Robust ASR Based On Clean Speech Models: An Evaluation of Missing Data Techniques For Connected Digit Recognition in Noise

Jon Barker, Martin Cooke, Phil Green

Department of Computer Science,
University of Sheffield, Sheffield, UK
{j.barker, m.cooke, p.green@dcs.shef.ac.uk}

Abstract

In this study, techniques for classification with missing or unreliable data are applied to the problem of noise-robustness in Automatic Speech Recognition (ASR). The techniques described make minimal assumptions about any noise background and rely instead on what is known about clean speech. A system is evaluated using the Aurora 2 connected digit recognition task. Using models trained on clean speech we obtain a 65% relative improvement over the Aurora clean training baseline system, a performance comparable with the Aurora baseline for multicondition training.

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 [5] and are summarised, with updated results, in [7].

The authors have developed and applied techniques for adapting Continuous-Density Hidden Markov Model recognisers to the incomplete data case [5, 13, 8, 1, 3, 2]. In brief, the likelihood estimation for an observation with some reliable and some unreliable components will be generated from a given distribution using the marginal distribution over the reliable components and integrating over the range of possible values for the unreliable components. For spectral energies, the true speech energy must lie between zero and the observed energy of the speech/noise mixture. Several other groups have reported related work [11, 10]. In this paper we briefly review our previously reported work. We then discuss a recent extension to our previous systems; specifically, gender dependent modelling. Finally, we report results of an evaluation of our system based on the Aurora 2 connected digit recognition task[9].

2. Identifying Reliable Speech Data

The missing-data technique operates on data in which the features have been labelled as either reliable or unreliable. This section describes two contrasting approaches for generating such labellings that have been used both alone and in combination in our current system.

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 Guy Brown whose ideas contributed greatly to the design of the harmonicity masks and to Ljubomir Josifovski who has been heavily involved in much of this work. Thanks also to Andrew Morris who has lent much mathematical insight.

2.1. Local SNR Estimates

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 can be labelled as being reliable. 'Oracle' masks derived from the true local SNR provide robust recognition approaching human levels of performance. In practice we do not have access to the true local SNR. However, the correct labelling can be approximated via 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. This approach assumes 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. We therefore 'soften' the reliable/unreliable decisions. Rather than labelling each point with a 0 or 1, a continuous value in the range [0.0, 1.0] is employed which is interpreted in the missing data probability calculation as "the probability that the point is dominated by the speech signal". Soft reliability decisions can be employed with a small extension to the equations employed for discrete decisions (see [3] for details).

2.2. Employing Voiced Speech Cues

Our earlier work on Computational Auditory Scene Analysis [6, 4] provided the original motivation for the missing data approach. In 'primitive' CASA, low-level constraints arising from 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 such 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 [12]. Having identified a harmonic group, some decision process is needed to decide whether it represents part of the speech source or part of the noise. We are currently investigating top down approaches to this problem (see [1]), but the work described in this paper takes the simplistic approach of attributing all harmonic groups to the speech source. This approach is quite reasonable at higher

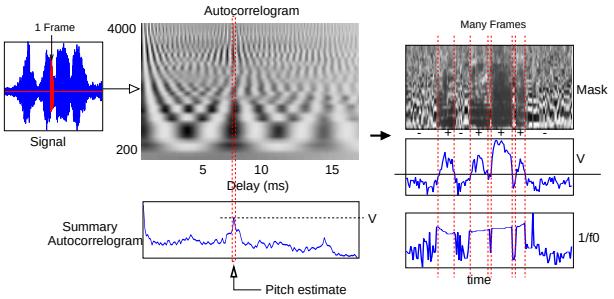

Figure 1: *The computation of the Harmonicity Mask. For each frame of speech a correlogram and a summary autocorrelogram are computed. A slice is then taken through the autocorrelogram at the lag of the biggest peak in the summary. This slice is rescaled using a sigmoid compression and then entered as a frame in the mask. A voicing decision and a pitch track can also be extracted from the representation.*

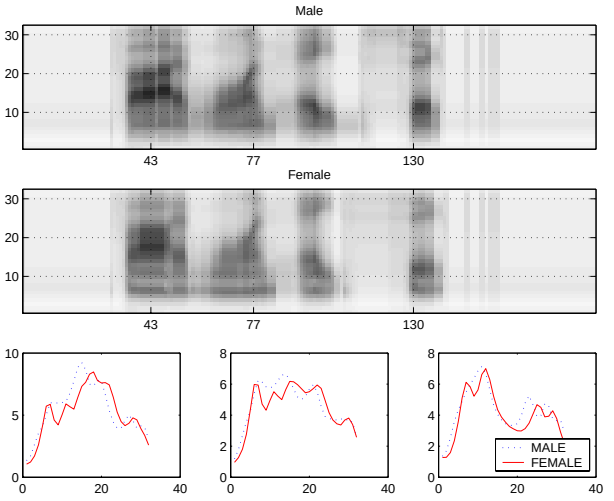

Figure 2: *Comparison of reconstructions of the utterance 'five one eight six' as spoken by 'prototypical male' and 'prototypical female' speakers (see text for details). The lower panels display spectral slices taken from the reconstructions at the points indicated by the x-axis tick marks.*

global SNRs where speech is usually the dominant source of voicing.

The harmonicity mask is constructed from the autocorrelogram representation[12]. An outline of the steps taken to produce the mask is shown Figure 1. For full details see [2].

The harmonicity mask is only valid during regions of the signal that are dominated by voiced speech. During unvoiced regions the harmonicity mask is invalid, and the SNR mask is a better indicator of reliability. The two masks can be effectively combined by interpolating between them according to the degree of voicing (see Section 4.3). An estimate of the degree of voicing can be extracted from the summary autocorrelogram (see Figure 1).

3. Gender Dependent Modelling

Missing data speech recognition is constrained to operate on spectrally-based features, because orthogonalising transforms spread unreliability over all features. Spectral features are particularly sensitive to gender differences. This is illustrated in Figure 2 which has been constructed by taking a clean speech utterance and force-aligning it to single-component Gaussian models trained on either male speakers or female speakers. By outputting the means of each model state in the alignment it is possible to effectively reconstruct the spectral representation as if the speech had been spoken by either the prototypical male or prototypical female speaker. This technique allows us to clearly see the differences in what the models have learnt about male and female speech. The following differences are observed: i) The female formants are slightly higher in frequency - this small frequency shift has a large effect on the mean values of every feature (its impact on a cepstral representation is probably less severe); ii) for female speech, harmonics are being resolved in the F1 region - this means that irrelevant F0 differences will add to the variance of the female features in this region; iii) the 4th formant is visible for male speech but not for the female speech.

There are two common techniques for dealing with gender specific differences i) gender dependent modelling, ii) vocal tract normalisation. We have employed the former, because vocal tract normalisation, although having many advantages (e.g. it requires adding fewer parameters to the system) will only compensate for the first of the three sources of gender difference listed above.

In our gender dependent system, models are trained separately on the male and female training data. At recognition time, a grammar is used to constrain the recognition hypotheses to select a sequence of male or female models. Note that the male/female decision is not explicitly imposed during recognition, since the Viterbi algorithm selects the path with the greatest likelihood.

4. Testing on Aurora 2 Data

The experiments reported here employ the Aurora 2 speaker independent connected digit recognition task [9].

Acoustic vectors were obtained via a 32 channel auditory filter bank [6] 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. The features were supplemented with their temporal derivatives, to form a 64 dimensional feature vector.

Whole word digit models were trained using the Aurora clean speech training set. The Aurora model topology and training regime has been adhered to, except our system employs 7-component mixture models rather than the 3-component mixtures suggested. Missing data techniques, as explained above, are constrained to operate with spectral features. The features they employ do not have the near independence of the cepstral features on which most traditional ASR systems are based. This means that diagonal Gaussian components are not themselves a good fit to the data distribution, and a greater number of components is required to produce good quality acoustic models.

To reduce the number of digit insertions, a grammar was used to constrain all hypotheses to start and end with the silence model.

Four system variants were evaluated. The first three differ only in their construction of the missing data mask: i) a system employing a discrete SNR based mask, ii) a system employing a soft SNR-based mask, iii) a system employing a combined SNR and harmonicity mask. The final system uses the com-

bined mask as in iii), but in conjunction with a set of gender dependent models. Some details of the mask construction are given in the sections that follow.

4.1. Discrete SNR Masks

The discrete SNR masks are based on an estimate of local SNR. Specifically, the filter bank 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. Features are labelled 'reliable' if they have a local SNR greater than a threshold of 7 dB, otherwise they are 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, the system accepts 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.

4.2. Soft SNR 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 centre.

Best recognition results are obtained using sigmoid parameter values around $\beta = 0.0$ (i.e. SNR 0 dB) 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.

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.

4.3. Combined Harmonicity and SNR Masks

Harmonicity masks were generated as illustrated in Figure 1. Frames of the mask are constructed from slices taken from the autocorrelograms at the lag of the largest peak in the summary autocorrelogram. These autocorrelogram slices are rescaled using a sigmoid with coefficients $\alpha = 30$ and $\beta = 0.5$. Again, the sigmoid coefficients were determined empirically through pilot experiments using a different set of test data and a different noise type to any of those in the Aurora data. The combined mask, M_c, is then determined by linearly interpolating between the harmonicity mask, M_h, and the soft SNR mask, M_s, according to a time varying weight based on the degree of voicing parameter V.

$$M_c = f(V)M_h + (1 - f(V))M_s$$

The function $f()$ is a sigmoid with parameters $\alpha = 60$ 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 (see [2] for further details).

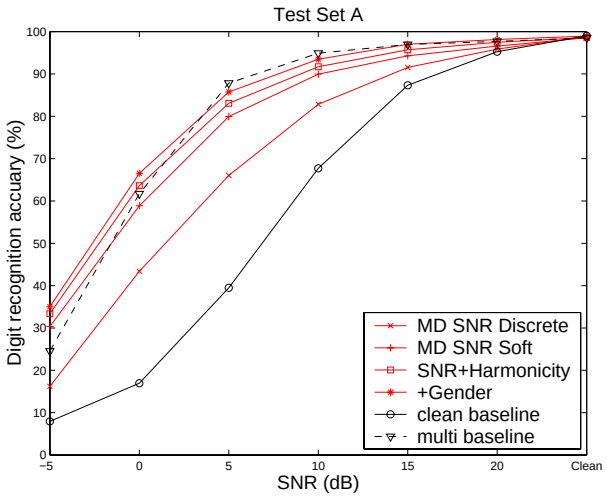

Figure 3: *Results for Test Set A comparing all four missing data system variants with the two Aurora baseline systems.*

5. Results and Conclusions

All of the missing data systems were tested on Aurora test set A, and results are shown in Figure 3. The best performing system is that which employs the combined harmonicity and SNR-based masks and uses gender dependent modelling. This system was also tested with Aurora tests sets B (Figure 4) and C. Results on all three test sets are summarised in Table 1 which also shows relative improvements compared to the Aurora clean training baseline system. The missing data system exhibits overall relative improvements of 69.4%, 68.5% and 46.2% for test sets A, B and C [1] respectively.

As can be seen from Figures 3,4, and 5, the missing data system, although trained on *clean speech* alone, performs on a par with the Aurora baseline for multi-condition training.

Comparison of the performace of the various missing data systems tested (Figure 3) shows the large gains in performance that can be made by increasing the quality of the feature reliability estimates (i.e. progressing from discrete SNR masks, to soft SNR masks, on to combined SNR + Harmonicity masks). Encouragingly, the techniques employed in this paper are all reletively crude and offer room for further improvement.

6. 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, M.P. Cooke, and P.D. Green. Linking auditory scene analysis and robust ASR by missing data techniques. In *Proc. WISP '01*, pages 295–307, Stratford, UK, April 2001.

[3] J.P. Barker, L. Josifovski, M.P. Cooke, and P. Green. Soft decisions in missing data techniques for robust automatic

[1] The intention of test set C is the consideration of mismatched filters (MIRS for the test data, G712 for the training data). Our current system makes no attempt to address the problems caused by this type of spectral distortion.

Aurora 2 Clean Training - Results															
	A					B					C				%
	Sub.	Bab.	Car	Exhib.	Ave.	Rest.	Street	Air.	Stat.	Ave.	Sub. M	Str. M	Ave.	Overall	Improv.
Clean	99.17	98.85	98.78	98.92	**98.93**	99.17	98.85	98.78	98.92	**98.93**	98.05	98.16	**98.11**	**98.76**	**-29.52%**
20 dB	98.13	98.16	98.30	98.03	**98.15**	98.22	97.22	98.30	97.81	**97.89**	96.07	95.95	**96.01**	**97.62**	**51.24%**
15 dB	96.29	97.37	97.17	97.10	**96.98**	96.47	95.47	97.44	96.58	**96.49**	93.80	93.41	**93.60**	**96.11**	**67.55%**
10 dB	92.05	93.47	94.33	94.14	**93.50**	92.05	90.78	93.74	92.07	**92.16**	89.19	87.30	**88.25**	**91.91**	**73.82%**
5 dB	83.11	84.25	88.10	87.54	**85.75**	80.96	81.86	85.48	82.66	**82.74**	79.43	75.12	**77.28**	**82.85**	**70.94%**
0 dB	64.81	58.59	69.37	73.31	**66.52**	54.96	61.76	64.03	62.94	**60.92**	57.26	50.18	**53.72**	**61.72**	**53.54%**
-5 dB	34.36	25.51	34.00	46.50	**35.09**	23.00	32.01	31.52	32.52	**29.76**	27.79	22.10	**24.94**	**30.93**	**24.48%**
Average	86.88	86.37	89.45	90.02	**88.18**	84.53	85.42	87.80	86.41	**86.04**	83.15	80.39	**81.77**	**86.04**	
% Imp.	**56.99%**	**72.79%**	**73.23%**	**71.17%**	**69.42%**	**67.37%**	**62.10%**	**73.90%**	**69.37%**	**68.45%**	**50.20%**	**42.13%**	**46.16%**		**65.05%**

Table 1: *Summary of results for the combined-mask gender-dependent missing data system. Results are presented in terms of word recognition accuracy, and relative improvements over the Aurora Clean Training baseline are also shown.*

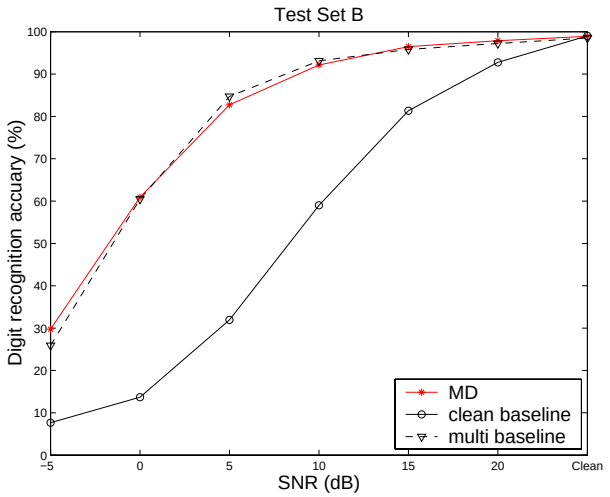

Figure 4: *Results for Test Set B comparing the combined-mask gender-dependent missing data system with the two Aurora baseline systems.*

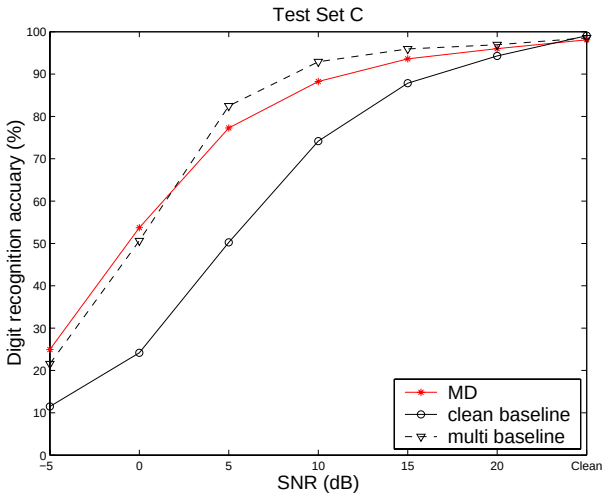

Figure 5: *Results for Test Set C comparing the combined-mask gender-dependent missing data system with the two Aurora baseline systems.*

speech recognition. In *Proc. ICSLP '00*, volume 1, pages 373–376, Beijing, China, October 2000.

[4] G.J. Brown and M. Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 8:297–336, 1994.

[5] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285, 2001.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.