Distant microphone speech recognition in a noisy indoor environment: combining soft missing data and speech fragment decoding

Ning Ma, Jon Barker, Heidi Christensen, Phil Green

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

Abstract

This paper examines the problem of distant microphone speech recognition in noisy indoor home environments. The noise background can be roughly characterised in terms of a slowly varying noise floor in which there are embedded a mixture of energetic but unpredictable acoustic events. Our solution to the problem combines two complementary techniques. First, a soft missing data mask is formed which estimates the degree to which energetic acoustic events are masked by the noise floor. This step relies on a simple adaptive noise model. Second, a fragment decoding system attempts to interpret the energetic regions that are not accounted for by the noise floor model. This component uses models of the target speech to decide whether fragments (time-frequency regions dominated by a single sound source) should be included in the target speech stream or not. This combined approach is able to achieve a performance that is modestly superior to that achieved using speech fragment decoding without an adaptive noise floor. Our experiments also show that speech fragment decoding performs far better than soft missing data decoding in variable noise, achieving 73% keyword recognition accuracy at -6 dB SNR on the Grid corpus task and substantially outperforming multicondition training.

Index Terms: Noise robust speech recognition; Fragment decoding; Missing data; Reverberation

1. Introduction

This paper considers the problem of distant microphone speech recognition in an everyday domestic environment. This problem is *interesting* because solutions would open the door to a new generation of applications. In particular, 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: 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 categorised 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 [1]), estimate the noise masking pattern (missing data techniques [2]), or to adapt the speech model via a model combination technique (e.g. [3, 4, 5]). These techniques clearly depend on the quality of the noise model and work well in situations where an accurate model can be easily 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 (as represented by Fig. 1b). These conditions are seldom met in everyday listening conditions, where the noise is itself a mixture of sources with unpredictably changing levels of activity.

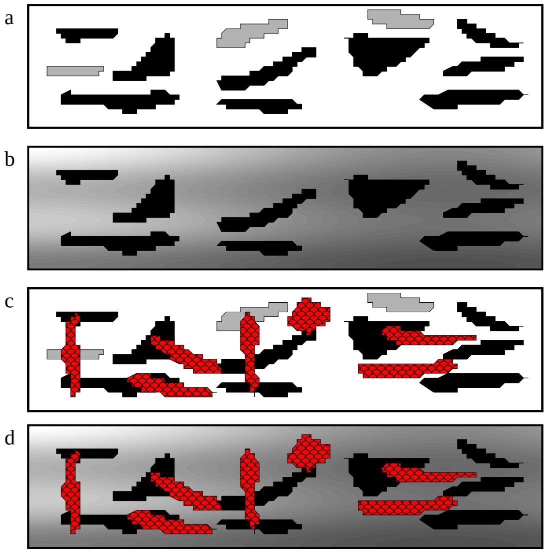

Figure 1: Schematic time-frequency representation of speech in different backgrounds: a) Speech with no background noise; b) Speech in quasi-stationary noisy; c) Simultaneous speech; d) Speech in natural noise conditions.

In conditions where the noise spectrum cannot be readily estimated, a signal separation based approach to robustness 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. In multi-microphone systems location estimates can be used. Alternatively, pitch can remain an effective cue even in single-channel mixtures. For example, pitch was exploited by the majority of systems competing in the recent Pascal Speech Separation Challenge evaluation [6]. However, by focusing on separation of instantaneous speech mixtures in noise free conditions (Fig. 1c) this challenge was not particularly representative of the demands of real noise-robust systems.

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

The domestic noise backgrounds that we employ in the current work are challenging because they do not have a single ‘character’. Instead, they can be broadly described as having an ambient, slowly varying noise floor that is overlaid by unpredictable acoustic events such as speech, human movement and mechanical sounds (Fig. 1d).

The current work studies distant microphone speech recognition in this environment, comparing a noise estimation approach – soft Missing Data (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 localisation cues.

This paper also examines ways in which the soft MD and SFD techniques may be combined to take advantage of the complementary strengths of noise modelling and signal separation approaches. Section 2 briefly describes the construction of the background noise corpus and the speech recognition task. Sections 3 reviews soft MD and SFD and compares their ASR performance. Section 4 provides an approach for combining the two techniques and presents novel ASR results. Analysis of the results drives a discussion of more sophisticatedly combined systems in Section 5. Section 6 concludes this paper.

2. Task

All the ASR experiments were conducted using the CHiME corpus [7], which accurately replicates natural contamination. Briefly, short speech utterances are reverberated with room impulse responses measured at various locations in different rooms. They are then mixed at a normal speaking level with domestic noise recorded at the corresponding locations. The SNRs range from -6 dB to 18 dB at an interval of 3 dB.

In CHiME, test material with a specific SNR is generated by selecting segments with noise at the desired and naturally occurring levels, rather than taking the same noise, adjusting its levels, and then adding it to clean speech. This is a more realistic procedure, because different SNR bands typically contain noises of different types. 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 are more stationary.

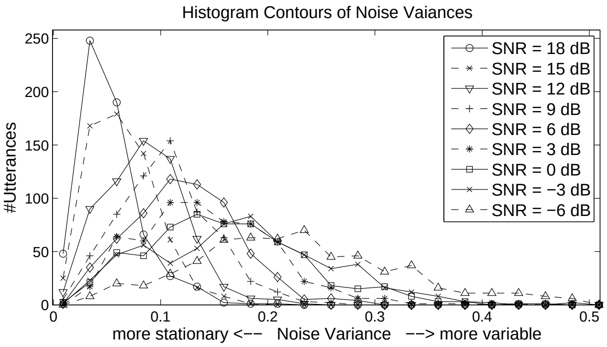

Figure 2: Each line shows the histogram contour of utterances in different noise variance bands for one SNR condition.

Fig. 2 shows histogram contours of test utterances in different noise variance bands. It is clear that each SNR has a substantially different noise profile. The noise is mostly stationary at the high SNR end and becomes more variable at lower SNRs.

Test speech is taken from the Grid corpus [8]. There are 51 words in the vocabulary and the ASR task is to identify 2 keywords – the letter and the digit – in each utterance. The average recognition accuracy for the 2 keywords is used to report the *Keyword Accuracy*, giving the chance level for this task at around 7%. In this paper the training and test condition is fixed to a single location in one room (lounge_200cm_az0). Therefore the training data and test data have matching reverberation.

The CHiME corpus provides binaural signals, but the ASR evaluation reported here employs monaural signals, which are formed by averaging the binaural signals in the time domain. No binaural cues are employed in this work.

3. Baseline Recognition Systems

3.1. Standard ASR systems

We first evaluated two standard ASR systems. The first system employed standard 39 dimensional MFCC features (with deltas and accelerations) plus Cepstral Mean Normalisation (CMN). Speaker-Dependent (SD) word-level Hidden Markov Models (HMMs) were used, following the ‘standard’ model setup of the 2006 Speech Separation Challenge [6]. The SD HMMs were produced by performing 4 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 7-component Gaussian mixtures with diagonal-covariance. There was no retraining on noisy data.

The second baseline employed multicondition training. The HMMs from the first baseline were retrained using noisy training data, constructed by mixing reverberated training speech with CHiME noise at SNR levels ranging from -6 dB to 21 dB. The noise used in multicondition training came from the same noise recordings used in mixing test data.

A standard Viterbi decoder was employed to recognise each utterance using the set of SD models corresponding to the speaker who spoke that utterance, with prior knowledge of speaker identities [9].

3.2. Missing data based systems

Both the soft MD system and the SFD system employ marginalisation based missing data techniques [2] at core. When an observed feature vector, x, may be partially corrupted by noise, we can denote the reliable part as x_r and the unreliable part as x_u. If the state distribution $p(x|q)$ is modelled by a mixture of Gaussian distributions with diagonal covariance, the distribution for each component k can be evaluated as the marginal distribution of x_r by integrating over x_u:

$$p(x|q,k) = \prod_{i \in r} p(x_i|q,k) \prod_{i \in u} \int_{-\infty}^{x_i} p(x'_i|q,k)dx'_i \qquad (1)$$

where $p(x_i|q,k)$ is the univariate Gaussian distribution.

Marginalisation based techniques require spectral features: missing features are localised in the spectral domain but not in the cepstral domain [2]. In this work We employed spectral features that are the auditory equivalent to a spectrogram, the cochleagram [10]. They were produced via a 32-channel Gammatone filterbank distributed in frequency between 50 Hz and 8000 Hz on the Equivalent Rectangular Bandwidth (ERB) scale [11], log-compressed and supplemented with deltas to form 64-dimensional feature vectors. Both missing data based systems employed speaker-dependent HMMs trained on reverberated speech, and there was no retraining for noisy conditions.

i. Soft missing data system

In Eq. 1 the missing data mask is assumed to be binary. Performance loss caused by irreversible Time-Frequency (T-F) labelling errors can be limited by introducing a soft missing data mask [12], in which each T-F pixel is associated with a probability value in the range of $[0, 1]$, expressing a degree of confidence in the reliability of the data. With a soft mask $p(x|q, k)$ can be evaluated as a weighted sum of likelihoods and marginals:

$$p(x|q,k) =$$

$$\prod_{i=1}^{N} \left(w_i p(x_i|q,k) + (1 - w_i) \frac{1}{x_i} \int_{-\infty}^{x_i} p(x'_i|q,k) dx'_i \right) \quad (2)$$

where w_i is the soft value for the i^{th} dimension and N is feature dimensionality. The use of soft masks has been shown to improve recognition accuracy significantly [13].

Previous attempts at deriving MD masks have often made use of local SNR estimates. For example, [13] assumed speech is absent at the beginning of each utterance on the Aurora 2 task, and noise spectrum is estimated by averaging the first 10 frames. The technique works well if the noise is sufficiently stationary – at least within the duration of each utterance. This is a poor assumption in many situations. We therefore employ an adaptive noise floor tracking technique in this work.

In brief, a Gaussian Mixture Model (GMM) with diagonal covariance was fitted to a rolling buffer of noisy speech, and the component with the lowest energy was assumed to be the noise floor. The GMM was updated with a half second increment, producing a noise floor for every half second. Since adjacent spectral channels are not independent, we chose only a subset (6 channels) of the full frequency band so that features were nearly independent. Our experiments show that recognition accuracies were not sensitive to the number of Gaussian components and the buffer size[1].

A typical output of this adaptive noise floor tracking technique is shown in Fig. 3. The upper panel is the cochleagram of a 5-second long 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 greater than 0 dB are displayed in the lower panel. Our preliminary experiments show that the MD system employing this technique performed substantially better than using masks generated by simply averaging 10 frames prior to the speech onset.

Soft missing data mask values were produced by applying a sigmoid function to the local SNR estimates. The centre of the sigmoid function serves as the SNR threshold for computing soft MD masks, which was fixed to 9 dB for all test conditions after optimisation on the development set.

ii. Speech fragment decoding system

The missing data system only considers a single segregation hypothesis, i.e. the missing data mask, which could be incorrect. A better solution would be to consider various segregation hypotheses and let the top-down models decide which one best explains the acoustic scene. To fully couple the segregation problem with recognition, the SFD framework searches for the word sequence (W) and segregation hypothesis (S) that together are most probable, given the noisy signal (Y):

$$\begin{aligned} \hat{W}, \hat{S} &= \arg\max_{W,S} P(W, S|Y) \\ &= \arg\max_{W,S} P(W|S, Y)P(S|Y) \end{aligned} \quad (3)$$

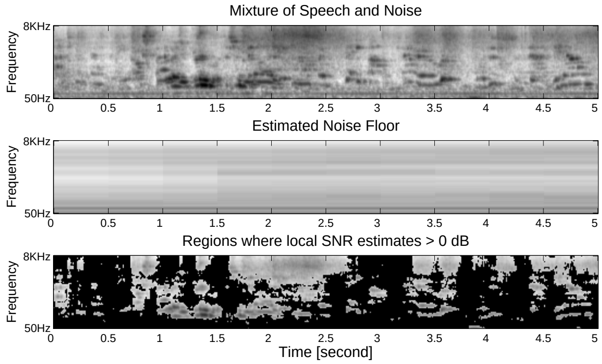
Mixture of Speech and Noise
Estimated Noise Floor
Regions where local SNR estimates > 0 dB

Figure 3: Illustration of adaptive noise estimation using a GMM based noise floor tracker. The GMM is adapted every half second using a 5-second buffer.

$P(W|S, Y)$ is equivalent to missing data decoding given a fixed mask, S, and $P(S|Y)$ is the segregation model [14]. The search is now being conducted over the joint space of word sequences and segregation hypotheses. In practice, given each segregation, the word sequence dimension of the search can be efficiently performed using missing data techniques. The segregation search is then equivalent to selecting the best missing data mask. An exhaustive search is clearly not practical. Fortunately, most of the segregation hypotheses do not need to be evaluated. Primitive grouping principles can be employed to group T-F pixels according to local correlations of their characteristics. This process results in the acoustic mixture being divided into multiple local *fragments* in the spectro-temporal plane. Barker et al. [14] show that decoding can be performed in an efficient manner using fragments.

The concept of fragments is consistent with the underlying principles of auditory scene analysis. They are the physical representation of the components from which perceptual ‘auditory streams’ are built [15]. In SFD, each fragment is represented by labelling all its T-F pixels with a unique positive integer. Finding such fragments is an easier task than separating the target utterance from the noise background directly because fragment foreground/background identities do not have to be decided until the recognition stage when top-down models are available.

In this work we employ techniques for tracking multiple pitches of simultaneous sounds and use this information to identify fragments [16, 17]. The idea of fragments has also been employed to integrate different auditory grouping cues for better localisation of sound sources in reverberant recordings [18].

3.3. Results

Fig. 4 shows the ASR results of these baseline systems. Firstly, the MFCC+CMN system performed reasonably well in conditions with little noise, but its performance decreased rapidly towards the low SNR end. Multicondition training provided considerably better resistance to noise corruption, with a more moderate decreasing rate in recognition accuracy.

[1] We used a GMM with 2 components and a buffer of 5 seconds.

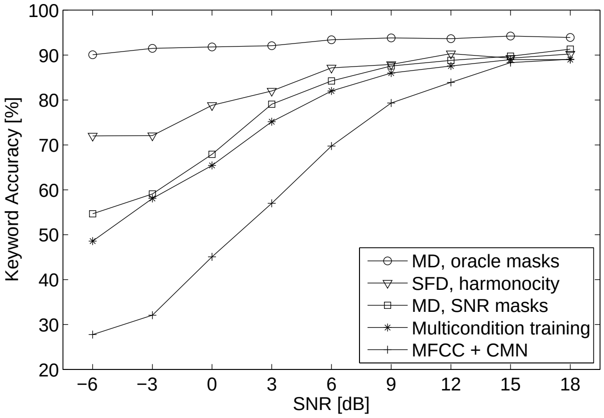

Figure 4: Recognition results for SFD compared against those for soft MD and various standard baseline ASR systems.

Secondly, the soft MD system produced higher recognition accuracies over the multicondition training system (stars) consistently across all SNR conditions, despite that the MD system did not have access to noisy speech during training.

Thirdly, the SFD system substantially outperforms the multicondition training system and the MD system at SNRs below 9 dB. This is not surprising given the non-stationarity nature of the noise. In these conditions the noise is not just louder but also less stationary (see Fig. 2). For the MD system the noise can become so unpredictable that it is almost impossible to track. Therefore many T-F regions may be incorrectly given high SNR estimates. We notice that for the soft MD system, in order to compensate the SNR estimation errors, a SNR threshold substantially higher than 0 dB was needed when computing the soft mask (9 dB was used).

The MD system produced slightly better results than SFD at SNRs of 15 dB and 18 dB where the noise is fairly stationary, but the difference is not significant. Better performance may be expected since the MD system makes narrow assumptions about the noise, which give it an advantage when the assumptions happen to be correct.

Finally, results of a MD system using 'oracle' masks [2] are also presented. The oracle masks were derived from the true local SNR for each T-F pixel with access to the premixed speech and noise. Those pixels with a local SNR $>$ 0 dB were labelled as 'reliable'. Although the oracle masks are artificial, their results demonstrate the upper boundary of missing data based ASR systems.

The oracle mask results remain almost flat across the SNR range. On previous tasks, such as Aurora 2, we observed a slight decrease at low SNRs. This is because in CHiME the SNR-dependent datasets are not artificially produced but relate to operating conditions in a real environment. The level of speech masking no longer increases linearly as the SNR decreases. In fact, at 0 dB SNR the area being labelled 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.

4. Combining MD and SFD

4.1. Motivation

A detailed analysis shows the error patterns for the soft MD system and the SFD system are different and complementary. Fig. 5 shows histogram contours of ASR error differences between the two systems in matched pairs for the 3 dB SNR condition. The squares show when the MD system correctly recognises more keywords than the SFD system for each utterance, while the triangles show the cases when SFD performs better. The histograms are computed against noise variances in the test set. Although the two systems produce close overall accuracy, it is clear that the MD system copes better with stationary noise and the SFD system performs 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. In this section we present some initial effort and results.

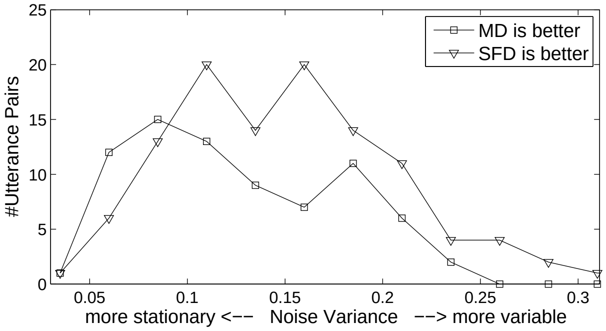

Figure 5: Histogram contours of utterance-level recognition error differences, showing different effects of noise stationarity on MD and SFD, SNR = 3 dB.

4.2. Method

In domestic settings and many other natural listening conditions the auditory scene can be approximately described as a slowly varying noise floor plus highly unpredictable acoustic 'events'. The idea of combining MD and SFD is to use the noise floor tracker to remove slowly varying noise and then employ SFD to deal with unpredictable acoustic events.

We separately generated soft MD masks (using the adaptive noise tracker) and fragments (using harmonicity based techniques). The SNR threshold for computing soft masks was optimised and the best results were obtained with a threshold of -3 dB. The T-F pixels with values < 0.5 in the SNR-based soft mask were then identified. These regions have low SNR estimates and the observations are most likely to have originated from some stationary noise sources. These T-F pixels were excluded from any fragments and were forced to be interpreted as part of the noise background during fragment decoding. The remaining fragments were employed by SFD as normal.

Fig. 6 illustrates this procedure. Fig. 6a is the cochleagram of a speech/noise mixture. The missing data mask derived from local SNR estimates is shown in Fig. 6b, where regions with soft value $<$ 0.5 are displayed in black. Fragments identified by harmonicity analysis are shown in Fig. 6c using different shades of grey. Fig. 6d shows the fragments used by the combined system, where regions in white have low SNR estimates and are forced into the background. The procedure is akin to using the missing data mask in Fig. 6b to filter the fragments in Fig. 6c.

Soft decisions were also employed in the combined system using Eq. 2. For the identified low SNR regions their corresponding soft MD mask values (< 0.5) were employed during decoding so that these pixels had the same contribution as in the MD system. For the rest T-F pixels their soft MD mask values

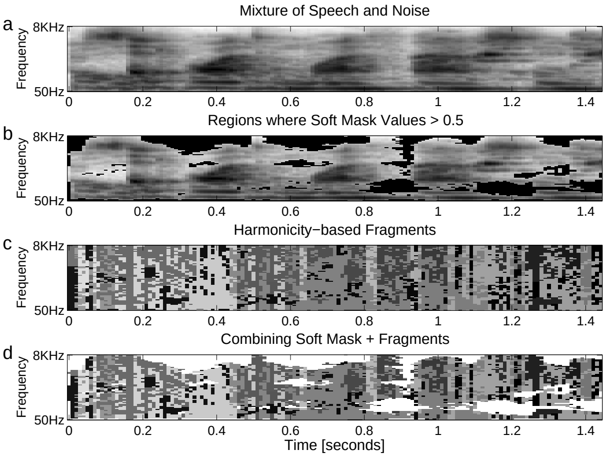

Figure 6: Combining soft missing data and speech fragment decoding techniques: a) Cochleagram of a speech/noise mixture; b) Missing data mask derived from local SNR estimates; c) Fragments identified by harmonicity analysis, represented as regions with different shades of grey; d) Fragments excluding low SNR regions (white).

($>$ 0.5) were not used. Instead we employed the confidence measures obtained from harmonicity analysis [17].

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

4.3. Results

Fig. 7 shows that the combined system exhibits improved performance over both standalone systems at SNRs below 6 dB. To further investigate the error reduction, we computed the number of utterances that produced keyword errors in different noise variance bands. Fig. 8 shows such histograms (contours) at 3 dB SNR for the MD system, the SFD system, and the combined system, respectively. Comparing the histograms of the MD system (squares) and the SFD system (triangles), we can again see that MD was prone to errors in more variable noise while SFD suffers more in stationary noise. The combined system (dashed line) improves over the SFD system by reducing keyword errors mostly for utterances with stationary noise (variance < 0.1), and the improvement over the MD system mainly comes from the cases in less stationary noise (variance > 0.1).

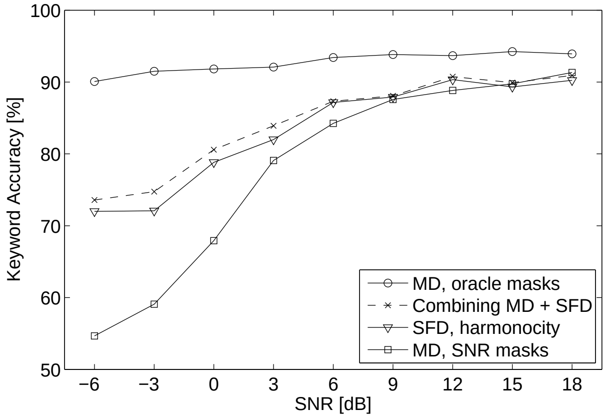

Figure 7: Results for the system combining soft MD and SFD in various noisy conditions.

For SNRs above 6 dB the combined system did not provide any significant improvement. A detailed error analysis (see Fig. 9) shows that the combined system actually reduced keyword errors for utterances with stationary noise over the SFD system alone. The improvement is, however, offset by increased errors for utterances with more variable noise, probably due to less accurate SNR estimation in these cases. An improved noise floor tracking component would help solve this issue.

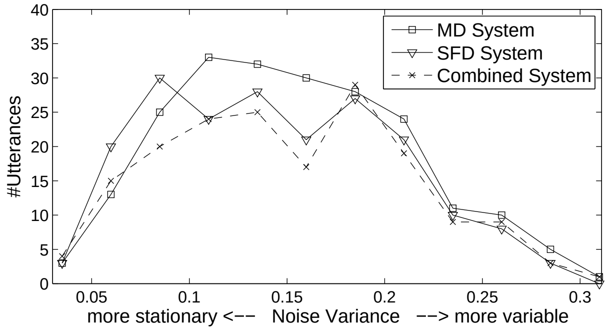

Figure 8: Histogram contours of utterances that produced keyword errors in different noise variance bands. SNR = 3 dB.

5. Discussion

We notice that in order to accommodate SNR estimation errors the soft MD system required an SNR threshold greater than 0 dB (9 dB was used) for computing soft MD masks. When the noise is less stationary the noise floor tends to be underestimated, causing many noise-dominated T-F pixels to be given high SNR estimates. By contrast the combined system did not need such a high SNR threshold (-3 dB was used). The regions with high local SNR estimates were divided into fragments, and the system has the ability to include it as part of the background if it is dominated by noise.

This paper has presented a relatively straightforward approach for combining a noise model-based approach and a source separation-based approach to robust speech recognition. Essentially the noise model is being allowed a first view of the data and then separation techniques are being reserved for el-

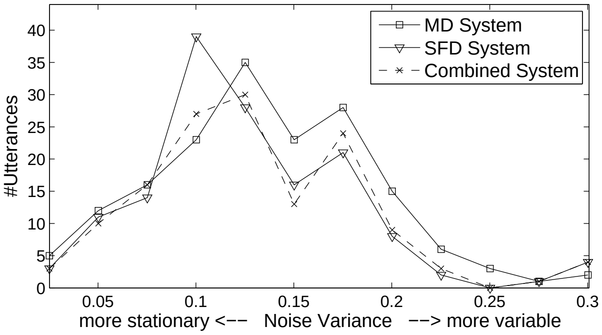

Figure 9: Histogram contours of utterances that produced keyword errors in different noise variance bands. SNR = 6 dB.

ements that are poorly predicted by the model. Although the combined system produces only small gains the work presents a firm baseline for investigating more sophisticated approaches:

Improved noise floor tracking

We have employed a simple frame-based approach for estimating the spectra of the noise floor. This relies on there being regular instances when no acoustic events are active. A more sophisticated estimator would be able to use spectro-temporal glimpses of the background in a more opportunistic manner. For example, performance might be improved using trackers that operate within frequency sub-bands.

Coupling noise floor estimation and fragment analysis

In the current system the noise tracking and fragment separation 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.

Statistical model combination

The work presented here employs an estimate of the mean noise floor spectra. However, if a reliable mean and variance could be estimated by the tracker, the noise floor model could potentially be combined with the target speech models within the fragment decoding framework in a more principled way.

6. Conclusions

This paper has presented a noise robust ASR system that combines aspects of the noise modelling and source separation approaches to the problem. The combined approach has been motivated by the observation that everyday listening noise backgrounds can be roughly characterised 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 energetic acoustic events are masked by the noise floor (represented by a soft missing data mask). Second, a fragment decoding system attempts to interpret the energetic 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. Although performance improvements are modest, we expect to be able to improve upon this baseline by using more sophisticated noise floor tracking approaches, exploiting noise floor variance estimates and introducing closer coupling of the noise floor estimation and fragment generation processes.

7. References

[1] P. Lockwood and J. Boudy, “Experiments with nonlinear spectral subtractor, Hidden Markov Models and the projection for robust speech recognition in cars,” *Speech Communication*, vol. 11, 1992.

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

[3] A. Varga and R. Moore, “Hidden Markov model decomposition of speech and noise,” in *Proc. IEEE ICASSP’90*, 1990, pp. 845–848.

[4] M. Gales and S. Young, “HMM recognition in noise using parallel model combination,” in *Proc. Eurospeech’93*, Berlin, 1993.

[5] 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’01*, Aalborg, Denmark, 2001, pp. 901–904.

[6] M. Cooke, J. Hershey, and S. Rennie, “Monaural speech separation and recognition challenge,” *Comput. Speech. Lang.*, vol. 24, no. 1, pp. 1–15, 2010.

[7] H. Christensen, J. Barker, N. Ma, and P. Green, “The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments,” in *Proc. Interspeech’10*, Makuhari, 2010.

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

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

[10] G. Brown and M. Cooke, “Computational auditory scene analysis,” *Comput. Speech. Lang.*, vol. 8, no. 4, pp. 297–336, 1994.

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

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

[13] 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’01*, Aalborg, 2001, pp. 213–216.

[14] J. Barker, M. Cooke, and D. Ellis, “Decoding speech in the presence of other sources,” *Speech Commun.*, vol. 45, no. 1, pp. 5–25, 2005.

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

[16] N. Ma, P. Green, and A. Coy, “Exploiting dendritic autocorrelogram structure to identify spectro-temporal regions dominated by a single sound source,” in *Proc. Interspeech’06*, Pittsburgh, PA, 2006, pp. 669–672.

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

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