Combining Speech Fragment Decoding and Adaptive Noise Floor Modeling

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

***Abstract*—This paper presents a novel noise-robust automatic speech recognition (ASR) system that combines aspects of the noise modeling and source separation approaches to the problem. The combined approach has been motivated by the observation that the noise backgrounds encountered in everyday listening situations can be roughly characterized as a slowly varying noise floor in which there are embedded a mixture of energetic but unpredictable acoustic events. Our solution combines two complementary techniques. First, an adaptive noise floor model estimates the degree to which high-energy acoustic events are masked by the noise floor (represented by a soft missing data mask). Second, a fragment decoding system attempts to interpret the high-energy regions that are not accounted for by the noise floor model. This component uses models of the target speech to decide whether fragments should be included in the target speech stream or not. Our experiments on the CHiME corpus task show that the combined approach performs significantly better than systems using either the noise model or fragment decoding approach alone, and substantially outperforms multicondition training.**

***Index Terms*—Adaptive noise floor modeling, fragment decoding, missing data decoding, noise robust speech recognition.**

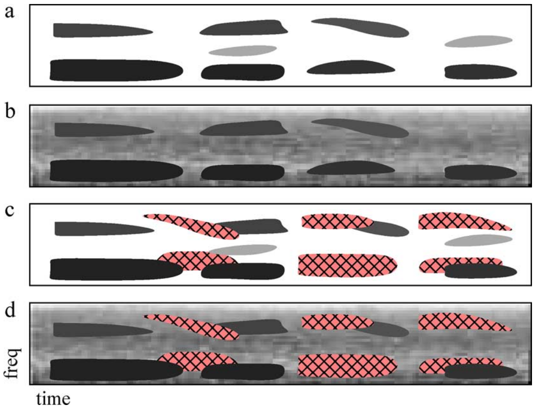

Fig. 1. Schematic time–frequency representation of speech in different backgrounds. (a) Speech with no background noise: the lighter shade of gray represents low energy. (b) Speech in quasi-stationary noisy: low speech energy is masked by noise. (c) Simultaneous speech: patches with the filled pattern represent another speaker. (d) Speech in more natural noise conditions.

I. Introduction

This paper considers the problem of distant microphone speech recognition in an everyday domestic environment. This problem is *important* because solutions would open the way to a new generation of applications [1]. In particular, such solutions would enable home-automation applications that would be valuable in the context of an increasingly ageing society. However, the problem is *difficult* because our homes tend to be noisy and unpredictable places that lie a long way outside the operating conditions of current speech recognition technology [1]: the target speech will be part of a heterogeneous mixture of competing sources; the combined noise energy may be comparable to or even greater than that of the speech; there will be significant room reverberation effects that will hinder source separation techniques.

There exists an extremely diverse set of techniques for noise-robust speech recognition but they can be loosely categorized into two broad approaches, which we will term *noise estimation* and *signal separation*. Noise estimation approaches rely on it being possible to estimate a model of the spectral characteristics of the noise background. This model, which might be as simple as an average noise spectrum, is then used to either "subtract" the noise from the mixture (e.g., spectral subtraction [2], [3]), estimate the noise masking pattern (missing data techniques [4]–[6] and uncertainty decoding [7]–[9]), or to adapt the speech model via a model combination technique (e.g., [10]–[14]). These techniques clearly depend on the quality of the noise model and work well in situations where an accurate model can be estimated, e.g., where the noise is known to be quasi-stationary or to have predictable dynamics that allow it to be tracked with some degree of certainty [Fig. 1(b)]. These conditions are seldom met in everyday listening conditions where the noise is itself a mixture of sources with unpredictably changing levels of activity.

In conditions where the noise spectrum cannot be readily estimated, a signal separation-based approach can sometimes be applied. Such approaches exploit the continuity of primitive signal properties (e.g., pitch or location) to allow some form of source separation prior to recognition. For example, pitch can remain an effective cue even in single-channel mixtures and it was exploited by the majority of systems competing in the first Pascal Speech Separation Challenge [15]. The non-negative matrix factorisation algorithm can also be applied in single microphone speech separation systems [16], [17]. In multi-microphone systems location estimates can be used [18], [19]. However, by focusing on separation of instantaneous speech mixtures in noise-free conditions, as depicted in Fig. 1(c), the Pascal Speech Separation Challenge was not particularly representative of the demands of real noise-robust systems.

Manuscript received January 13, 2011; revised June 07, 2011; accepted August 03, 2011. Date of publication August 22, 2011; date of current version January 18, 2012. This work was supported by U.K. EPSRC Grant EP/G039046/1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr Mark J. F. Gales.

The authors are with the Department of Computer Science, University of Sheffield, Sheffield S1 4DP, U.K. (e-mail: n.ma@dcs.shef.ac.uk; j.barker@dcs.shef.ac.uk; h.christensen@dcs.shef.ac.uk; p.green@dcs.shef.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2165945

Recently, a domestic noise corpus, the CHiME corpus [20], was developed aiming to accurately replicate natural noise contamination. The domestic noise backgrounds present challenging conditions for speech recognition because they are highly unpredictable: they can be broadly described as having an ambient, slowly varying noise floor that is overlaid by potentially overlapping acoustic events from a range of sources including speech, human movement and mechanical sounds. Section V-A gives more details about the CHiME corpus.

This work studies automatic speech recognition in such an environment. In particular, it investigates a noise estimation approach—adaptive noise floor estimation combined with a soft missing data recognizer (ANF-MD), and a separation-based approach—speech fragment decoding (SFD). The former is able to perform well during segments where the background is relatively "uneventful" and good noise floor approximations can be estimated. The latter approach uses cues to affect a partial separation of sources, but may struggle to handle the ambient noise floor which often exhibits weak pitch and localization cues. This paper also examine ways in which the ANF-MD and SFD techniques may be combined to take advantage of the complementary strengths of noise modeling and signal separation approaches. The proposed approach is to apply soft missing data decoding for time–frequency (T-F) regions with stationary noise, identified by an adaptive noise floor tracker, and employ fragment decoding to deal with the remaining unpredictable acoustic events.

Section II reviews the speech fragment decoding approach to robust ASR. The proposed adaptive noise floor modeling algorithm is introduced in Section III. Section IV discusses our approach for combining the fragment decoding and noise floor modeling. Section V briefly describes the speech recognition task and various ASR systems, and presents novel ASR results. Analysis of the results drives a discussion, presented in Section VI, which considers more sophisticated system combination approaches. Section VII concludes this paper.

II. Fragment Decoding Framework

Speech is a highly modulated signal with energy which is sparsely distributed in time and frequency, concentrated in narrowband structures such as formants and harmonics, or short duration events such as bursts [21]. These properties are clearly evident in the compressed spectro-temporal representations commonly used in computational models of the auditory system [22], [23]. As the speech energy is unevenly distributed, when a speech signal is corrupted by additive noise, in some sparse time–frequency regions the speech energy will be far greater than that of the noise, even if globally the noise is more energetic than the speech. In these local regions, the corrupted speech signal is well-modeled by the noise-free speech signal. We refer to these regions as *reliable* speech evidence. Furthermore, the information encoding in clean speech is redundant such that speech still remains intelligible even when a large spectro-temporal region of the speech is removed [24]–[26]. This redundancy essentially allows human listeners to recognize speech in noise based on the relatively sparse "glimpses" of reliable evidence [27]. Note that if the noise is itself modulated, the opportunities for glimpsing patches or reliable speech evidence may be even greater.

The sparseness and redundancy of spectro-temporal speech representations motivate the missing data approach to robust ASR [4]. This approach assumes there is a process that can identify the spectro-temporal regions that may be considered reliable (missing data mask estimation), and it then matches these reliable regions to models of clean speech. The unreliable regions are treated by either imputation [28]—replacing noise corrupted regions with estimates of the clean speech signal, or marginalization [4]—considering all possible values that the clean speech may have taken given the noisy observation. Imputation-based techniques are a form of feature compensation and have the advantage of allowing integration with conventional recognition systems. Marginalization approaches, however, have better theoretical justification and form the basis of the systems used in this paper.

The marginalization-based missing data techniques provide the foundation of the fragment decoding framework which incorporates mask estimation as part of the decoding process. This section will review these techniques and their incorporation within the fragment decoding framework.

A. Marginalization-Based Missing Data Techniques

Let $\mathbf{Y}$ represent a sequence of noisy speech observations $\{\mathbf{y}^1,\ldots,\mathbf{y}^T\}$ where each $\mathbf{y}^t$ is a feature vector representing a spectral energy component at time t. Let $\mathbf{S}$ denote a corresponding binary segmentation (or "missing data mask") composed of a sequence of frames $\{\mathbf{s}^1,\ldots,\mathbf{s}^T\}$, each frame being a vector of binary indicator variables stating whether the corresponding time-frequency element is reliable (1) or unreliable (0). The missing data ASR task is to find the best underlying acoustic model state sequence $\mathbf{Q}=\{\mathbf{q}^1,\ldots,\mathbf{q}^T\}$ given both the sequence of noise-corrupted speech observations and the segmentation

$$\hat{\mathbf{Q}} = \operatorname*{argmax}_{\mathbf{Q}} P(\mathbf{Q}|\mathbf{Y},\mathbf{S}). \tag{1}$$

The sequence of noise-free target speech vectors $\mathbf{X}$ is not directly observed but can be introduced by integrating over all possibilities

$$\begin{aligned}\hat{\mathbf{Q}} &= \operatorname*{argmax}_{\mathbf{Q}} \int_{\mathbf{X}} P(\mathbf{Q},\mathbf{X}|\mathbf{Y},\mathbf{S})d\mathbf{X} &(2)\\ &= \operatorname*{argmax}_{\mathbf{Q}} \underbrace{\int_{\mathbf{X}} P(\mathbf{X}|\mathbf{Q})\frac{P(\mathbf{X}|\mathbf{Y},\mathbf{S})}{P(\mathbf{X})}d\mathbf{X}}_{f_{\mathbf{Q}}(\mathbf{Y},\mathbf{S})} P(\mathbf{Q}). &(3)\end{aligned}$$

Using the frame independence assumptions, the integral is factorized into the product of integrals over individual frames

$$\begin{aligned}f_{\mathbf{Q}}(\mathbf{Y},\mathbf{S}) &= \prod_{t=1}^{T} f_{\mathbf{q}^t}(\mathbf{y}^t,\mathbf{s}^t) &(4)\\ &= \prod_{t=1}^{T}\left(\int_{\mathbf{x}^t} p(\mathbf{x}^t|\mathbf{q}^t)\frac{p(\mathbf{x}^t|\mathbf{y}^t,\mathbf{s}^t)}{p(\mathbf{x}^t)}d\mathbf{x}^t\right). &(5)\end{aligned}$$

For the sake of simplicity the index t is sometimes omitted in the paper and $\mathbf{y}$ still represents a vector and $\mathbf{y}_i$ is a scalar representing the i^{th} dimension of $\mathbf{y}$. The best state sequence $\hat{\mathbf{Q}}$ can then be computed within the standard hidden Markov model framework using Viterbi decoding. The only alteration required is to substitute the usual state emission probability calculations, $p(\mathbf{x}|\mathbf{q})$, with evaluations of the integrals $f_{\mathbf{q}}(\mathbf{y},\mathbf{s})$.

The distributions $p(\mathbf{x}|\mathbf{q})$ are estimated in the usual way by training hidden Markov models (HMMs) on clean speech. They are typically represented using Gaussian mixture models (GMMs) with components having diagonal covariance. The term $p(\mathbf{x}|\mathbf{y},\mathbf{s})$ is the masking model: the speech observation $\mathbf{x}_i$ must take the value of the noisy observation $\mathbf{y}_i$ when $\mathbf{s}_i$ indicates that the observation is reliable; the speech $\mathbf{x}_i$ must have a value less than $\mathbf{y}_i$ if the observation is marked as unreliable. For values of $\mathbf{x}$ which violate these rules the probability is set to 0. It can then be reasonably assumed that $p(\mathbf{x}|\mathbf{y},\mathbf{s})$ is otherwise proportional to $p(\mathbf{x})$.

Substituting these distributions into (5) it is seen that the evaluation of $f_{\mathbf{q}}(\mathbf{y},\mathbf{s})$ follows the computation of the GMM likelihood, $p(\mathbf{x}|\mathbf{q})$, except that the contribution of each mixture component, k, now involves an (one-dimensional) integration over the possible values of the masked speech features

$$f_{k,\mathbf{q}}(\mathbf{y},\mathbf{s}) = \prod_{i\in\mathbb{R}^{\mathsf{s}}} \mathsf{R}(\mathbf{y}_i,k,\mathbf{q}) \prod_{j\in\mathbb{U}^{\mathsf{s}}} \mathsf{U}(\mathbf{y}_j,k,\mathbf{q}) \qquad (6)$$

where $\mathbb{R}^{\mathsf{s}}$ represents the set of reliable feature dimensions of $\mathbf{y}$ according to segmentation $\mathbf{s}$, i.e., $\mathbb{R}^{\mathsf{s}} = \{i : \mathbf{s}_i = 1\}$, $\mathbb{U}^{\mathsf{s}}$ represents the set of unreliable dimensions, i.e., $\mathbb{U}^{\mathsf{s}} = \{i : \mathbf{s}_i = 0\}$, and $\mathsf{R}(\cdot)$ and $\mathsf{U}(\cdot)$ are defined, respectively, as

$$\mathsf{R}(\mathbf{x}_i,k,\mathbf{q}) = \frac{p(\mathbf{x}_i|k,\mathbf{q})}{p(\mathbf{x}_i)} \qquad (7)$$

$$\mathsf{U}(\mathbf{x}_i,k,\mathbf{q}) = \xi_i \int_{-\infty}^{\mathbf{x}_i} p(\mathbf{x}_i|k,\mathbf{q}) d\mathbf{x}_i \qquad (8)$$

where $p(\mathbf{x}_i|k,\mathbf{q})$ is the univariate Gaussian distribution trained on clean speech, and ξ_i is a normalization constant to keep the integral a probability density function

$$\xi_i = \frac{1}{\int_{-\infty}^{\mathbf{y}_i} p(\mathbf{x}_i) d\mathbf{x}_i}. \qquad (9)$$

In practice, neither the denominators $p(\mathbf{x}_i)$ nor the normalizing constants ξ_i need to be computed because their effect is constant across all state sequence hypotheses.

In the unreliable part of $\mathbf{y}$, for static features the noisy observation serves as an effective upper bound for the range over which the unreliable features are integrated, as the energy that speech is contributing must be less than the observed energy. Unreliable delta features are ignored because it is not possible to compute an effective bound.

B. Soft Missing Data Decoding

In (6), the missing data mask is assumed to be binary. Performance loss caused by irreversible time–frequency labeling errors can be limited by introducing a soft mask [29]—a spectro-temporal map in which each element is associated with a real value ω_i in the range of $[0, 1]$, expressing a degree of confidence in the reliability of the data. With a soft mask, $f_{k,\mathbf{q}}(\mathbf{y},\omega)$ can also be evaluated using (7) and (8)

$$f_{k,\mathbf{q}}(\mathbf{y},\omega) = \prod_{i=1}^{D} \Big(\omega_i \mathsf{R}(\mathbf{y}_i,k,\mathbf{q}) + (1-\omega_i)\mathsf{U}(\mathbf{y}_i,k,\mathbf{q})\Big) \qquad (10)$$

where D is feature dimensionality.

C. Speech Fragment Decoding

Missing data decoding only considers a single segregation hypothesis, i.e., that represented by the missing data mask, $\mathbf{S}$. If the noise cannot be tracked reliably then the mask is hard to estimate correctly. A better solution is to consider various segregation hypotheses and let the top-down models decide which one best explains the acoustic scene. To couple the segmentation problem with recognition, the speech fragment decoding (SFD) framework [6] searches for the acoustic model state sequence $\mathbf{Q}$ and segmentation hypothesis $(\mathbf{S})$ that jointly maximize the probability

$$\hat{\mathbf{Q}}, \hat{\mathbf{S}} = \underset{\mathbf{Q},\mathbf{S}}{\operatorname{argmax}} P(\mathbf{Q},\mathbf{S}|\mathbf{Y}) \qquad (11)$$

$$= \underset{\mathbf{Q},\mathbf{S}}{\operatorname{argmax}} \int_{\mathbf{X}} P(\mathbf{Q},\mathbf{X},\mathbf{S}|\mathbf{Y}) d\mathbf{X} \qquad (12)$$

$$= \underset{\mathbf{Q},\mathbf{S}}{\operatorname{argmax}} \int_{\mathbf{X}} P(\mathbf{Q},\mathbf{X}|\mathbf{Y},\mathbf{S}) d\mathbf{X}\, P(\mathbf{S}|\mathbf{Y}). \qquad (13)$$

Note that the SFD framework uses exactly the same acoustic model as the marginalization-based missing data approach, i.e., (2) –(6). There is, however, an additional complexity arising from the fact that the decoder is now searching across competing segmentations: different segmentation hypotheses will have different normalization terms (i.e., $p(\mathbf{x}_i)$ and ξ_i), these terms therefore can no longer be ignored.

$P(\mathbf{S}|\mathbf{Y})$ is the segmentation model and the segmentation search is equivalent to selecting the best missing data mask. An exhaustive search is clearly not practical. Fortunately, most of the segmentation hypotheses do not need to be evaluated. Primitive grouping principles can be employed to group T-F elements according to local correlations of their characteristics. This process results in the acoustic mixture being divided into multiple *fragments* in the spectro-temporal plane—each fragment consists of a group of T-F elements that are considered to have originated from a single source. The fragments are imposing a form for $P(\mathbf{S}|\mathbf{Y})$ in (13)—it assigns equal probability to any foreground/background segmentation that can be constructed from the set of fragments, i.e., the region covered by each fragment must be either allocated exclusively to the foreground or to the background. All other segmentations are assigned a probability of 0. Barker *et al.* [6] have shown that the maximization over state sequence $\mathbf{Q}$ and segmentation $\mathbf{S}$ can be achieved via a Viterbi search over a lattice of segmentation and state sequence hypotheses [6].

SFD, like missing data decoding, can be generalized by the use of soft masks to express uncertainty about whether a T-F element is reliable according to a segmentation hypothesis $\mathbf{S}$

$$f_{k,\mathbf{q}}(\mathbf{y},\omega) = \prod_{i=1}^{D} \Big(\omega_i^{\mathsf{s}} \mathsf{R}(\mathbf{y}_i,k,\mathbf{q}) + \big(1-\omega_i^{\mathsf{s}}\big)\mathsf{U}(\mathbf{y}_i,k,\mathbf{q})\Big) \qquad (14)$$

where $\omega_i^{\mathbf{s}}$ is the probability of $\mathbf{y}_i$ being reliable according to $\mathbf{S}$. Soft missing data decoding (10) can be seen as a special case of (14) when the segmentation hypothesis $\mathbf{S}$ is fixed.

III. Adaptive Noise Floor Modeling

In many natural listening conditions such as domestic settings, the auditory scene can be approximately described as a slowly varying noise floor plus highly unpredictable acoustic "events." The idea of combining SFD and noise floor modeling is to build a "universal" noise model to account for slowly varying noise, and employ SFD to deal with unpredictable acoustic events. In this section we first investigate adaptive noise floor modeling.

The output of noise floor tracking can be expressed as a spectro-temporal map holding local signal-to-noise ratio (SNR) estimates. Such SNR maps have formed the basis of missing data mask estimation in many previous missing data ASR systems [29]–[31]. Typically, the noise power spectrum is estimated from regions where speech is assumed absent. For example, in [29] speech was assumed to be absent at the beginning of each utterance on the Aurora 2 task, and the noise power spectrum was estimated by averaging the first ten frames. This technique works well if the noise is sufficiently stationary—at least within the duration of an utterance. However, this is a poor assumption in many situations.

In the field of speech enhancement, several algorithms have been proposed for estimating spectra of nonstationary noise [3], [32], [33]. Many of these methods are based on tracking the minimum of a smoothed noisy spectrum over a finite window. The noise estimate for each frequency bin is obtained by scaling the minimum with a biasing factor based on the minimum statistics. These adaptive methods may be robust to nonstationary noise, as the noise estimate is updated continuously by averaging the noisy speech power spectrum with time–frequency dependent smoothing factors.

In this work, we employ an adaptive noise floor tracking algorithm similar to the minimum tracking-based methods. However, instead of tracking the minimum by averaging the previous noisy spectra over a finite window, the tracker models the noisy spectra as a mixture of Gaussians and the distribution that has the minimum mean value is used to obtain the noise estimate. This kind of adaptive tracking mixture models is often used in image processing for segmenting moving regions from background in image sequences [34].

A. Method

Let $\mathbf{Y}^w = \{\mathbf{y}^1, \ldots, \mathbf{y}^L\}$ represent a sequence of noisy speech vectors in a window w of L frames. In this work the model operates directly at the feature level, i.e., $\mathbf{y}^t$ represents a $\log_{10}$ compressed spectral feature vector obtained from an auditory filterbank. Delta features are not used. A GMM with diagonal covariance was fitted to the rolling window of noisy speech, using the expectation maximization (EM) algorithm. Since adjacent spectral dimensions are correlated, a well-separated subset of frequency channels, $\hat{\mathbf{Y}}^w = \{\hat{\mathbf{y}}^1, \ldots, \hat{\mathbf{y}}^L\}$, was chosen from the full frequency band, so that features are nearly independent. The maximum-likelihood estimate (MLE) of the parameters for $\hat{\mathbf{Y}}^w$ is

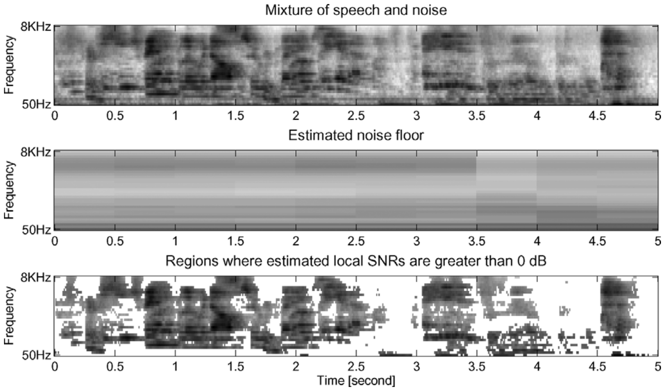

Fig. 2. Output of the adaptive noise floor tracker based on a mixture model. The white regions in the lower panel have SNR estimates less than 0 dB.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(\hat{\mathbf{Y}}^w|\theta). \tag{15}$$

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

$$\mu_k^w = \frac{\sum\limits_{t=1}^{L} P(\hat{\theta}_k|\hat{\mathbf{y}}^t)\mathbf{y}^t}{\sum\limits_{t=1}^{L} P(\hat{\theta}_k|\hat{\mathbf{y}}^t)}. \tag{16}$$

The noise floor estimate of $\mathbf{Y}^w$ is assumed to be the full-band mixture component mean that has the lowest energy

$$\hat{k} = \underset{k}{\operatorname{argmin}} \left|10^{\mu_k^w}\right| \tag{17}$$

$$\hat{\mathbf{n}} = \mu_{\hat{k}}^w. \tag{18}$$

In the current work, the noise floor tracking model employed two mixture components. The length of the rolling window L was set to five seconds. The GMM was continuously updated with a half-second increment, producing a fresh noise floor estimate for every half second. The subset of frequency channels were equally spaced on the equivalent rectangular bandwidth (ERB) scale [35] between 50 and 8000 Hz, with center frequencies located at: 118, 439, 1057, 2247, and 4538 Hz, respectively. These parameters were chosen after optimization on a development set.

B. Local SNR Estimation

The local SNR estimate is computed in decibels (dB) as

$$\mathrm{SNR}_i = 20\Big(\log_{10}\big(10^{\mathbf{y}_i} - 10^{\hat{\mathbf{n}}_i}\big) - \hat{\mathbf{n}}_i\Big) \tag{19}$$

where $\mathbf{y}$ is the $\log_{10}$ compressed noisy feature for each frame, and $\hat{\mathbf{n}}$ is the $\log_{10}$ compressed noise floor estimate using (18).

A typical output of this adaptive noise floor tracking technique is shown in Fig. 2. The upper panel is the cochleagram of a speech/noise mixture in the CHiME corpus. The middle panel shows the estimated noise floor, updated every half second. The regions where local SNR estimates are less than 0 dB are displayed in white in the lower panel.

IV. Combining SFD and Noise Floor Modeling

In order to handle both the quasi-stationary and unpredictable event components of the noise background we wish to combine the adaptive noise modeling and fragment decoding techniques. The combined technique will be termed adaptive noise floor speech fragment decoding (ANF-SFD). We use the adaptive noise floor model to estimate the degree to which high-energy acoustic events are masked by the noise floor. This is represented by a soft missing data mask. A fragment decoding system then attempts to interpret the high-energy regions that are not accounted for by the noise floor model. The first step is to separately generate soft missing data masks (using the adaptive noise tracker) and fragments (using harmonicity-based techniques [36]) from the noisy signals.

A. Soft Mask Generation

The soft missing data mask is produced by applying a sigmoid function to the local SNR estimates. This simple technique has been shown to produce effective soft masks in previous work [37]. For each T-F element, ω_i in (10) is computed as

$$\omega_i = \frac{1}{1+e^{-\alpha(\mathrm{SNR}_i-\beta)}} \tag{20}$$

where α determines the slope of the sigmoid function and the center β serves as the SNR threshold when computing soft MD masks. A higher SNR threshold will cause more T-F regions to be biased towards being interpreted as the noise background during decoding. The effect of different SNR thresholds will be discussed in more detail in Section VI-A.

B. Fragment Generation

This work employs techniques for tracking multiple pitches of simultaneous sounds in the autocorrelogram domain and use this information to identify fragments [36]. In brief, a running short-time autocorrelation is computed on the output of each gammatone filter using a 30-ms Hann window. For periodic sounds, the autocorrelogram [38], [39] exhibits symmetric tree-like structures whose stems are located on the delays that correspond to multiple pitch periods. 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 through local pitch estimates across time using the rule-based multi-pitch tracker developed in [36]. Each pitch track is then used to recruit a spectro-temporal fragment.

As discussed in Section II-C, fragment decoding can also make use of soft masks. In practice, the soft decoder takes a spectro-temporal confidence map $\mathbf{c}_i^{\mathsf{F}}$ as an additional input, which encodes the degree of belief that each T-F element is a true member of the fragment F. Confidence map values range from 0.5 (no confidence) to 1 (high confidence), and they are combined with the fragment labels to make a soft mask: in regions covered by foreground fragments, the soft missing data mask takes values directly from the confidence map $\mathbf{c}_i^{\mathsf{F}}$; in regions covered by background fragments, the mask takes the values $1-\mathbf{c}_i^{\mathsf{F}}$. Thus, the soft mask value $\omega_i^{\mathbf{s}}$ in (14) is

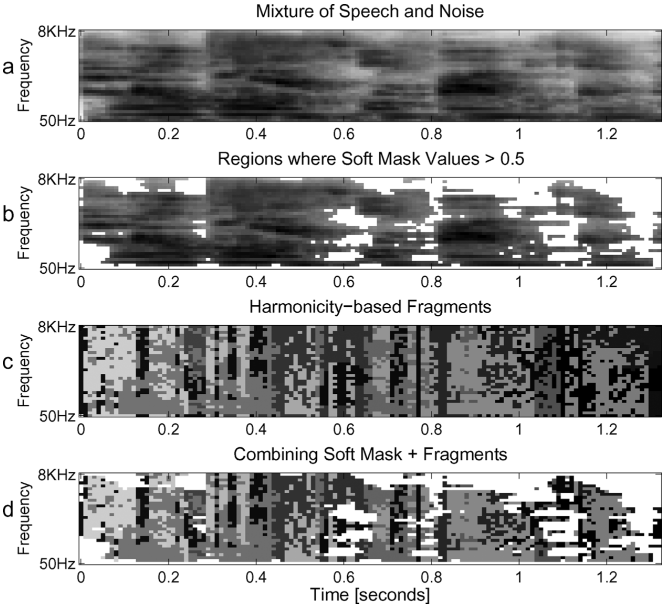

Fig. 3. a) Cochleagram of a speech/noise mixture. (b) Missing data mask derived from local SNR estimates. (c) Fragments identified by harmonicity analysis, represented as regions with different shades of gray. (d) Fragments excluding low SNR regions (white).

$$\omega_i^{\mathbf{s}} = \begin{cases} \mathbf{c}_i^{\mathsf{F}}, & \text{if } \mathsf{F} \in \mathbb{F}^{\mathbf{s}} \\ 1-\mathbf{c}_i^{\mathsf{F}}, & \text{otherwise} \end{cases} \tag{21}$$

where $\mathbb{F}^{\mathbf{s}}$ represents the set of fragments being considered as foreground according to segmentation hypothesis $\mathbf{S}$. The confidence map values were obtained from harmonicity analysis using the same algorithm and parameterization as in [36].

C. Combination Method

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

Soft decisions are also employed in the combined SFD system using (14). The low SNR regions directly take the SNR-based soft mask values ω_i from (20). For the remaining T-F elements that are included in a fragment F, the confidence values $\mathbf{c}_i^{\mathsf{F}}$ is used as in (21)

$$\omega_i^{\mathbf{s}} = \begin{cases} \omega_i, & \text{if } \omega_i \le 0.5 \\ \mathbf{c}_i^{\mathsf{F}}, & \text{if } \omega_i > 0.5, \mathsf{F} \in \mathbb{F}^{\mathbf{s}} \\ 1-\mathbf{c}_i^{\mathsf{F}}, & \text{otherwise} \end{cases}. \tag{22}$$

Fig. 3 illustrates this process. Fig. 3(a) is the cochleagram of a speech/noise mixture. The missing data mask derived from local SNR estimates is shown in Fig. 3(b), where regions with soft mask value less than 0.5 are displayed in white. Fragments identified by harmonicity analysis are shown in Fig. 3(c) using different shades of gray. Fig. 3(d) shows the fragments used by the combined system, where regions in white have low SNR estimates and are forced into the background. The process is akin to using the missing data mask in Fig. 3(b) to filter the fragments in Fig. 3(c).

The combined ANF-SFD approach improves upon the ANF-MD system in that the regions assigned high SNR estimates by the adaptive noise floor model are no longer always considered to be part of the foreground. Instead, they are divided into fragments, each of which may belong to either the speech foreground or the noise background. The foreground versus background identities of these fragments are decided with top-down knowledge from speech HMMs. So, fragments which are due to some unexpected noise source (e.g., a child shouting) will generally be rejected during fragment decoding because they are unlikely to match the speech HMMs.

The ANF-SFD system also differs from standard SFD 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 labeling hypotheses. Standard SFD 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.

V. Experiments and Results

A. *Speech Recognition Task*

All ASR experiments have been conducted using noise background taken from the CHiME corpus [20]. In brief, the corpus provides binaural audio recorded from a real domestic living room. The background has been recorded in multiple sessions over a period of several weeks using a binaural manikin that has remained in a fixed position within the room.

The experiments employ a 600 utterance test set that has been taken from the Grid corpus [40]. The Grid 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 occurs in every utterance. The average recognition accuracy for the two keywords is used to report the *keyword accuracy*.

The Grid utterances are processed in such a way as to simulate the effect of them having been recorded in the CHiME living room. Reverberation is added by convolving them with a binaural room impulse response (BRIR) measured in the CHiME living room at a position 2 m directly in front of the manikin [20]. The reverberated Grid utterances are then added to segments of the CHiME background audio. The SNRs are controlled so as to produce versions of the test set that range from −6 dB to 9 dB at intervals of 3 dB. Note, since CHiME recordings are binaural, the definition of the SNR is

$$\mathrm{SNR}_{\mathrm{dB}} = 20 \log_{10}\Big(\frac{A_{s,l} + A_{s,r}}{A_{n,l} + A_{n,r}}\Big) \qquad (23)$$

where A is root mean square amplitude, s and n represent speech and noise, and l and r represent the left and right channels, respectively.[1]

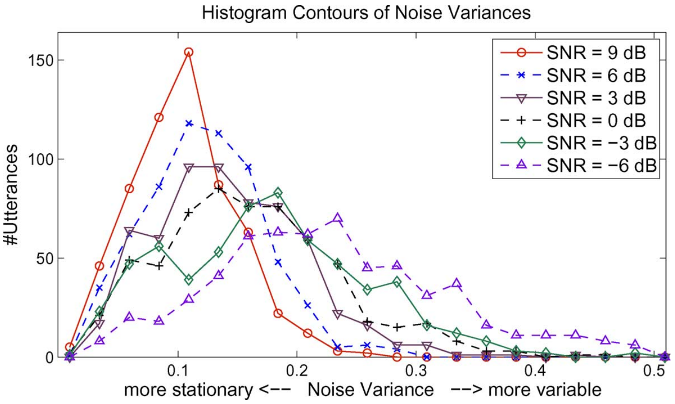

Fig. 4. Histogram contours of utterances in different noise variance bands for each SNR condition.

The SNR is controlled by selecting segments of the domestic audio recording that happen to be at the correct energy level with respect to the reverberated test utterance level, i.e., the −6 dB mixtures are using segments of the background recording that are naturally at a higher energy level than the segments used in the 9-dB mixtures. Controlling SNR in this manner rather than artificially amplifying or attenuating the background noise level results in mixtures that are more representative of situations that might occur in a real domestic ASR scenario, however, note it also means that different SNR conditions will typically contain different types of noise: in low SNR conditions it is common to find loud but often short-duration noises (e.g., a child shouting), while in quieter conditions noises tend to be more stationary.

Fig. 4 shows histogram contours of utterances in different noise variance bands. The noise variance was estimated from noise energy (using pre-mixed signals) of frames over each test utterance. It is clear that each SNR condition has a substantially different noise profile. The noise is mostly stationary at the high SNR end and becomes more variable at lower SNRs.

The CHiME corpus provides binaural signals, but the ASR evaluation reported here employs a single channel signal, which has been formed by averaging the pair of binaural signals in the time domain. No binaural cues are employed in this work. The same binaural room impulse response is used for both the training and the test condition. Therefore the training data set and test data set contain matching reverberation.

B. *Recogniser Setup*

Speaker-dependent word-level HMMs were used in all the ASR systems, following the "standard" model setup of the First Speech Separation Challenge [15]. The models were produced by performing four more iterations of EM training over a set of well-trained speaker-independent HMMs, using the 500 training utterances for each speaker. Each HMM state employed seven-component Gaussian mixtures with diagonal-covariance. All the ASR systems decoded each test utterance using the set of models corresponding to the speaker who spoke the utterance, with prior knowledge of speaker identities.

Two MFCC-based baseline systems have been considered. The first baseline employs the typical 39-dimensional MFCC features (with deltas and delta-deltas) and cepstral mean normalization (CMN). The second baseline employs "multicondition training" with the same MFCC+CMN features, but the

[1] Note in [20] the SNR calculation (1) contains a typographical error.

models used in the MFCC+CMN baseline have been retrained using noisy training data, constructed by mixing reverberated training speech with CHiME noise at SNR levels ranging from -6 dB to 9 dB, i.e., following the same procedures used in the construction of the test set. The noise signals used in multicondition training were taken from the same noise recordings used to mix the test data but not exactly the same noise samples were used.

All three missing-data-based recognizers employ models trained in noise-free conditions and there was no retraining on noisy data. ANF-MD is a soft missing data system, which employs soft SNR-based masks produced using the adaptive noise floor tracker as discussed in Section IV-A. The soft mask values were computed using (20): α was fixed as 0.1 and β was fixed to 12 dB for all test conditions after optimisation on the development set. SFD represents a standard soft fragment decoding system which employed fragments identified from multi-pitch analysis, as discussed in Section IV-B. The final system, ANF-SFD, is a soft fragment decoding system combined with adaptive noise floor modeling, as discussed in Section IV. Note the SNR threshold for computing the soft mask values in this system was optimized separately, and the best results on the development set were obtained with the SNR threshold of -3 dB.

Marginalization-based techniques require spectral features—missing features are localized in the spectral domain but not in the cepstral domain [4]. In this work the missing-data-based systems employed spectral features that are the auditory equivalent to a spectrogram, the cochleagram. They were produced with a 32-channel Gammatone filterbank distributed in frequency between 50 Hz and 8 kHz on the ERB scale [35], log-compressed and supplemented with deltas to form 64-dimensional feature vectors.

C. Results and Analysis

Table I shows the keyword accuracies of various ASR systems. They are also plotted in Fig. 5. First, the performance of the standard MFCC+CMN system decreases rapidly towards the low SNR end. Multicondition training exhibits considerably greater resistance to noise corruption, with a more moderate rate of decrease in recognition accuracy. The ANF-MD system has a performance that is better than that of the multicondition training system across all SNR conditions, despite the MD system not having access to noisy speech during training.

Second, both the combined ANF-SFD system and the standard SFD system substantially outperform the multicondition training system and the ANF-MD system at SNRs below 9 dB (p-value < 0.001 according to the McNemar test [41] on keyword-pair errors). This is not surprising given the nonstationary nature of the noise. In these conditions the noise is not just louder but also less stationary (see Fig. 4). For the ANF-MD system the noise can become highly unpredictable and difficult to track reliably. Therefore, many T-F regions may be incorrectly given high SNR estimates. It was observed that for the ANF-MD system, in order to compensate the SNR estimation errors, it was necessary to use an SNR threshold substantially higher than 0 dB when computing the soft mask (12 dB was used). Reasons for needing this high threshold are discussed in Section VI-A.

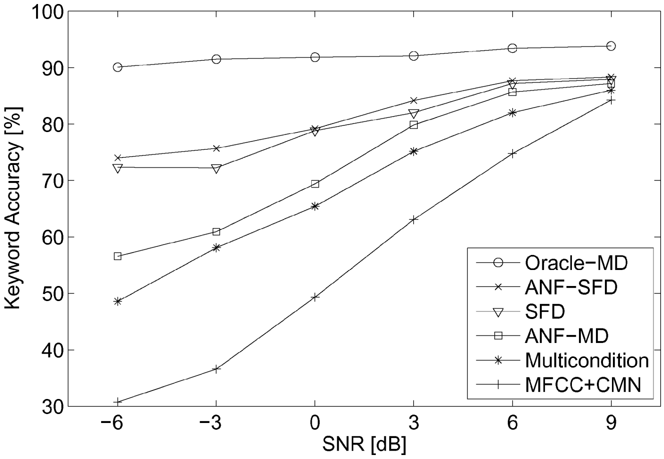

Fig. 5. Keyword recognition accuracies of typical MFCC-based recognizers and various missing-data-based systems across various SNRs.

TABLE I
KEYWORD RECOGNITION ACCURACIES (%) OF VARIOUS ASR SYSTEMS ON THE CHiME CORPUS TASK. THE RESULTS FOR EACH SNR ARE PRESENTED AS WELL AS THE AVERAGE (AVG.) ACROSS ALL SNRS

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
MFCC+CMN	30.75	36.58	49.33	63.08	74.75	84.25	56.46
Multicondition	48.58	58.08	65.42	75.17	82.00	86.00	69.21
Oracle-MD	90.08	91.50	91.83	92.08	93.42	93.83	92.12
ANF-MD	56.58	60.92	69.42	79.83	85.67	87.17	73.27
SFD	72.33	72.25	78.83	82.00	87.17	87.92	80.10
ANF-SFD	74.00	75.67	79.17	84.17	87.67	88.33	81.50

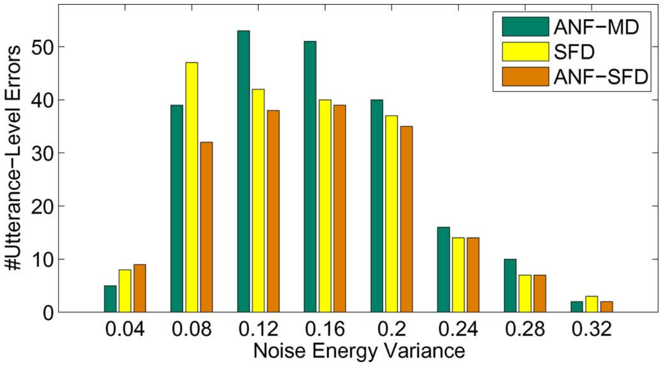

Fig. 6. Histograms of utterance-level keyword recognition errors in different noise variance bands, showing different effects of noise stationarity to the ANF-MD system, the standard SFD system, and the combined ANF-SFD system. The SNR is 3 dB.

The performance of the ANF-MD system was very similar to that of the SFD and ANF-SFD systems at the SNR of 9 dB where the noise is more stationary. This is expected since the ANF-MD system makes narrow assumptions about the noise, and assumptions which allow good performance when they happen to be correct.

Third, the combined ANF-SFD system also exhibits improved performance significantly over the SFD system (p-value $<$ 0.001 according to the McNemar test). A detailed analysis shows the error patterns of the ANF-MD system and the standard SFD system are different and complementary. Fig. 6 shows histograms of keyword recognition errors at the 3-dB SNR for the ANF-MD system, the standard SFD system, and the combined ANF-SFD system, respectively.

The histograms were computed against noise energy variances across each utterance in the test set. It is clear that ANF-MD produced fewer ASR errors in more stationary noise (variance less than 0.12), while the SFD system performed better when noise becomes more variable.

The complementary error pattern suggests that it may be possible to combine the two systems to produce better results, and this is clearly demonstrated by the ASR error histogram of the ANF-SFD system in Fig. 6. The combined ANF-SFD technique improves over standard SFD by reducing keyword errors mostly for utterances in more stationary noise, and the improvement over the ANF-MD system mainly comes from the cases in more variable noise.

Finally, results of a missing data system using "oracle" masks [4] are also presented (Oracle-MD). The oracle masks were derived from the true local SNR for each T-F element with access to the premixed speech and noise. Those elements with a local SNR > 0 dB were labeled as "reliable." Although the oracle mask results do not represent the performance of genuine recognition systems, they have been include as they give some indication of the potential performance of missing-data-based ASR systems. The Oracle-MD results remain almost flat across the SNR range. On previous tasks, such as Aurora 2, we observed a slight decrease at low SNRs. The difference can be explained by the observation that in CHiME the SNR-dependent datasets are not artificially produced but relate to operating conditions in a real environment. The degree of speech masking no longer increases linearly as the SNR decreases. In fact, at 0-dB SNR the area being labeled as reliable in the oracle mask on this task is 67% of that at 15 dB SNR, compared to only 39% on the Aurora 2 task.

VI. Discussion

A. *Use of Noise Floor Modeling in ANF-SFD Versus ANF-MD*

In both ANF-MD and ANF-SFD systems, SNR-based soft masks were produced by applying a sigmoid function to local SNR estimates produced by the noise floor tracking component. The sigmoid function center serves as a soft SNR threshold. A higher SNR threshold will cause more T-F mask elements to be assigned values of less than 0.5 and hence more weight is given to the 'masked by background' interpretation of the corresponding T-F features. However, the differing manner in which the noise floor model is used in the two systems has consequences for the tuning of the threshold.

As discussed in Section V-A, in the CHiME corpus, within each SNR condition the background has a different degree of stationarity. In low SNR conditions the noise is less stationary: the trackable noise floor is mixed with high energy noise events. For the ANF-MD system the local SNR estimation is based purely on the degree to which the observed energy exceeds the noise floor: any unpredicted high energy noise events will be erroneously considered to be T-F regions of high local SNR. This in turn means they will be assigned mask values that incorrectly bias their interpretation to be part of the speech foreground. This problem can be somewhat ameliorated by using an SNR threshold greater than 0 dB when computing soft masks. The higher threshold will remove some of the noise events from the foreground, but at the expense of also placing some speech dominated regions into the background.

TABLE II
KEYWORD RECOGNITION ACCURACIES (%) OF ANF-MD AND ANF-SFD SYSTEMS USING MASKS COMPUTED WITH VARIOUS SNR THRESHOLDS. THE RESULTS ARE LISTED FOR EACH SNR CONDITION WITH THE OPTIMIZED PARAMETERS HIGHLIGHTED IN BOLD

ANF-MD	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
SNR Threshold = 15 dB	58.67	61.50	70.42	79.75	85.08	86.75
SNR Threshold = 12 dB	**56.58**	**60.92**	**69.42**	**79.83**	**85.67**	**87.17**
SNR Threshold = 9 dB	54.67	59.08	67.92	79.08	84.25	87.58
SNR Threshold = 6 dB	52.67	57.50	67.00	78.75	84.92	87.33
SNR Threshold = 3 dB	50.75	57.17	65.50	76.83	84.00	87.08
ANF-SFD	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
SNR Threshold = 3 dB	73.00	73.50	80.08	83.75	87.67	88.50
SNR Threshold = 0 dB	72.83	74.83	79.83	83.92	87.50	88.75
SNR Threshold = -3 dB	**74.00**	**75.67**	**79.17**	**84.17**	**87.67**	**88.33**
SNR Threshold = -6 dB	73.42	75.08	79.42	84.42	87.17	88.25

The benefit to the ANF-MD system of an increased threshold for the low-SNR conditions is illustrated by Table II. However, in the high-SNR conditions where the noise is more stationary, the noise floor estimate is generally a good model of the noise, so the local SNR estimation is reliable and a sigmoid threshold of 0 dB will correctly discriminate between foreground and background dominated regions. In this case, increasing the threshold will incorrectly label speech regions as part of the background without the benefit of removing noise from the foreground, and the net result may be a reduction in recognition performance.

Although the same estimated SNR maps were used in both the ANF-MD and ANF-SFD systems, they use the information in the maps somewhat differently. Specifically, whereas the ANF-MD system uses the map to attempt to distinguish speech and background in a single thresholding stage, ANF-SFD applies a fragment-based segmentation step to regions that are have higher energy than the noise floor, and the foreground/background nature of these fragments is considered during decoding. In contrast to the ANF-MD system, it is therefore unnecessary to use a high threshold to exclude noise-dominated fragments because these fragments can be assigned the background interpretation during the decoding stage when they poorly match the statistical speech models. Using a lower threshold avoids the adverse consequence that, on average, unnecessary amounts of reliable speech are treated as missing data. For the ANF-SFD system -3 dB was found to be an optimum SNR threshold, but the system was not very sensitive to different thresholds around 0 dB as suggested in Table II.

B. *Fragment Decoding Versus Model Combination*

It is instructive to contrast the fragment decoding approach with model combination (MC) approaches [10]–[12]. The success of the fragment decoding technique requires that the spectro-temporal fragments can be robustly classified as speech or non-speech using knowledge embodied in the speech models. Importantly, it functions without the need for explicit noise event models. This makes the approach highly attractive in situations where it is not possible to construct good predictive models of the competing noise sources. However, in situations where data exists to train noise events models, then alternative model combination techniques can be applied.

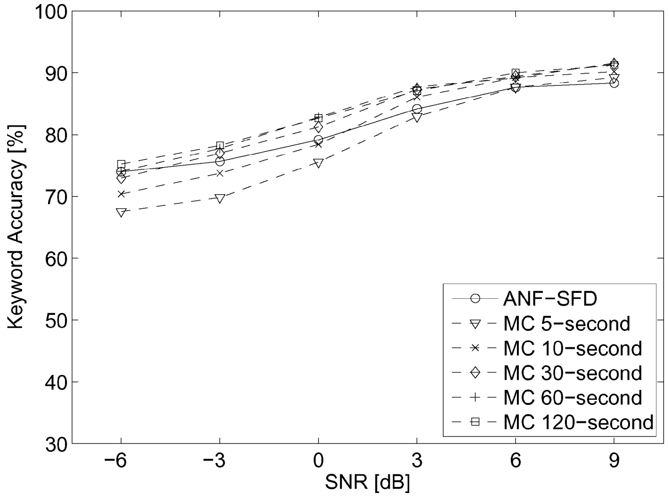

Fig. 7. Keyword recognition accuracies of fragment decoding and model combination (MC) systems. For model combination separate GMMs were trained for each utterance in the test set using a window of audio immediately preceding each utterance.

Given the variability of the CHiME noise it appears unlikely that a noise model constructed from a background training session would generalize usefully to an independent test session. However, the noise background, although highly varies, is perhaps locally predictable. For example, although the noise event "vocabulary" changes in response to transitions in acoustic scene context (e.g., people may leave or enter the room), significant state changes accrue only over long time scales, and the acoustic context can stay fixed for significant periods of time. It may therefore be possible to train *utterance-specific* noise models using short segments of audio immediately preceding any particular noisy speech utterance.

To test this idea, separate GMMs were trained for each utterance in the test set using a window of audio immediately preceding each utterance. Window sizes varying from 5 seconds up to 120 seconds were tested. The optimal number of mixture components for each window size was determined empirically from 5 Gaussians up to 256 Gaussians. The GMMs were trained on the same spectral-temporal representations used in earlier experiments. The noise GMMs were then combined with the clean speech HMMs as separate Markov chains in a factorial HMM [42].

Fig. 7 shows performance of the model combination systems with various window sizes. It is clear that the results improve as window size gets large with performance plateauing around 120 seconds. With 30 seconds or more of pre-speech audio, performance of model combination systems is significantly better than fragment decoding systems at most SNRs, with larger improvement at high SNRs where the noise is more stationary (thus better modeled by the noise GMMs).

It is surprising that the technique performs so well given the nonstationary nature of the background noise especially at low SNRs. It seems two minutes of audio is sufficient to represent the acoustic context. The combination approach we have applied is crude and better results would be expected with more sophisticated approaches, e.g., [12]–[14].

Although MC and SFD performance is similar, the approaches are using very different sorts of information (one uses pitch grouping, another uses noise model). There is a potential opportunity to form a combined approach—the existing fragment decoding framework can be generalized in the manner of a factorial HMM, but with a constraint offered by the segmentation hypothesis. The fragment-constrained model combination approach will be investigated in the future.

VII. Conclusion

This paper has presented a noise robust ASR system that combines aspects of previously separate noise modeling and source separation approaches to the problem. The combined approach has been motivated by the observation that everyday listening noise backgrounds can be roughly characterized in terms of a slowly varying noise floor in which there are embedded a mixture of energetic but unpredictable acoustic events. Our solution proceeds in two steps. First, an adaptive noise floor model estimates the degree to which high-energy acoustic events are masked by the noise floor (represented by a soft missing data mask). Second, a fragment decoding system attempts to interpret the high-energy regions that are not accounted for by the noise floor model. This component uses models of the target speech to decide whether fragments should be included in the target speech stream or not. The combined approach is able to outperform comparable systems using either the noise model or fragment decoding approach alone. The results have shown that model combination techniques also perform well on the CHiME task. Future work will explore the potential combination of model combination techniques with fragment decoding to exploit detailed knowledge of the background sound sources where available.

References

[1] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding, Part 1," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 75–80, May 2009.

[2] P. Lockwood and J. Boudy, "Experiments with nonlinear spectral subtractor (NSS), hidden Markov models and the projection for robust speech recognition in cars," *Speech Commun.*, vol. 11, pp. 215–228, 1992.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech. Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

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

[5] M. Seltzer, B. Raj, and R. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Commun.*, vol. 43, pp. 379–393, 2004.

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

[7] J. Droppo, L. Deng, and A. Acero, "Uncertainty decoding with splice for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 57–60.

[8] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech. Audio Process.*, vol. 13, no. 3, pp. 412–421, May 2005.

[9] H. Liao and M. Gales, "Issues with uncertainty decoding for noise robust automatic speech recognition," *Speech Commun.*, vol. 50, pp. 265–277, 2008.

[10] A. Varga and R. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1990, pp. 845–848.

[11] M. Gales and S. Young, "HMM recognition in noise using parallel model combination," in *Proc. Eurospeech*, Berlin, 1993.

[12] B. Frey, L. Deng, A. Acero, and T. Kristjansson, “ALGONQUIN: Iterating Laplace’s method to remove multiple types of distortion for robust speech recognition,” in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 901–904.

[13] J. R. Hershey, S. J. Rennie, and P. A. Olsen, “Super-human multi-talker speech recognition: A graphical modeling approach,” *Comput. Speech. Lang.*, vol. 24, pp. 45–66, 2010.

[14] S. J. Rennie, J. R. Hershey, and P. A. Olsen, “Single-channel multitalker speech recognition,” *IEEE Signal Process. Mag.*, vol. 27, pp. 66–80, 2010.

[15] M. Cooke, J. Hershey, and S. Rennie, “Monaural speech separation and recognition challenge,” *Comput. Speech. Lang.*, vol. 24, pp. 1–15, 2010.

[16] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Trans. Audio. Speech.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[17] M. N. Schmidt and R. K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in *Proc. Interspeech*, Pittsburgh, PA, 2006, pp. 2614–2617.

[18] M. Seltzer, B. Raj, and R. Stern, “Likelihood-maximizing beamforming for robust hands-free speech recognition,” *IEEE Trans. Speech. Audio Process.*, vol. 12, no. 5, pp. 489–498, Sep. 2004.

[19] R. Takeda, S. Yamamoto, K. Komatani, T. Ogata, and H. Okuno, “Missing-feature based speech recognition for two simultaneous speech signals separated by ICA with a pair of humanoid ears,” in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2006, pp. 878–885.

[20] 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*, 2010.

[21] S. Greenberg, W. Ainsworth and S. Greenberg, Eds., “Understanding speech understanding: Towards a unified theory of speech perception,” in *Proc. ESCA Workshop Auditory Basis Speech Percept.*, U.K., 1996, pp. 1–8.

[22] M. Slaney and R. Lyon, “On the importance of time – A temporal representation of sound,” in *Visual Representations of Speech Signals*, M. Cooke, S. Beet, and M. Crawford, Eds. Sussex, U.K.: Wiley, 1993, pp. 95–116.

[23] B. C. J. Moore, “Temporal integration and context effects in hearing,” in *J. Phonetics*, 2003, vol. 31, pp. 563–574.

[24] H. Fletcher, *Speech and Hearing in Communication*. New York: Van Nostrand, 1953.

[25] R. Warren, K. Riener, J. Bashford, and B. Brubaker, “Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits,” *Percept. Psychophys.*, vol. 57, pp. 175–182, 1995.

[26] K. Kasturi, P. C. Loizou, M. Dorman, and T. Spahr, “The intelligibility of speech with “holes” in the spectrum,” *J. Acoust. Soc. Amer.*, vol. 112, pp. 1102–1111, 2002.

[27] M. Cooke, “A glimpsing model of speech perception in noise,” *J. Acoust. Soc. Amer.*, vol. 119, pp. 1562–1573, 2006.

[28] B. Raj, M. Seltzer, and R. Stern, “Reconstruction of missing features for robust speech recognition,” *Speech Commun.*, vol. 43, pp. 275–296, 2004.

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

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

[31] 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.

[32] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Trans. Speech. Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[33] S. Rangachari and P. C. Loizou, “A noise-estimation algorithm for highly non-stationary environments,” *Speech Commun.*, vol. 48, pp. 220–231, 2006.

[34] C. Stauffer and W. Grimson, “Learning patterns of activity using real-time tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[35] B. Glasberg and B. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hearing Res.*, vol. 47, pp. 103–138, 1990.

[36] N. Ma, P. Green, J. Barker, and A. Coy, “Exploiting correlogram structure for robust speech recognition with multiple speech sources,” *Speech Commun.*, vol. 49, pp. 874–891, 2007.

[37] J. Barker, M. Cooke, and P. Green, “Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise,” in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 213–216.

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

[39] M. Slaney and R. Lyon, “A perceptual pitch detector,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Albequerque, NM, 1990, pp. 357–360.

[40] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *J. Acoust. Soc. Amer.*, vol. 120, pp. 2421–2424, 2006.

[41] L. Gillick and S. Cox, “Some statistical issues in the comparison of speech recognition algorithms,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1989, pp. 532–535.

[42] Z. Ghahramani and M. I. Jordan, “Factorial hidden Markov models,” *Mach. Learn.*, vol. 29, pp. 245–273, 1997.

Ning Ma received the B.Sc. degree from the South China University of Technology, Guangzhou, China, in 2002, and the M.Sc. and Ph.D. degrees from the University of Sheffield, Sheffield, U.K., in 2003 and 2008, respectively.

He has been a Visiting Research Scientist at the University of Washington, Seattle, studying the graphical models for robust conversational speech recognition. He is currently a Postdoctoral Research Associate in the Speech and Hearing Research Group, University of Sheffield, and his research interests include noise-robust automatic speech recognition, computational auditory scene analysis, and speech perception in noise.

Jon Barker received the B.A. degree in electrical and information sciences from the University of Cambridge, Cambridge, U.K., in 1991 and the Ph.D. degree in computer science from the University of Sheffield, Sheffield, U.K., in 1998.

He has spent a year as a Researcher at ICP, Grenoble, France, studying audiovisual speech perception and has spent time as a Visiting Research Scientist at IDIAP, Martigny, Switzerland, and ICSI, Berkeley, CA. He is currently a Senior Lecturer in Computer Science at the University of Sheffield. His research interests include audiovisual speech perception, robust automatic speech recognition, and audiovisual speech processing. He has authored or coauthored over 50 papers in these areas.

Heidi Christensen received the M.Sc. and Ph.D. degrees from Aalborg University, Aalborg, Denmark, in 1996 and 2002, respectively.

She is a Research Associate at the School for Health and Related Research, University of Sheffield. Before that she was a Research Associate at the Department of Computer Science, University of Sheffield for ten years working on numerous European and U.K.-funded project. Her research interests mainly concern the spoken language processing, clinical application of speech technology, and binaural machine listening.

Phil Green received the B.Sc. degree from the University of Reading, Reading, U.K., in 1967 and the Ph.D. degree from the University of Keele, Keele, U.K., in 1971.

He founded the speech research group at the University of Sheffield, Sheffield, U.K., in 1985. He has around 100 publications on a variety of topics within speech science and technology. In recent years, he has focused on the links between computational auditory scene analysis and ASR and on clinical applications of speech technology. He has been principle investigator for around 12 U.K.- and EC-funded projects.