MISSING DATA TECHNIQUES FOR ROBUST SPEECH RECOGNITION

Martin Cooke, Andrew Morris & Phil Green
{m.cooke,a.morris,p.green}@dcs.shef.ac.uk
Computer Science, University of Sheffield, Regent Court, 211, Portobello Street, Sheffield, UK

ABSTRACT

In noisy listening conditions, the information available on which to base speech recognition decisions is necessarily incomplete: some spectro-temporal regions are dominated by other sources. We report on the application of a variety of techniques for missing data in speech recognition. These techniques may be based on marginal distributions or on reconstruction of missing parts of the spectrum. Application of these ideas in the Resource Management task shows performance which is robust to random removal of up to 80% of the frequency channels, but falls off rapidly with deletions which more realistically simulate masked speech. We report on a vowel classification experiment designed to isolate some of the RM problems for more detailed exploration. The results of this experiment confirm the general superiority of marginals-based schemes, demonstrate the viability of shared covariance statistics, and suggest several ways in which performance improvements on the larger task may be obtained.

1. BACKGROUND

The missing data problem arises naturally in many pattern recognition tasks [2,8] where elements of data vectors to be classified are unavailable during training and/or recognition. The causes of incomplete evidence include unreliable sensors, band-restricted data transmission (e.g. the spectral filtering action of a telephone channel), or partial occlusion of the desired pattern by an interfering signal. In the latter case, it is assumed that some preprocessor is able to determine which parts of the mixed observation correspond to the source to be classified.

Our motivation for studying the missing data problem derives from ongoing studies at Sheffield and elsewhere [1] on computational auditory scene analysis (CASA), in which evidence for different sound sources is separated using auditory grouping principles. CASA is an attractive paradigm for robust ASR. It makes no assumptions about the type and number of acoustic sources which make up the mixture, and does not require prior exposure to these sources. However, separation will never be able to recover *all* the evidence: there will be some regions where other sound sources dominate. CASA-based robust ASR requires that the resulting missing data problem be confronted.

In previous work [4,9] we demonstrated that it is possible to remove high proportions (up to 90%) of the input spectrum without significant deterioration in recognition rates. In ICASSP-95, we reported (using NOISEX) noise tolerance comparable to that of human listeners when only those spectro-temporal regions with a favourable local SNR were retained. Subsequently, we have applied missing data techniques to the Resource Management (RM) task [5]. The main results of that study are outlined in section 3 of this paper. The RM experiments highlight a number of outstanding problems with the practical application of missing data ideas. Here, we address these issues with a more focussed problem, that of TIMIT vowel identification using a Gaussian classifier (section 4). This task allows for a comparison of missing data techniques which would have been computationally infeasible on RM, and decouples the observation probability estimation problem from the problem of finding the best model sequence.

2. MISSING DATA TECHNIQUES FOR MULTIVARIATE GAUSSIAN DISTRIBUTIONS

Missing components of pattern vectors can either be estimated or ignored. Estimates assume an importance in situations where reconstruction of the data vector is required, possibly for further processing (e.g. further pattern transformation prior to classification), or for regeneration (e.g. resynthesis). Ignoring missing data means attempting to classify the observation solely on the basis of the information present. It has been argued [2] that it can be inappropriate to replace missing values with *any* estimate.

Both kinds of approach benefit from some model for the process giving rise to the observations. Here, we assume that the observation vector x belongs to one of a number of classes, each of which is modelled as a mixture of K multivariate Gaussian distributions:

$$f(x|\, S_j) = \sum_{i=1}^{K} c_{ij}\phi(x, \mu_{ij}, C_{ij}) \quad (1)$$

S_j represents model j (or, e.g., an emitting state of an HMM for class j), c_{ij} is the weight of mixture i for model j, and ϕ is the n-dimensional Gaussian distribution (mean μ, covariance C)

$$\phi(x, \mu, C) = \frac{1}{(2\pi)^{0.5n}|C|^{0.5}} exp(-0.5\,(x-\mu)^t C^{-1}(x-\mu)) \quad (2)$$

The missing data problem for pattern classification is the computation of $f(x|\, S_j)$ for an incomplete vector x. It will be convenient to re-order x as $x = (x_p x_m)$, where x_p and x_m represent, respectively, the subvectors of present and missing components. To simplify things further, we will drop model subscript j and present the required formulae for the single mixture condition. All arguments presented here are applicable to the multiple-mixture case.

The mean and covariance matrix are similarly partitioned:

$$\mu = (\mu_p \mu_m)\,,\; C = \begin{bmatrix} C_{pp} & C_{pm} \\ C_{mp} & C_{mm} \end{bmatrix} \quad (3)$$

One simple estimation technique is to replace missing values by *unconditional* model means (so-called *mean imputation*) i.e.

$$x_m = \mu_m \quad (4)$$

This approach makes no use of present components and hence cannot exploit information in the covariance. An alternative is to calculate model means *conditioned* on those components present. For multivariate Gaussians, this conditional distribution is also Gaussian [12], with mean and covariance:

$$x_{m|\,p} = \mu_m + C_{pm}^t C_{pp}^{-1}(x_p - \mu_p) \quad (5)$$

$$C_{m|\,p} = C_{mm} - C_{pm}^t C_{pp}^{-1} C_{pm} \quad (6)$$

For unconditional and conditional mean replacement techniques, classification proceeds by computing

$$f((x_p \mu_m) \mid \mu; C) \quad (7)$$

$$\text{or } f((x_p x_{m\mid p}) \mid \mu; C) \quad (8)$$

respectively – that is, the probability of the reconstructed observation conditioned on the model. The conditional covariance can be used to indicate the certainty of each missing component.

In contrast, classification on the basis of x_p alone can be achieved using the marginal distribution, $f(x_p)$, which is particularly simple for multivariate Gaussians, since any marginal of a Gaussian is itself Gaussian. In our case, we estimate

$$f(x_p \mid \mu_p; C_{pp}) \quad (9)$$

Missing data techniques represented by (7), (8) and (9) form the basis for the studies reported here. Unfortunately, application of the more principled (8) and (9) in CDHMM ASR is computationally intractable due to the need to compute the inverse of C_{pp} for each frame and for each active emitting state. In a typical triphone-based system like the one we report on in the next section, there may be several thousand active states.

Several ways to sidestep this problem exist. One is to use diagonal-only covariance, for which the inverse is formed by simply reciprocating the diagonal elements of C_{pp}. However, the underlying assumption of componentwise independence for filterbank energies is unjustified. Further, transformation via PCA or DCT to achieve exact or approximate independence is not possible for incomplete observations. Another method is to use a single common covariance matrix shared by all states of all models. These so-called *grand covariance* schemes have been used for both robust and normal ASR and offer several advantages. Covariances can be estimated from the whole data set, and thus provide more reliable estimates. The overriding benefit for us is computational: a common covariance requires a single matrix inversion for each *frame* of processed data, rather than an inversion for each HMM state.

3. EXPERIMENTS WITH RM

The HTK HMM Toolkit [13] was adapted for the missing data methods described in section 2. Three-state, single mixture triphone models were trained conventionally on clean speech from the RM corpus parameterised by a filterbank energy acoustic representation, and tested in various missing data conditions using the RM feb89 test set. For details of the results, see [5]. In summary,

1. Unconditional mean imputation generally performs badly.
2. For *random* deletions, both marginal and condition mean techniques hold up encouragingly well: *word accuracy does not fall significantly until around 80% of each observation vector is removed.* This result resembles the performance we have obtained in smaller tasks [4,9], is similar to that reported in other domains [2] and is all the more remarkable because of the simplifications we made to expedite processing: a fixed global covariance and, in the conditional mean case, a fixed global mean.
3. For more realistic deletions based on *local SNR* (i.e. add noise to the clean speech and retain only those channels where local SNR is favourable; see [9]), results for both marginals and conditional reconstruction were much worse: with a global SNR of 20dB, marginal estimation gave an accuracy of 19%, compared to 78% for the equivalent random deletion case.
4. The key difference between random and SNR-based deletions is in the distribution of missing data across time and frequency. To study this effect, we tested the performance of the system in the face of *randomly-deleted blocks* of spectro-temporal energy, a situation which crudely approximates that which is seen in the local SNR case, but allows us to control the deletion block size. We found that removal of contiguous spectral regions is far more harmful than removal of contiguous frames in a narrow spectral region, and that removal of sizeable spectro-temporal blocks leads to quite a sharp decrease in recognition rate for the same overall deletion rate. For instance, with a fixed overall deletion rate of 80%, word accuracy for conditional reconstruction was around 55% when removing blocks of 10 frames by 1 channel but 25% with blocks of 10 frames by 10 channels.

Why do block deletions produce rapid deterioration in recognition performance on RM? The effect might be due to (i) poor estimates of observation possibilities or (ii) the way these estimates are combined across time in the Viterbi algorithm, or to a combination of these factors. We next report on an experiment designed to isolate (i). In this we attempt identification of vowels from single spectral slices using a Gaussian classifier, thus eliminating (ii)'s effects.

4. VOWEL SPECTRA CLASSIFICATION

Gaussian classifiers were trained for the eight most frequent vowels (TIMIT symbols: aa, ae, ah, eh, er, ih, iy, uw) in the TIMIT database [7]. Training and test data was taken from all male speakers in dialect regions dr1 to 7 of TIMIT, a total of 2192 utterances. Two thirds of these utterances, chosen uniformly across speakers, were used for training. The spectrum was obtained from 32 channel mel-scaled filterbank energies. A test set was assembled from spectra for the 3 central frames of vowels in the remaining third of the selected part of TIMIT.

Baseline performance on non-occluded test data was around 60%, and was insensitive to the choice of acoustic vector (filterbank energies or MFCCs), to the addition of overall energy, and to the number of mixtures in the distributions.

A number of missing data conditions and methods were applied to this task, including several variants developed from eqns (7)-(9). Specifically, we addressed two issues:

- is the discrepancy between random deletions and 'energy-based' deletions present at the single-frame level?
- what effect do suboptimal choices of covariances have on recognition performance?

Additionally, several other forms of missing data simulating the effect of low, high and bandpass filtering were tested.

4.1. Random versus energy-based deletion

The top row of figure 1 compares the performance of the Gaussian classifier as a function of missing data technique and spectral deletion type. Four missing data methods were used: unconditional mean estimation (eqn. 7), conditional mean estimation (eqn. 8), estimation using marginals with full covariance structure (eqn. 9), and estimation using marginals with diagonal-only covariance matrices. The latter condition was included because it represents a computationally-tractable approach to the use of marginals (albeit inappropriate for correlated filterbank energies) and for comparison with our previous work. Three forms of spectral deletion were assessed: 'pointwise-random' refers to independent random deletion of spectral components; 'energy' denotes removal of spectral regions with low energy (and roughly corresponds to the SNR-based deletions of section 3); 'random blocks' means the random removal of spectral regions. In the latter case, the spectral blocks deleted had the same region size distribution as for the energy-based deletions. Deletion of blocks at random (for a fixed overall deletion rate e.g. 50% of the spectrum removed) provides a fair comparison with deletions based on energy, since the block-based deletion studies on RM reported in the previous section indicate that random removal of contiguous spectral regions has a greater

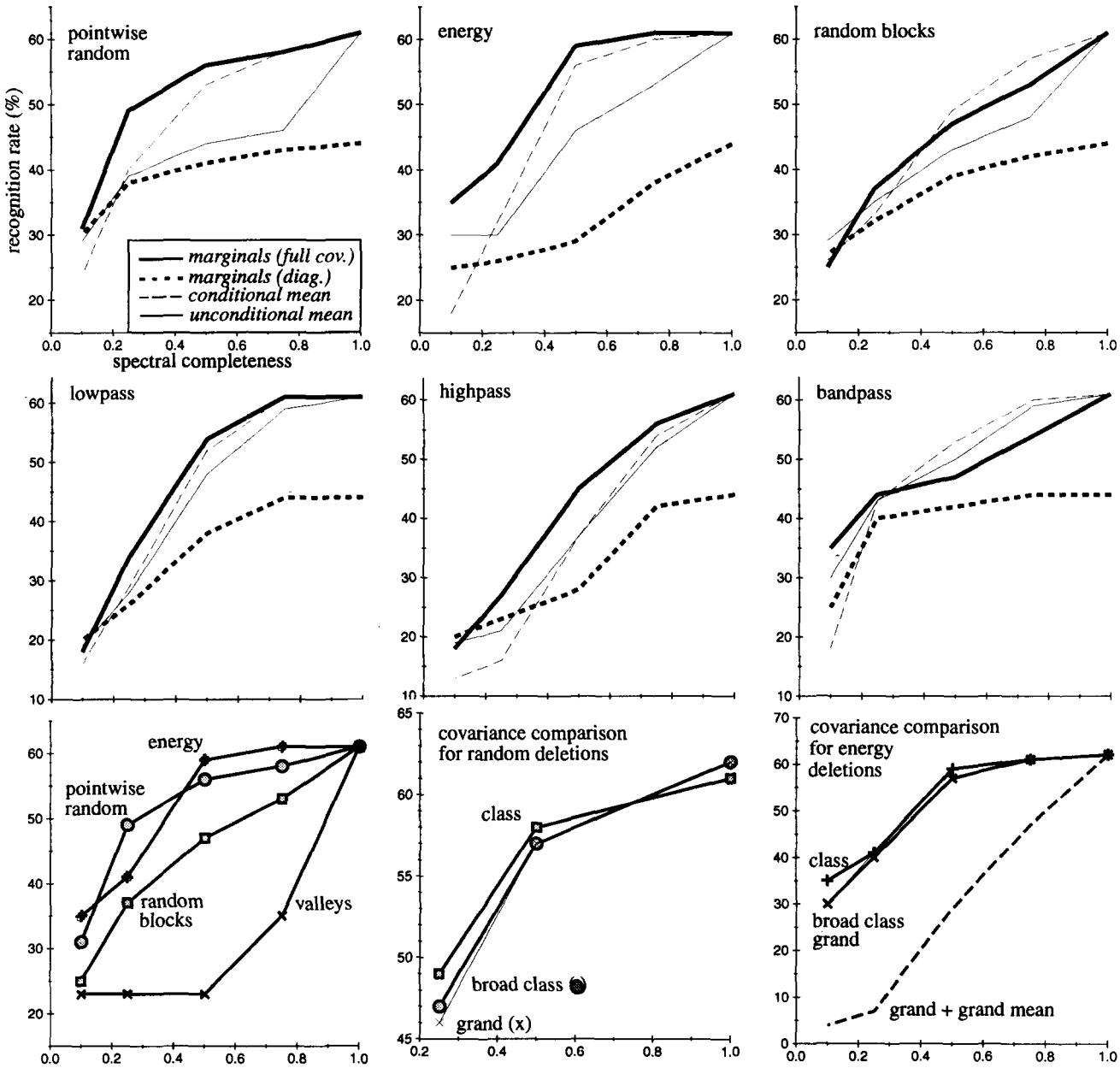

Figure 1: Vowel recognition scores as a function of spectral element deletion (1.0 = no deletion, 0.0 = all deleted). Key in top left panel serves for top and middle rows of figure. ***Top row*** *compares recognition rates for various missing data techniques for deletions: random pointwise (left), low energy regions (middle) and random blocks (right).* ***Middle row*** *shows results for simulated lowpass, highpass and bandpass filtering.* ***Bottom row, left****: comparison of 4 deletion types for full covariance marginal estimation.* ***Bottom row, middle****: comparison of conditional mean estimation for 3 types of covariance matrix on random deletions.* ***Bottom row, right****: compares conditional mean estimation for 3 types of covariance matrix, and for grand means, on energy-based deletion. Note the different scales.*

effect on recognition rate than pointwise deletion.

The three main findings of this comparison are: (i) missing data handling is best achieved with full covariance marginals, with conditional mean imputation working reasonably well except at high deletion rates; (ii) diagonal-only marginals perform poorly (but from a lower baseline), as expected for this acoustic parameterisation; (iii) energy-based deletion works as well as pointwise-random deletion, and significantly better than random block deletion. Point (iii) is made clearer in the bottom row of figure 1 (left panel), which replots one deletion method (full covariance marginals) for pointwise-random, energy and random-block deletion. In addition, results for a further condition ('valleys') in which only low energy regions of the spectrum are retained is shown. As one might expect, for a fixed deletion rate, it matters which parts of the spectrum are removed.

4.2. Which covariance?

The computational intractability of performing many thousands of

matrix inversions per acoustic data frame has been mentioned, and the workaround involving the use of a common covariance matrix for all models was necessary for practical application of these techniques to RM. To assess the effect of this sub-optimal missing data strategy on performance, three different covariance schemes were compared. The first, 'class', represents the optimal strategy which employs a separate covariance matrix for each class. The 'grand' scheme uses a single covariance matrix, estimated using speech across all phones (not just the vowels). An intermediate scheme, 'broad class', uses a single covariance matrix estimated using just the vowels. This intermediate condition tests the possibility of improving performance using more specific covariance matrices whilst still retaining computational tractability.

The main result of this comparison (illustrated in the middle and right panels on the bottom row of figure 1) is that both broad class and grand covariances perform marginally worse than class-specific covariances in conditions of pointwise-random and energy-based spectral deletion. This is a boost for the practical application of missing data techniques for filterbank energy representations of speech, but it does not explain the poor performance of grand covariance schemes on RM. However, the RM studies employed both grand *mean* and covariance to reconstruct a single speech spectrum (as opposed to different reconstructions for each class). A further condition (lower right, fig. 1) utilising both grand mean and covariance resulted in poorer performance, pointing to a possible explanation for the RM results. We have yet to confirm this.

4.3. Simulating spectral filtering

In order to further demonstrate the potential of missing data techniques, a set of deletion conditions simulating various forms of spectral filtering were tested. In the lowpass condition, elements were removed from the higher frequencies. Here, as earlier, the spectral completeness axis refers to the proportion of the spectrum removed, and can be interpreted as the cutoff frequency, on a mel-scale, of an ideal lowpass filter. The highpass and bandpass conditions can be interpreted similarly. In the bandpass case, the band retained was centred on the middle of the mel-frequency range.

In the lowpass condition, recognition performance degrades gradually up to the removal of the top half of the spectrum. A more rapid degeneration is seen for highpass and bandpass, although for the greatest deletion (approximately 90% removal, or 4 points in our 32-element spectral vector), bandpass deletions retain a reasonable level of performance. Further, marginals are not always superior in these conditions. Additionally, whilst conditional mean imputation works well for moderate amounts of spectral deletion, it (predictably) performs poorly when pressed to provide conditional means based on a small number of points present.

5. DISCUSSION POINTS

1. Missing data techniques based on marginal distributions of multivariate Gaussians outperform other techniques in most deletion conditions we have investigated.
2. However, there may be a gain in using a combination of techniques, deploying the technique most suitable to the type and extent of deletion e.g., conditional mean estimation works well across deletion types when the deletion rate is not too severe; for 90% deletion, the unconditional mean is generally superior.
3. As reported in [5], conditional mean estimation can be much improved by making use of the conditional covariance estimate (eqn. 6) as an indication of the confidence. Performance on RM improved from 4% to 37% using this device. These estimates may also lead to an improvement in marginal-based schemes.
4. No significant difference was observed when grand or broad class covariances were used instead of model-specific covariances. This is encouraging because model-specific covariance inversion is infeasible in any realistic ASR task.
5. A performance penalty resulted from the use of grand means.
6. Energy-based deletions (simulating masked data) outperformed random removal of spectral segments with equivalent size distribution. This suggests that estimation from incomplete data is a viable method for obtaining phone likelihoods at the frame level.
7. When estimates from incomplete data are used in the Viterbi algorithm as the basis for recognition decisions across time, random deletions will lead to uncorrelated errors from frame-to-frame. We conjecture that it is easier to recover from these errors than the time-correlated errors induced by realistic deletions, which may explain our RM results.
8. The missing data approach may be useful in models of listeners' performance on the recognition of distorted speech [6,11].
9. An alternative is multiband ASR [3,10], in which outputs from independent recognisers for each frequency band are combined. Multiband recognition is potentially important when it is known that some frequency channels are corrupted by noise. For combination with CASA, it would be necessary to use a relatively large number of bands, and to perform the combination using only a time-varying subset of band-recognisers.

Acknowledgements: Work supported by EPSRC GR/K18962.

REFERENCES

[1] *Proceedings of the 1st Workshop on Computational Auditory Scene Analysis, Int. Joint Conf. Artificial Intelligence*, Montreal, 1995.

[2] Ahmed, S. & Tresp, V. (1993), "Some solutions to the missing feature problem in vision", in: *Advances in Neural Information Processing Systems 5* (eds: S.J. Hanson, J.D. Cowan & C.L. Giles), Morgan Kaufmann, San Mateo, CA.

[3] Bourlard, H. & Dupont, S (1996), "A new ASR approach based on independent processing and recombination of partial frequency bands", *Proc. ICSLP-96*, Philadelphia.

[4] Cooke, M.P., Green, P.D. & Crawford, M.D. (1994), "Handling missing data in speech recognition", *Proc. Int. Conf. Spoken Language Processing*, Yokohama, 1555-1558.

[5] Cooke, M.P., Morris, A.C. & Green, P.D. (1996) "Recognising occluded speech", *ESCA ETRW on The Auditory Basis of Speech Perception*, Keele.

[6] Cooke, M.P. (1996) "Auditory organisation and speech perception: Arguments for an integrated computational theory", *ESCA ETRW on The Auditory Basis of Speech Perception*, Keele.

[7] Garofolo, J.S. & Pallett, D.S. (1989), 'Use of the CD-ROM for speech database storage and exchange', *Proc. Euro. Conf. Speech Communication and Technology*, Paris, 309-315.

[8] Ghahramani, Z. & Jordan, M.I. (1994), "Supervised learning from incomplete data via an EM approach", in: *Advances in Neural Information Processing Systems 6* (eds: J.D. Cowan, G. Tesauro & J. Alspector), Morgan Kaufmann, San Mateo, CA.

[9] Green, P.D., Cooke, M.P. & Crawford, M.D. (1995), "Auditory scene analysis and HMM recognition of speech in noise", *Proc. ICASSP*, 401-404.

[10] Hermansky, H, Tibrewala, S. & Pavel, M. (1996), "Towards ASR on partially corrupted speech", *proc. ICSLP-96*, Philadelphia.

[11] Lippmann, R.P., pers. comm.

[12] Morrison, D.F. (1990), *Multivariate Statistical Methods (3rd ed)*, McGraw Hill.

[13] Young, S.J. & Woodland, P.C. (1993), "HTK Version 1.5", Cambridge University Engineering Department.