Proceedings of the Institute of Acoustics

LINKING AUDITORY SCENE ANALYSIS AND ROBUST ASR BY MISSING DATA TECHNIQUES

Jon Barker Department of Computer Science, University of Sheffield, Sheffield, S1 4DP
Phil Green Department of Computer Science, University of Sheffield, Sheffield, S1 4DP
Martin Cooke Department of Computer Science, University of Sheffield, Sheffield, S1 4DP

1 INTRODUCTION

The 'missing data' approach to robust Automatic Speech Recognition assumes that when the speech is one of several sound sources, some spectral-temporal regions will remain uncorrupted and can be used as 'reliable evidence' for recognition. Identification of these regions can be thought of as placing a 'mask' over the spectral data. Arguments for the missing data premise can be found in [6] and are summarised in [10]. The authors have developed and applied techniques for adapting Continuous-Density Hidden Markov Model recognisers to the incomplete data case [6, 20, 11, 1, 2]. Several other groups have reported related work [17, 12, 16]. In this paper we briefly review the technique, and show how softening the decision about which regions are deemed to be reliable improves results. In the work reported up to this point, simple noise estimates have bee using to estimate local signal-to noise ratio, and this estimate has been used to define the masks. We now introduce harmonicity constraints as an additional cue (motivated by auditory scene analysis [3, 7, 5]) for identifying source-specific regions. Results are reported on the AURORA task [14]. We also discuss the idea of 'multi-source decoding', which generalises recognition algorithms for the case where initial grouping processes identify spectral temporal regions dominated by a single source but do not decide on the identity of the source.

2 AUTOMATIC SPEECH RECOGNITION WITH UNRELIABLE DATA

The classification problem in general is to assign an observation vector x to a class C. In the missing data case, a preceding process has partitioned x into reliable and unreliable parts, (x_r, x_u). Since the features x_u are unreliable it is not possible to obtain a reliable classification employing the likelihood $f(x|C)$. However, a marginal distribution can be formed by integrating over the possible values of the unreliable features and classification can proceed on the basis of the likelihood of the reliable features alone, $f'(x_r|C)$.

When considering noisy speech, the 'true' values of the unreliable data can be confined by *bounds* : if x is a spectral energy vector in which the unreliable channels are contaminated by additive noise, the speech energy in these channels must lie between 0 and the observed value x_u . This forms an additional constraint that can be applied by bounding the range over which the unreliable features are integrated.

In conventional Continuous Density Hidden Markov Model Speech Recognition, each chosen speech unit is represented by a trained HMM with a number of states. The states correspond to the classes of the last section. Each state is characterised by a multivariate mixture Gaussian distribution over the components of x, from an observation sequence X. The parameters of these distributions, together with state transition probabilities within models, are estimated in an EM fashion, commonly using the Baum-Welch algorithm. In our work, the models are trained on clean data: there is no *re-training for* noise conditions. A decoder (usually implementing the Viterbi algorithm) finds the state sequence having the highest probability of generating X. We show in [6] that in these conditions the *'bounded-marginal'* estimation of $f(x|C_i)$ can be written as:

$$f'(x|C_i) = \sum_{k=1}^{M} P(k|C_i) f(x_r|k, C_i) \int f(x_u|k, C_i) dx_u$$

Where the $P(k|C_i)$ are the M mixture coefficients for the distribution associated with C_i.

The first term in this equation is the marginal distribution over the reliable vector components. The integral term introduces constraints on the true values of the unreliable components. In the complete ignorance case it reduces to 1. In the bounded case it represents counter-evidence against the hypothesis of class C_i. For multivariate Gaussians, the integral required to evaluate the bounded marginal can be approximated by a difference of error functions.

3 IDENTIFYING RELIABLE AND UNRELIABLE DATA

With complete knowledge of the noise signal it would be possible to calculate the true local SNR at each time-frequency point in the spectral representation. Points with a high local SNR are dominated by signal energy and can be labelled as being reliable. Unfortunately, in practise we do not have access to the true local SNR. However, we can approximate the correct labelling by using a local SNR estimate. In previous work local SNR estimates have been obtained by averaging the noise spectrum over a short period in which there is no speech present. It is then hoped that the noise remains reasonably stationary over the duration of the utterance.

In realistic conditions, our local SNR estimate may be quite poor. Real noise is never totally stationary, and even stationary noise exhibits statistical variation around the mean energy estimate. A poor estimate of local SNR will lead to errors when labelling the data as reliable or unreliable. These errors are made concrete and irreversible when discrete labelling decisions are employed.

One approach to ameliorating the effects of the poor noise estimate is to 'soften' the reliable/unreliable decisions [2]. Rather than labelling each point with a 0 or 1, we use a continuous value in the range

[0.0, 1.0] which is interpreted in the missing data probability calculation as "the probability that the point is dominated by the speech signal".

For missing data with **discrete** masks, each component of the feature vector x is first classified as either reliable or unreliable. The contribution each feature makes to the likelihood of the observation, $f(x|C)$, will depend on how that feature is classified. Assuming an $f(x|C)$ where the components of x are independent:

$$\overline{f(x|C)} = \prod_{i\in r} f_i(x_i|C) \prod_{j\in u} \frac{1}{x_j} \int_0^{x_j} f_j(x_j|C)dx_j$$

With a mask containing **soft decisions** the probability due to each feature vector component becomes a weighted sum of the reliable and unreliable probability terms:

$$\overline{f(x|C)} = \prod_{i=1}^{N} (w_i f_i(x_i|k) + (1-w_i)\frac{1}{x_i} \int_0^{x_i} f_i(x_i|k)dx_i)$$

The above can be extended to a Gaussian mixture model in which the full distribution $f(x|C)$ is composed of a weighted sum of Gaussian distributions. Each individual Gaussian models the features independently, but the overall mixture distribution does not make this assumption.

4 SPEECH RECOGNITION EXPERIMENTS

4.1 Experimental Setup

The experiments reported here employ the AURORA 2000 speech recognition task [14]. This task is speaker independent recognition of digit sequences. All speech data is obtained from the TIdigits data base, downsampled to 8 kHz and filtered with a G71 characteristic.

Acoustic vectors were obtained via a 32 channel auditory filter bank [7] with centre frequencies spaced linearly in ERB-rate from 50 to 3750 Hz. The instantaneous Hilbert envelope at the output of each filter was smoothed with a first order filter with an 8 ms time constant, and sampled at a frame-rate of 10 ms. Finally, a cube root compression was applied to the frame of energy values.

Twelve whole word HMMs were trained ('1'-'9', 'oh', 'zero' and 'silence') with the 16 state topology suggested for the Aurora task. 7-component diagonal Gaussian mixture models were employed for each state. These models were trained using HTK [22] and data from the clean Aurora training set.

The first set of experiments evaluates the missing data approach using either discrete decisions or soft decisions. Recognition was performed using in-house software. Experiments were run using Aurora test set 'A', which consists of utterances with 4 different types of noise artificially added at several SNRs.

Generating Discrete Masks

First, an estimate of the local SNR is obtained by using the first 10 frames (preceding the speech) to derive a noise estimate for each frequency band. Specifically, the features are converted into the spectral amplitude domain, the first ten frames are averaged to form a stationary noise estimate and this estimate is subtracted from the noisy signal to form a clean signal estimate. The ratio of these two estimates forms the local SNR. We then label the feature as 'reliable' if the local SNR is greater than a threshold of 7 dB, otherwise it is labelled as 'unreliable'.

The 7 dB threshold has been empirically shown to be near optimal in previous work using different data and different noise types. By using a local SNR threshold very much greater than 0 dB, we accept labelling some reliable data as 'unreliable', in order to be more confident of the data that has been labelled as 'reliable'. This high threshold offers a safety margin that reduces the impact of the errors introduced by a poorly fitting stationary noise assumption.[1]

Generating Soft Masks

The values for the soft masks have been generated by compressing the local SNR with a sigmoid function with empirically derived parameters. The mapping is of the form:

$$f(x) = \frac{1}{1 + e^{-\alpha(x-\beta)}}$$

where α is the sigmoid slope, and β is the sigmoid center. For large values of α the sigmoid becomes steep and the resultant fuzzy mask approximates a discrete missing data mask. In this case we are implicitly assuming a small variance in the noise estimation error. At the other extreme, as the value of α tends to 0, we approach a mask where all values are 0.5. If $\alpha = 0$, we are assuming no knowledge of the noise and admitting maximum uncertainty into the mask.

Our use of fuzzy masks has parallels with the work of Renevey and Drygajlo in which a fuzzy mask is used in conjunction with missing data imputation [17].

Good recognition results are obtained using sigmoid parameter values around $\beta = 0.0$ and $\alpha = 3.0$. These values have been found empirically through recognition experiments employing a different set of utterances and a noise which is not one of those employed in the Aurora test set.

[1] Previous work has experimented with more advanced noise estimators [21].

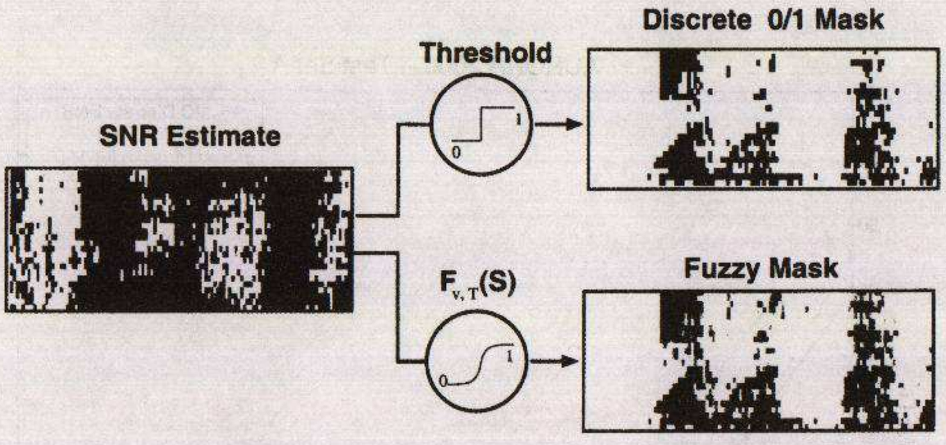

Figure 1: The difference between discrete and fuzzy missing data masks.

Whereas in the discrete mask the threshold is at 7 dB, for the fuzzy mask the sigmoid is centred around 0 db. When using discrete decisions much reliable data has to be discarded to avoid admitting incorrect points into the mask. In contrast, with the fuzzy interpretation, more points can be let through without the damage caused by admitting noise outweighing the benefit gained by the extra reliable information recovered.

4.2 Results

The curve labelled *'HTK clean training'* in Figure 2 shows the results obtained using the Aurora 2000 baseline system trained on clean data. This system employs MFCC features and is based on traditional ASR techniques. As expected its performance degrades rapidly with even modest amounts of noise. Word error rates (WER) are over 30% at 10 dB.

In comparison, results using spectral representations and missing data techniques with discrete decisions (*'MD Discrete SNR'*) do not degrade so rapidly. The WER at 10 dB is almost halved, with a reduction from over 30% to around 17%.

Applying soft decisions (*'MD Soft SNR'*) leads to a further performance gain. Relative to the system using discrete decisions, the WER at 15, 10 and 5 dB SNR is reduced by a further 40%. The 10 dB WER is now down to 10%. This extra robustness is not accompanied by a reduction in the clean speech performance.

The final line (*'HTK multi-condition'*), shows the performance of the Aurora baseline system when employing multi-condition training. This system has been trained using noisy data, with the noises employed in the training set matching those on which the system is tested. Unsurprisingly, this system outperforms the best missing data system. However, the multi-condition system is fitted to

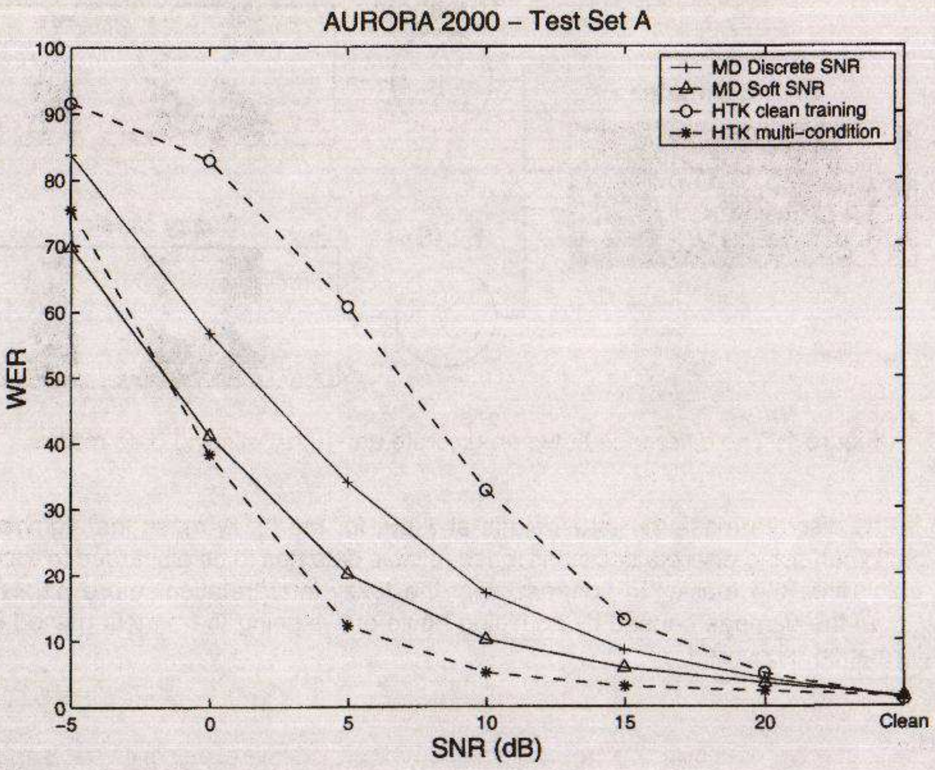

Figure 2: A comparison of discrete and soft missing data decisions: Word Error Rate as a function of global Signal to Noise Ratio.

the noises in the test set, and would not perform well if it were tested on different noise types. In contrast, the missing data system is trained only on clean speech, and should achieve a similar level of performance with a wide variety of noise types.

5 EMPLOYING PRIMITIVE CONSTRAINTS

Our earlier work on Computational Auditory Scene Analysis [7, 4, 5] provided the original motivation for the missing data approach. In 'primitive' CASA, low-level constraints which reach back to the physics of sound and the properties of the auditory system are used to group together spectral-temporal regions which are dominated by a single source. One grouping constraint is harmonicity: in voiced speech the energy will be organised around the harmonics of the fundamental frequency. Harmonic groups can be found using the autocorrelogram, a computational model of auditory pitch analysis [18]. Since harmonicity will do no grouping in unvoiced regions, it must be used in combination with SNR estimates.

5.1 Harmonicity Masks

The harmonicity masks are constructed from the autocorrelogram representation.[2]

Figure 3 shows the steps taken to compute the autocorrelogram and the summary autocorrelogram. The initial filterbank uses the same set of auditory filters as used for computing the speech features described in the Section 4.1. The mechanical to neural transduction of the inner hair cells is modelled by half-wave rectification followed by square root compression. The autocorrelations are computed using a modified short-time autocorrelation function of the form:

$$\hat{R}_n(k) = \sum_{m=-\infty}^{\infty} x(n+m)\hat{w}_1(m)x(n+m+k)\hat{w}_2(m+k) \tag{1}$$

where the window $\hat{w}_2$ is chosen to include samples outside the nonzero interval of window $\hat{w}_1$, i.e.:

$$\begin{aligned} \hat{w}_1 &= 1 \quad 0 \le m \le N-1 \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{2}$$

and

$$\begin{aligned} \hat{w}_2 &= 1 \quad 0 \le m \le N-1+K \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{3}$$

where K is the greatest lag of interest. This definition counters the problem of the amplitude falling-off as the overlap of the nonzero intervals becomes smaller [15]. In our experiments best results were achieved with N equal to 300 samples, and by employing a Hanning window for $\hat{w}_1$. The autocorrelation was computed with a maximum lag of 150 samples (corresponding to a minimum pitch frequency of about 50Hz).

The autocorrelogram was normalised by dividing each channel by the value of that channel at 0 lag. Hence the values contained are all in the range 0 to 1.

The procedure for computing a frame of the harmonicity mask from the autocorrelelogram was

1. Sum over the frequency channels to compute a summary autocorrelogram.

[2] Our use of the autocorrelogram has parallels with that of Glotin and Berthommier who employ a similar measure of harmonicity to derive an estimate of subband SNR [9].

2. Find the lag of the largest peak in the summary autocorrelogram.
3. Take a slice through the autocorrelogram at this lag.
4. Rescale the slice using a sigmoid function to obtain a frame of the soft harmonicity mask.

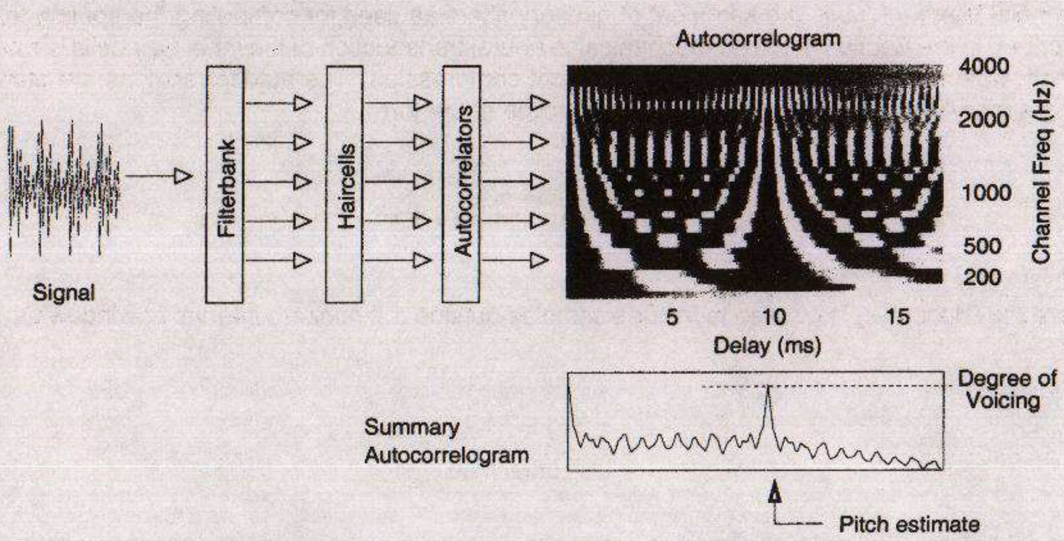

Figure 3: Computing the autocorrelogram (adapted from Summerfield and Culling [18]).

5.2 Combining Harmonicity and SNR Masks

The harmonicity mask is only valid during regions of the signal that are dominated by a harmonic source. During non-harmonic regions the temporal gaps in the mask can be filled in with frames taken from the SNR mask. So, in the simplest scheme, the combined mask is created by selecting from either the harmonicity mask or the SNR mask according to a discrete voicing decision. This voicing decision can be obtained by comparing the height of the largest summary autocorrelogram peak, V, to a fixed threshold, V_0. However, better results can be obtained by replacing the discrete decision with a soft weighting. In this case the voicing parameter, V, is passed through a sigmoid function to produce a weight, w, and then a combined mask is constructed from a weighted sum of the harmonicity mask and the SNR mask. This is illustrated in figure 4.

5.3 Experimental Setup

An experiment was conducted to compare the performance of the combined harmonicity/SNR masks against that of the SNR masks tested in the previous section. As before, the experiments employed

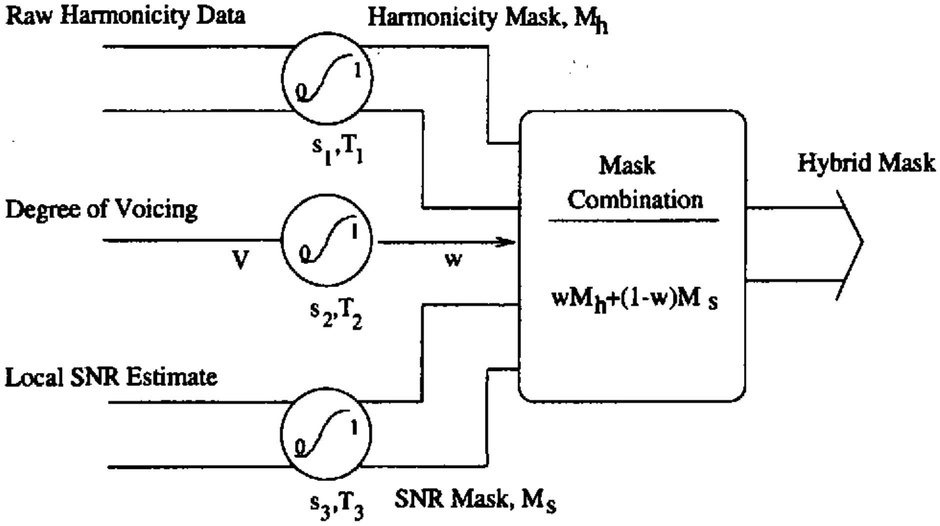

Figure 4: Combining Harmonicity and SNR Mask

Aurora II. The same 32-channel auditory filter bank representation was used, along with the HMM models trained on the clean Aurora training data.

The harmonicity masks were generated as described in Section 5.1. The slices through the autocorrelograms were rescaled using a sigmoid with coefficients $\alpha = 30$ and $\beta = 0.5$. These values were determined empirically through pilot experiments using a different set of test data and a different noise type to any of those in Aurora test set 'A'. The harmonicity mask was combined with the soft SNR masks from the previous experiment. The mixing weight, w, was produced by passing the voicing parameter V, through a sigmoid with coefficients $\alpha = 60$ and $\beta = 0.78$. These values were taken from an analysis of the distribution of V as measured over each frame of the clean training data.[3]

5.4 Results

The line *'MD Soft SNR/Harmonicity'* in Figure 5 shows the Aurora test set 'A' results obtained with the combined SNR and Harmonicity-based mask. There is a small but consistent improvement at all SNRs relative to the mask based on SNR alone (*'MD Soft SNR'*). The results shown here are the average across the 4 test set 'A' noise types. Roughly equal improvements were seen in the results for each noise taken individually.

[3] For clean speech it is observed that the voicing parameter, V, is strongly bimodal, with clear fairly well separated modes for regions of voiced and unvoiced speech. A sigmoid which maps V to w can be constructed by placing the centre to lie between these two modes, and adjusting the slope to reflect their variance.

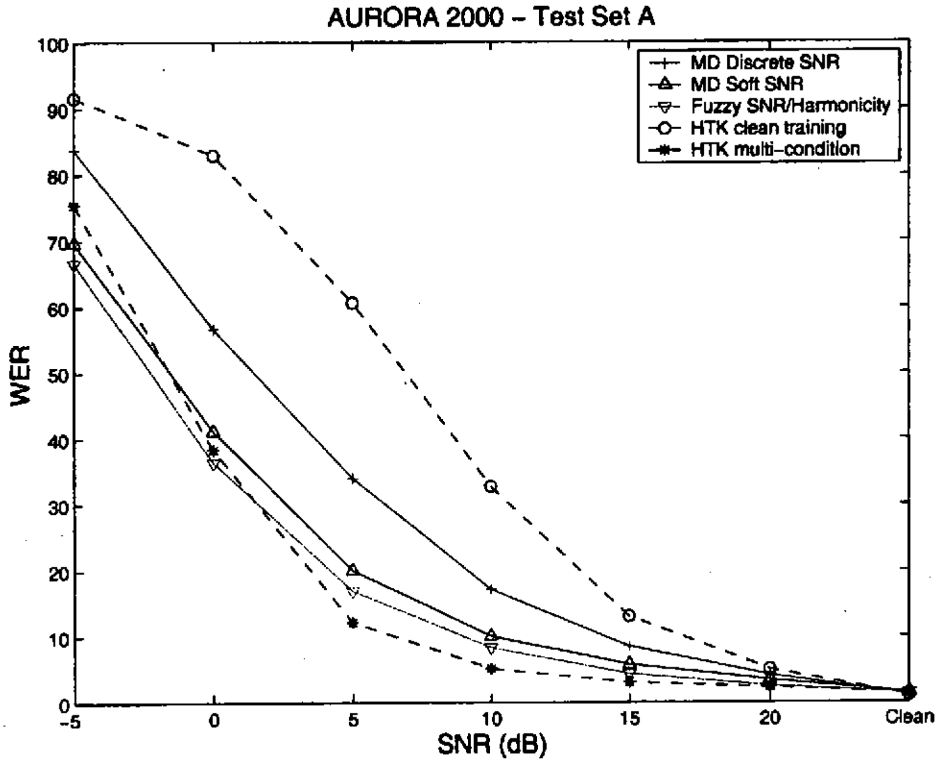

Figure 5: Results obtained using the combined SNR/Harmonicity-based soft decisions: Word Error Rate as a function of global Signal to Noise Ratio.

The *'MD Soft SNR'* and the *'MD Soft SNR/Harmonicity'* use exactly the same set of clean speech models and exactly the same missing data recognition technique. The only difference is in the quality of the soft data reliability estimate. This improvement in the data reliability estimate has almost halved the remaining gap between the missing data and multicondition training techniques at 5 and 10 dB SNR, and makes missing data superior to multicondition training at 0 dB SNR and below.

The technique's employed to gain the improvements illustrated in Figure 5 are very simplistic. For instance, no use is made of pitch information: it is assumed that all harmonic energy is due to the speech source, regardless of whether or not it has a pitch in the range of voiced speech; there is assumed to be a single harmonic source regardless of jumps in the pitch that would indicate two competing sources. That consistent and significant recognition performance improvements can be made despite these obvious deficiencies is extremely encouraging.

6 APPLYING TOP-DOWN CONSTRAINTS: MULTI-SOURCE DECODING

The technique for employing harmonic constraints described in the previous section makes the assumption that the speech source is the dominant harmonic sound source. In general this is not the case, and this assumption will lead to harmonic components in the noise background being incorrectly labelled as reliable speech. This is symptomatic of a more general problem - although primitive grouping constraints may be able to locate spectral-temporal fragments that are dominated by a common source, it is not known a-priori which of these fragments belongs to the speech source in particular.

A solution to the problem of selecting the correct subset of fragments is to let the selection be governed by the models (i.e. to use 'top-down' information). This can be implemented by employing a 'multi-source' decoder. In contrast to conventional decoding, where all the observations are assumed to belong to the source being recognized, the task of the multi-source decoder is to determine the most likely model state sequence at the same time as deciding which observations to use, and which to ignore as 'background'. We assume that we have models for the speech source, but in contrast with approaches such as Parallel Model Combination [8] and HMM decomposition [19] we do not require models for the acoustic background.

In essence, the multi-source decoder attempts to recognise speech from a set of evidence fragments by evaluating every possible combination of fragments over an entire utterance. Unfortunately, there are 2^N subsets of N fragments, and N could typically become rather large. However, many of these subset are very similar (e.g. many pairs differ only by the inclusion of a single fragment). By careful arrangement of the computation, the computational complexity can be reduced so that it scales with 2^M where M is the maximum number of *simultaneous* fragments. This is tractable if primitive grouping processes deliver evidence fragments above some minimum granularity, say over some tens of milliseconds duration. Crucially, although N increases with utterance length, M remains essentially constant. For further details see [1].

Pilot experiments using this approach have delivered promising results [1]. These experiments used simulated fragments derived from a fairly ad-hoc dissection of the local SNR-based missing data mask. Work is now planned to repeat these experiments using information extracted from the autocorrelogram representation to produce more meaningful fragments.

7 CONCLUSION

Most work on robust speech recognition has been based on the idea of 'reducing the mismatch' between training and test conditions. This leads to the use of noise models and their deployment in techniques such as Spectral Subtraction [13], HMM decomposition [19] and Parallel Model Combination [8]. If some predictable noise source is present, then it is appropriate to use its statistics in this way, and we have indeed made use of simple noise models in defining our 'missing data masks'. However, if speech is to be recognised within an arbitrary 'auditory scene', such as in a street or at a meeting, the sound sources will not be pre-determined, and they will change in location and with time. For this general case, missing data coupled with multi-source decoding has a number of attracticns:

- The first stage, the identification of reliable evidence fragments, can be based on processing which exploits only the predictable noise components and low-level constraints such as harmonicity and common onset/offset. There is no need, for instance, to make any assumption about how many sources are present.
- The reliable evidence decision does not have to be right all the time, because it can be expressed probabilistically rather than in a discrete way.
- Multisource decoding provides a way in which primitive processing can interact with schema-driven processing, so that the initial grouping stage does not have the responsibility of deciding what belongs to what source.

The scheme we have presented is, arguably, both a competitive robust ASR system and a computational implementation of a psycho-acoustic model.

Acknowledgements

This research is supported by the EC Transfer and Mobility of Researchers network SPHEAR, the EC ESPRIT long term research project RESPITE and Motorola. Thanks to Ljubomir Josifovski who has contributed greatly to the ideas in this paper, and to Guy Brown for his help with the implementation of the harmonicity model.

References

[1] J.P. Barker, M.P. Cooke, and D.P.W. Ellis. Decoding speech in the presence of other sound sources. In *Proc. ICSLP '00*, volume 4, pages 270–273, Beijing, China, October 2000.

[2] J.P. Barker, L. Josifovski, M.P. Cooke, and P. Green. Soft decisions in missing data techniques for robust automatic speech recognition. In *Proc. ICSLP '00*, volume 1, pages 373–376, Beijing, China, October 2000.

[3] A.S. Bregman. *Auditory Scene Analysis*. MIT Press, 1990.

[4] G.J. Brown. *Computational auditory scene analysis: A representational approach*. PhD thesis, Department of Computer Science, University of Sheffield, 1992.

[5] G.J. Brown and M. Cooke. Computational auditory scene analysis. *Computer speech and language*, 8:297–336, 1994.

[6] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, June 1999. In press.

[7] M.P. Cooke. *Modelling auditory processing and organisation.* PhD thesis, Department of Computer Science, University of Sheffield, 1991.

[8] M. J. F. Gales and S. J. Young. HMM recognition in noise using parallel model combination. In *Eurospeech'93*, volume 2, pages 837–840, 1993.

[9] H. Glotin and F. Berthommier. Test of several external posterior weighting functions for multiband Full Combination ASR. In *Proc. ICSLP '00*, Beijing, China, October 2000.

[10] P.D. Green, J. Barker, M.P. Cooke, and L. Josifovski. Test of several external posterior weighting functions for multiband Full Combination ASR. In *Proc. AI and Statistics*, pages 49–56, Key West, FA, 2001.

[11] L. Josifovski, M. Cooke, P. Green, and A. Vizinho. State based imputation of missing data for robust speech recognition and speech enhancement. In *Eurospeech'99*, volume 6, pages 2837–2840, sep 1999.

[12] R. P. Lippmann and B. A. Carlson. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise. In *Eurospeech'97*, pages 37–40, 1997.

[13] P. Lockwood and J. Boudy. Experiments with a non-linear spectral subtractor (NSS) Hidden Markov Models and the projection, for robust speech recognition in cars. In *Eurospeech'91*, volume 1, pages 79–82, 1991.

[14] D. Pearce and H.-G. Hirsch. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ICSLP '00*, volume 4, pages 29–32, Beijing, China, October 2000.

[15] L.R. Rabiner and R.W. Schafer. *Digital processing of speech signals.* Prentice-Hall, London, 1978.

[16] B. Raj, M. Seltzer, and R. Stern. Reconstruction of damaged spectrographic features for robust speech recognition. In *Proc. ICSLP '00*, volume 1, pages 357–360, Beijing, China, October 2000.

[17] P. Renevey and A. Drygajlo. Introduction of a reliability measure in missing data approach for robust speech recogntion. In *Proc. EUSPICO'2000*, 2000.

[18] Q. Summerfield and J. F. Culling. Auditory computations that separate speech from competeting sounds: a comparison of monaural and binaural processes. In Keller, editor, *Fundamentals of speech synthesis and speech recognition.* J.Wiley and Sons, Chichester, 1994.

[19] A. P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *ICASSP'90*, pages 845–848, 1990.

[20] A. Vizinho, P. Green, M. Cooke, and L. Josifovski. Missing data theory, spectral subtraction and signal–to–noise estimation for robust ASR: an integrated study. In *Eurospeech'99*, pages 2407–2410, 1999.

[21] A. Vizinho, P. D. Green, M. P. Cooke, and L. Josifovski. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study. In *Proceedings of EuroSpeech'99*, pages 2407–2410, Budapest, 1999.

[22] S. J. Young and P. C. Woodland. *HTK Version 1.5: User, reference and programmer manual.* CUED, Speech Group, 1993.