the segmentation model: if the direction of the target source is known, then regions of energy coming from other directions could be labelled as part of the background. Since the CHiME challenge provides binaural recordings, this section will present 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, it is difficult to measure binaural cues reliably within single frequency filter channels due to phase ambiguity and room reverberation (Stern et al., 2008). Reliability can be increased, however, by integrating estimates over extended spectro-temporal regions. Indeed, Christensen et al. (2007) 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 the Sayers and Cherry (1957) implementation of the Jefress model. At each time t and filterbank frequency channel f, the cross-correlation of the filterbank left ear output gl(t, f) and the right ear output gr(t, f) can be computed iteratively with a decay window for temporal integration

$$CC(t, f, \tau) = gl(t, f)gr(t - \tau, f) + e^{-t/\lambda}CC(t - 1, f, \tau)$$

where the exponential term serves as an exponential decay and the decay window \(\lambda\) was set to 8 ms which was found to be a good trade off – long enough to produce robust correlations and short enough to satisfy the assumption of stationarity over the correlation window (Christensen et al., 2009). The cross-correlation is then normalised by autocorrelation functions of the left and right ear signals respectively (Faller and Merimaa, 2004).

The standard approach to estimating the location of a single source is to sum the cross-correlation functions across frequency to form summary cross-correlogram and then to find the delay of the largest peak. With multiple sources that may overlap in time, the summary cross-correlogram would exhibit ambiguous peaks (Fig. 5, left). Christensen et al.
(2007) therefore generalised this idea so that the summary is computed by integrating the cross-correlation functions over a spectro-temporal fragment $F$,

$$\text{ITD} = \arg\max_\tau \sum_{t,f \in F} \text{CC}(t, f, \tau)$$

(10)

Since a fragment is dominated by energy from a single source, the summary cross-correlogram will exhibit a large peak indicating the source location$^2$ (Fig. 5, right).

To address the problem of low frequency bands having very broad peaks, the cross-correlogram was skeletonised by replacing the largest peak in each channel by a Gaussian before the integration over an entire fragment. This is akin to computing a weighted histogram with variable bin widths. Replacing only the largest peak was found more effective than replacing all the peaks as reported in (Palomäki et al., 2004).

This work does not use interaural level difference cues as Ma et al. (2011) suggest that they provide little discrimination on the CHiME task.

4.2. Combining binaural cues and fragment decoding

We integrate binaural cues and acoustic models in a probabilistic framework via the segmentation model $P(S|Y)$ in Eq. (5). By assuming independence of fragments, the segmentation model can be approximated as:

$$P(S|Y) = \prod_{F \in \mathbb{F}^S} P(F) \prod_{F \notin \mathbb{F}^S} 1 - P(F)$$

(11)

where $\mathbb{F}^S$ is the subset of fragments labelled as the foreground under hypothesis $S$, 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 lower for fragments that do not come from the same direction as the target source, and higher if they do.

4.3. Combining binaural cues and adaptive noise floor

The fragment localisation techniques can be combined with the adaptive noise floor modelling component described in Section 3 to further improve the segmentation model. First, the noise floor model can be used to identify spectro-temporal regions that are likely to be part of the noise background. The ambient, slowly varying noise captured by the noise floor tends to exhibit weak localisation cues and fragments are therefore localised after removing regions dominated by the noise floor. Second, the remaining fragments are localised based on the ITD cue extracted from binaural recordings at the fragment level, as discussed in Section 4. These fragments are ‘purer’ as potential noise regions have been removed from them, therefore their location estimates can be more reliable. This approximate location information is used to bias the fragment decoder towards selecting the fragments that come from the same direction of the target source while rejecting the rest.

5. Missing data imputation by fragment decoding

The fragment decoding technique is constrained to work with spectral features whereas the current state-of-art speech recognition frameworks work with cepstral features. One popular approach to treating missing data while exploiting the strength of more-orthogonalised features is by imputation which replaces noise-corrupted regions with estimates of the clean speech signal in the spectral domain (Cooke et al., 2001; Raj et al., 2004; Gemmeke et al., 2010; Segbroeck and Hamme, 2011; Kim and Stern, 2011). The restored full-band spectral features can then be processed to generate cepstral features for processing by a conventional ASR system, which may use different acoustic models or system architectures.

A widely used imputation method exploits the distribution of unreliable features conditioned on the reliable features to estimate the masked clean speech (Cooke et al., 2001; Raj et al., 2004). Let $x$ and $y$ represent feature vectors of

$^2$ It assumes the source remains stationary in space over the duration of the fragment.
Fig. 6. Imputed spectrogram based on the output of fragment decoding. The black colour in the lower-left panel represents the reliable speech regions identified by fragment decoding. It is clear that in the imputed spectrogram the noise energy (e.g. at 0–0.2 s) has been largely removed.

log-spectral energy of the underlying clean speech and the observed noisy mixture, respectively. If the clean speech is modelled by a Gaussian mixture model, the conditional distribution of unreliable features can be obtained by marginalising over each Gaussian component. Cooke et al. (2001) showed that imputation of unreliable features can be approximated as the expected value of such a distribution, by a weighted sum of the mean of each mixture component,

\[
\hat{x}_u = \sum_{k=1}^{M} P(k|x_r, y_u, q) \int_{-\infty}^{y_u} p(x_u|k, q) dx_u
\]

where \(y_u\) is the unreliable regions of the noisy spectra, \(q\) represents the Gaussian mixture, \(k\) is the mixture component, and \(M\) is the number of mixture components. The weights can be computed using the bounded marginalisation technique (Cooke et al., 2001). When the GMM employs a diagonal covariance matrix, the expectation is effectively the mean value of the mixture component \(k\) at the index of \(u\), but bounded by the observation.

In (Cooke et al., 2001) each HMM state in the acoustic models is represented by a GMM and therefore produces different imputed values for the same frame. Decoding is then based on these state-dependent feature vectors. In the cluster-based imputation proposed by Raj et al. (2004) and used by many others (Faubel et al., 2009, e.g.), a single GMM with a large number of Gaussian components is employed to impute the unreliable features and no temporal information is exploited. Decoding is based on cepstral features transformed from the reconstructed features. However, all these methods assume that a missing data mask is readily available, in which the identified reliable speech regions are labelled. In fact, identifying the reliable speech regions is one of the main challenges for the imputation methods.

As discussed in Section 2, the fragment decoding framework simultaneously searches for the best word sequence as well as the best foreground/background segmentation. The winning segmentation output resembles a missing data mask, as illustrated Fig. 7d. Such a mask is exploited in this work to reconstruct full-band features for further data transformation in order to take advantage of orthogonalised features. At each frame, the unreliable features are identified according to the winning segmentation. The corresponding state (represented by a GMM) in the winning state path is then used to impute the unreliable features in each frame based on the technique discussed above. An example of the imputed spectrogram based on the output of fragment decoding is shown in Fig. 6. The imputed spectrogram (middle) closely resembles the clean speech spectrogram (right) and the noise energy (e.g. at 0–0.2 s) has been largely removed.

6. Experiments and results

The systems were evaluated in the context of the PASCAL CHiME Speech Separation and Recognition Challenge (Barker et al., submitted for publication). The speech utterances obey a strict grammar constructed using a 51-word vocabulary. The ASR task is to identify two keywords, a letter-digit grid reference, that occur in every utterance. The average recognition accuracy for the two keywords is used to report the keyword accuracy. The speech materials were
6.1. Fragment generation

Fragments can be identified by exploiting the distinctness and continuity of signal-level properties of the individual sound sources. 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 (Ma et al., 2007; Barker et al., 2010). In this work a similar technique was employed but extra care was taken to handle reverberation.

The periodicity-based grouping was implemented via the computation of an autocorrelogram (Licklider, 1951; Slaney and Lyon, 1990). The autocorrelogram was formed from the short-time autocorrelation computed on the output of each Gammatone filter using a 50 ms Hann window. The longer window (30 ms used in previous studies) appeared to enhance the periodicity pattern in autocorrelograms hindered by reverberation. 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 were exploited to label channels as belonging to the same spectral group. Further, local pitch estimates were computed from the local spectral groups and by tracking the \( F_0 \) trajectory of sound sources it is possible to extend cross-frequency grouping through time. Energy not accounted for by the pitch-based fragments is segmented into disjoint ‘inharmonic fragments’ using the watershed segmentation algorithm described in (Ma et al., 2007).

6.2. Baseline experiments

The binaural data were reduced to a single channel by averaging the left and right ear signals for feature extraction. The baseline system employed 13-dimensional mel-frequency cepstral coefficients (MFCC) plus delta and delta-delta features, and with cepstral mean normalisation applied. 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 (Glasberg and Moore, 1990), log-compressed and supplemented with deltas to form 64-dimensional feature vectors.

Speaker-dependent word-level hidden Markov models (HMMs) were used in all the ASR systems with a left-to-right topology that was standardised by the CHiME Challenge. The number of states for each word was based on 2 states per phoneme. Each state employed seven-component Gaussian mixtures with diagonal-covariance. The recognition models were trained on the noise-free CHiME training set. 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.

Table 1 compares the keyword accuracies of the MFCC baseline system and the standard SFD system 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. The significance level p-value is less than 0.001 according to the McNemar test (Gillick and Cox, 1989) on keyword-pair errors.

Among the errors made by the SFD system, the most confusable pairs of words are letters that share a common vowel, e.g. $b|d|e|v$ are only recognised correctly around 30% at the $-3$ dB SNR. This is partly because the fragments corresponding to unvoiced energy fail to be identified during the primitive grouping stage, which primarily uses the periodicity cue. Further, the less-coherent unvoiced fragments tend to be confused with noise. The unvoiced energy is also difficult to group in noisy conditions for human listeners, who often exploit different grouping cues (e.g. location cues) and strong language context in noisy conditions.

An example of a typical fragment decoding is shown in Fig. 7. Fig. 7a shows an auditory spectrogram of a noisy utterance from the CHiME corpus at 0 dB SNR. The background in this example consists of mainly child speech. The set of fragments identified by harmonicity analysis (Section 6.1) is shown in Fig. 7b. Fig. 7c shows the ‘oracle’ foreground/background segmentation, obtained with access to premixed signals. The most likely segmentation selected by the fragment decoder is displayed in Fig. 7d, which closely resembles the oracle segmentation. The event at 1.2–1.6 s has been correctly identified as part of the background. The target speech is recognised correctly in this case.

6.3. Incorporation of adaptive noise floor modelling

The noise floor tracking model in this work employed two Gaussian components. The length of the rolling window was 5 s and the GMM was continuously updated with a half second increment. As discussed in Section 3.2 the ANF-SFD system employed a soft missing data mask for T–F regions whose SNR estimates are below a threshold $\beta$. The soft mask value for these regions was optimised on the development set. The best results were obtained with a soft mask value of 0.2 and the system was not sensitive to values in the range of $[0.1,0.3]$.

Table 2 compares the keyword recognition accuracy rates with SNR thresholds $\beta$ ranging from $-9$ dB to 3 dB, while fixing the soft mask value for low SNR regions to 0.2. The standard SFD results are also given for comparison (equivalent to having $-\infty$ as the SNR threshold). The ANF-SFD system exhibits significant improvement over the standard SFD system with all the SNR thresholds ($p$-value < 0.001). The best results on the development set are obtained with the SNR threshold of 0 dB, but the system does not appear sensitive to the parameter.

Further analysis of the results shows that the ANF-SFD system still mostly confuses letters such as $b$, $d$ and $e$, and it makes more recognition errors for letter $d$ than the standard SFD system. However, it reduces word error rates over
Table 2
Keyword recognition accuracy (%) of the ‘ANF-SFD’ systems with various SNR thresholds $\beta$ for the development set. The standard SFD results are given for comparison.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$-6,\text{dB}$</th>
<th>$-3,\text{dB}$</th>
<th>$0,\text{dB}$</th>
<th>$3,\text{dB}$</th>
<th>$6,\text{dB}$</th>
<th>$9,\text{dB}$</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rightarrow \infty$ (SFD)</td>
<td>70.25</td>
<td>72.92</td>
<td>79.33</td>
<td>82.67</td>
<td>86.67</td>
<td>88.08</td>
<td>79.99</td>
</tr>
<tr>
<td>$\beta = -9,\text{dB}$</td>
<td>72.42</td>
<td>76.08</td>
<td>80.08</td>
<td>84.00</td>
<td>88.25</td>
<td>88.25</td>
<td>81.51</td>
</tr>
<tr>
<td>$\beta = -6,\text{dB}$</td>
<td>72.67</td>
<td>75.33</td>
<td>79.83</td>
<td>84.08</td>
<td>88.17</td>
<td>88.42</td>
<td>81.42</td>
</tr>
<tr>
<td>$\beta = -3,\text{dB}$</td>
<td>72.42</td>
<td>75.50</td>
<td>80.00</td>
<td>84.33</td>
<td>88.42</td>
<td>88.00</td>
<td>81.45</td>
</tr>
<tr>
<td>$\beta = 0,\text{dB}$</td>
<td>71.75</td>
<td>75.83</td>
<td>80.42</td>
<td>85.00</td>
<td>87.83</td>
<td>88.75</td>
<td>81.60</td>
</tr>
<tr>
<td>$\beta = 3,\text{dB}$</td>
<td>72.08</td>
<td>76.17</td>
<td>80.33</td>
<td>84.00</td>
<td>87.92</td>
<td>88.33</td>
<td>81.47</td>
</tr>
</tbody>
</table>

the SFD system mainly for letters such as $e$, $f$, $g$, $i$, $q$ and $y$. The relative improvement for these letters is around 20% at the $-3\,\text{dB}$ SNR.

6.4. Incorporation of localisation cues

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 $0^\circ$. Clearly, originating from $0^\circ$ does not imply that the source is the target speaker. However, originating from a direction other than $0^\circ$ 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.

A fragment analysis was first 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. Fig. 8 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.

The ITD-SFD system employs the ITD azimuth estimates to inform the probability $P(F)$ for the segmentation model in Eq. (11). A value larger than 0.5 will cause the fragment to be biased towards the foreground (FG) during decoding, while a value smaller than 0.5 means the fragment will be biased towards the background (BG). $P(F)$ can be inferred as a classification problem for fragments (FG class vs. BG class) using a Bayesian approach from the probability distributions of $p(F|FG)$ and $p(F|BG)$. Fragment azimuths were identified for each fragment of a set of mixed training utterances. The label of each fragment (foreground or background) was produced by comparing the overlapping with the oracle mask (produced from pre-mixed signals). The distributions $p(F|FG)$ and $p(F|BG)$ can either

![Fig. 8](image-url) Proportion of speech fragments that would be falsely rejected as background and background fragments that would be truly rejected if all fragments whose absolute azimuth is greater than a given threshold are treated as part of the background.
be approximated by histograms obtained from the training data, or estimated by fitting GMMs to the data. However, both methods produced very poor classification results. The reason for the poor results is believed to be two-fold. First, as the a priori segmentations of different noise sources are not available (overlapping noises were recorded as a mixture), the background fragments may often contain energy from multiple noises as well as reverberations. As a result the azimuth estimates of background fragments tend to be averaged to be at 0° (the distribution peaks at 0° azimuth), reducing discrimination with the foreground distribution which is also peaked at 0°. Second, due to the existence of reverberation and multiple overlapping sources in this task, the azimuth estimates are often less reliable, especially for smaller fragments. The modelling errors therefore outweighed any benefit a statistical modelling technique may introduce.

It was found in this study that the key to achieving overall lower ASR word error rates is to soften the binary decision of assigning fragments to either foreground or background, and in this task a determinist approach works better. Lower word error rates can be achieved by directly optimising respective $P(F)$ values for fragments located inside an azimuth threshold and for those located outside the threshold using the development set. Table 3 compares the keyword recognition accuracy rates of the ITD-SFD system for the development set with various $P(F)$ values for fragments located inside and outside a azimuth threshold, respectively. For small fragments with less than 8 T–F elements, $P(F)$ is set to 0.5, i.e. they are not biased towards either foreground or background, since small fragments tend to have less reliable location estimates. The azimuth threshold was empirically fixed to be 18° after optimisation using the development set.

The standard SFD system is equivalent to assigning $P(F) = 0.5$ to all the fragments. It can be seen that the ITD-SFD system produces improvement over the SFD baseline across all SNRs ($p$-value < 0.001). 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 are obtained with $P(F)$ set to 0.53 for fragments inside the azimuth threshold and 0.45 for the others. However, Table 3 shows that the system is not very sensitive to these parameters in the range listed. Result analysis indicates that the ITD-SFD system makes the most errors for letters b and e, but improves over the standard SFD system for letters d, f, h, n and y. This pattern is complementary to that of exploiting noise floor modelling, which suggests that it may be possible to combine the two techniques within the fragment decoding framework.

### 6.5. Fusion of noise floor modelling and localisation cues

The adaptive noise floor modelling and fragment localisation techniques improve the standard SFD system in different ways and the two complementary techniques can be combined together to further improve its recognition accuracy. Tables 2 and 3 suggest that the performance of both techniques is not sensitive to the parameters tested. Therefore the best performing parameters for the respective techniques are chosen independently for the combined ANF-ITD-SFD system. Table 4 shows the results of the combined system for the development set, together with other tested systems. The combined system improves over all the SFD systems across various SNR conditions ($p$-value <
Table 4
Keyword recognition accuracy rates (%) of various ASR systems for the development set.

<table>
<thead>
<tr>
<th></th>
<th>−6 dB</th>
<th>−3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>31.08</td>
<td>36.75</td>
<td>49.08</td>
<td>64.00</td>
<td>73.83</td>
<td>83.08</td>
<td>56.33</td>
</tr>
<tr>
<td>SFD</td>
<td>70.25</td>
<td>72.92</td>
<td>79.33</td>
<td>82.67</td>
<td>86.67</td>
<td>88.08</td>
<td>79.99</td>
</tr>
<tr>
<td>ANF-SFD</td>
<td>71.75</td>
<td>75.83</td>
<td>80.42</td>
<td>85.00</td>
<td>87.83</td>
<td>89.75</td>
<td>81.60</td>
</tr>
<tr>
<td>+Imputation</td>
<td>73.00</td>
<td>77.25</td>
<td>82.00</td>
<td>85.75</td>
<td>89.08</td>
<td>90.17</td>
<td>82.88</td>
</tr>
<tr>
<td>ITD-SFD</td>
<td>72.25</td>
<td>75.08</td>
<td>81.00</td>
<td>84.92</td>
<td>87.92</td>
<td>89.08</td>
<td>81.71</td>
</tr>
<tr>
<td>+Imputation</td>
<td>72.83</td>
<td>76.42</td>
<td>82.58</td>
<td>87.17</td>
<td>90.08</td>
<td>90.75</td>
<td>83.31</td>
</tr>
<tr>
<td>ANF-ITD-SFD</td>
<td>74.67</td>
<td>77.33</td>
<td>82.17</td>
<td>86.17</td>
<td>88.83</td>
<td>89.58</td>
<td>83.17</td>
</tr>
<tr>
<td>+Imputation</td>
<td>75.75</td>
<td>79.33</td>
<td>83.83</td>
<td>87.00</td>
<td>90.58</td>
<td>91.33</td>
<td>84.64</td>
</tr>
</tbody>
</table>

0.001). In particular, most error reductions over the standard SFD system occur for letters e, f, g, h, i, s and y. This improvement is the result of a combined effort. Table 5 summarises recognition keywords (letters only) for which each system improves or decreases the recognition accuracy the most (>15%) over the standard SFD system. It can be seen that the ANF-ITD-SFD system combines the strength of both ANF-SFD and ITD-SFD systems. More importantly, the extra constraints brought by the noise floor tracking and binaural localisation components contributed in the reduction of letter confusion observed in previous systems.

Table 5
Recognition keywords (letters only) for which each system improves or decreases the recognition accuracy the most (>15%) over the standard SFD system for the development set.

<table>
<thead>
<tr>
<th></th>
<th>Improved words</th>
<th>Degraded words</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANF-SFD</td>
<td>e, f, g, i, q, y</td>
<td>d</td>
</tr>
<tr>
<td>ITD-SFD</td>
<td>d, f, h, n, y</td>
<td>b</td>
</tr>
<tr>
<td>ANF-ITD-SFD</td>
<td>e, f, g, h, i, s, y</td>
<td>None</td>
</tr>
</tbody>
</table>

6.6. Missing data imputation

The output from each of the 3 proposed SFD systems was used to impute the missing data as described in Section 5. Note that once the full-band spectral features are restored, they can be converted to cepstral features for processing by a conventional ASR system. In this work the restored features were orthogonalised by applying the DCT transformation to produce 13-dimensional MFCC-like cepstral features, which were further processed with mean normalisation and supplemented with delta and delta-delta features. Multicondition training was also employed as it has been reported by the majority of the systems participating in the CHiME Challenge as a technique that produced large improvement (Barker et al., submitted for publication). The multicondition trained models were obtained by retraining the word-level acoustic models using the noise-free training data mixed with CHiME noise at SNRs ranging from −6 dB to 18 dB, as well as the noise-free dataset.

Although performance of the imputation system is limited by the quality of the winning state sequence and the winning segmentation from SFD, it is believed the advantage of having more-orthogonalised features and processing such as mean normalisation and multicondition training would bring extra improvement. This is clearly evident from Table 4, which lists the imputation results from each SFD system for the development set. The results for the final test set are given in Table 6. The parameters for these systems were fixed to those optimised for the development set.

The imputation systems significantly outperformed each respective SFD system based on which the missing feature were restored (p-value < 0.001). The improvements occurred in all the SNR conditions, especially at the high SNR end where the noise was quieter and more stationary. Comparing the imputation systems tested, it is clear that the performance trend is in general agreement with that of the SFD base systems, and the best results were obtained by imputation based on output of the ANF-ITD-SFD system. A better SFD base system would produce a more correct HMM state sequence as well as better segmentation, therefore the restored features would approximate the underlying clean speech features better.
Table 6
Keyword recognition accuracy rates (%) for the final test set.

<table>
<thead>
<tr>
<th>Method</th>
<th>−6 dB</th>
<th>−3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>30.33</td>
<td>35.42</td>
<td>49.50</td>
<td>62.92</td>
<td>75.00</td>
<td>82.42</td>
<td>55.93</td>
</tr>
<tr>
<td>SFD</td>
<td>71.75</td>
<td>72.75</td>
<td>78.75</td>
<td>82.83</td>
<td>85.08</td>
<td>87.25</td>
<td>79.74</td>
</tr>
<tr>
<td>ANF-SFD</td>
<td>74.17</td>
<td>76.33</td>
<td>81.25</td>
<td>85.00</td>
<td>86.92</td>
<td>87.00</td>
<td>81.78</td>
</tr>
<tr>
<td>+Imputation</td>
<td>75.75</td>
<td>78.33</td>
<td>83.83</td>
<td>87.08</td>
<td>88.83</td>
<td>90.33</td>
<td>84.03</td>
</tr>
<tr>
<td>ITD-SFD</td>
<td>74.00</td>
<td>75.83</td>
<td>81.33</td>
<td>85.33</td>
<td>87.50</td>
<td>88.67</td>
<td>82.11</td>
</tr>
<tr>
<td>+Imputation</td>
<td>76.50</td>
<td>78.75</td>
<td>83.08</td>
<td>87.75</td>
<td>91.83</td>
<td>91.50</td>
<td>85.54</td>
</tr>
<tr>
<td>ANF-ITD-SFD</td>
<td>76.00</td>
<td>78.00</td>
<td>83.17</td>
<td>86.08</td>
<td>88.33</td>
<td>88.42</td>
<td>83.33</td>
</tr>
<tr>
<td>+Imputation</td>
<td>78.08</td>
<td>80.58</td>
<td>85.75</td>
<td>88.08</td>
<td>91.00</td>
<td>91.50</td>
<td>85.83</td>
</tr>
</tbody>
</table>

The SFD-based imputation has a clear advantage over a standard missing data imputation setup, in which the segmentation has to be identified independently and prior to the imputation stage. The benefit of considering multiple segmentation hypotheses and being able to exploit constraints from the top-down speech models has been demonstrated in previous work (Ma et al., 2012; Ma and Barker, 2012).

7. Conclusion and discussion

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 biases the decoder towards accepting fragments that are believed to originate from a known target source location.

Both the noise floor model and the localisation cues have been shown to effectively inform a better segmentation model within the fragment decoding framework, which has resulted in significant word error reduction. The fragment decoding framework has been able to combine strengths of both techniques to further improve its robustness to noise. Finally, the paper has shown that the SFD framework can be employed for missing data imputation to reconstruct full-band spectral features that can then be processed further and used as input to conventional ASR systems. Using the orthogonalised features based on the output of the best performing SFD system (ANF-ITD-SFD), the imputation system achieves an average keyword recognition accuracy of 85.83% for the final test set on the PASCAL CHiME Challenge task.

7.1. Computational complexity

The full system processes each test utterance in a multi-pass manner. On an Intel 2.27 GHz machine running Linux, the fragment generation stage (C++ implementation) runs 5 times faster than real-time. The noise floor tracking component (Matlab implementation) runs 2 times faster than real-time. The binaural localisation component (C implementation) runs 2 times slower than real-time. The fragment decoding and imputation stage (C++ implementation) runs 7 times slower than real-time. The final decoding stage (using HTK 3.4.1) runs 12 times faster than real-time.

7.2. Future directions

In the current system the fragment separation, noise tracking and binaural fragment localisation are conducted independently 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. Another future work will explore the possibility of learning the parameters used in converting noise floor
estimation and binaural localisation cues to probabilities. Currently the parameters have been empirically tuned on the development set, which have been found effective on this task. A more systematic fashion of learning the parameters statistically from the data will be explored.

References


Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P. The PASCAL CHiME speech separation and recognition challenge. Computer Speech and Language, submitted for publication.


