SOME SOLUTIONS TO THE MISSING FEATURE PROBLEM IN DATA CLASSIFICATION, WITH APPLICATION TO NOISE ROBUST ASR

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

ABSTRACT

We address the theoretical and practical issues involved in ASR when some of the observation data for the target signal is masked by other signals. Techniques discussed range from simple missing data imputation to Bayesian optimal classification. We have developed the Bayesian approach because this allows prior knowledge to be incorporated naturally into the recognition process, thereby permitting us to go beyond the simple "integrate over missing data" or "marginals" approach reported elsewhere, which we show to be inadequate for dealing with realistic patterns of missing data. After deriving general techniques for recognition with missing data, these techniques are formulated in the context of an HMM based CSR system. This scheme is evaluated under both random and more realistic patterns of missing data, with speech from the DARPA RM corpus and noise from NOISEX. We find that a key problem in real world recognition with missing data is that efficient ASR requires data vector components to be independent, and incomplete data cannot be orthogonalised in the usual way by projection. We show that use of spectral peaks only can provide an effective solution to this problem.

1. INTRODUCTION

Our motivation for studying the "missing data" (MD) problem derives from ongoing studies at Sheffield and elsewhere on source separation by computational auditory scene analysis (CASA). CASA based separation prior to recognition is an attractive paradigm for robust ASR because it makes no assumptions about source type (as does parallel model combination) or number (as does blind separation). However, the potential for MD techniques in ASR does not stand or fall with CASA based separation. In the case of unstructured and near stationary noise there are alternative means for local snr estimation [10], while in some applications, such as band limited or multi-stream ASR [4], the deletion pattern is imposed a priori. In each case this approach requires a solution to the MD problem. This paper is a summary of a larger TR [8].

A number of recent studies have examined the MD problem [1,7,8,11,13]. A method is presented in [1] for MD with stationary noise, and a general maximum likelihood framework for both training and recognition with MD is presented in [7]. However, these methods are not Bayesian, and do not address practical implementation issues.

2. CLASSIFICATION WITH MISSING DATA

In recognition with missing data we are confronted with the problem that whereas with complete data the optimal class decision $d(x)$ given data vector x is[1] $C^\circ = argmax_C P(C|x)$ (see derivation of Eq.14 in section 3), with MD $P(C|x)$ cannot be directly evaluated because x does not have an exact value. Instead both x and $y = P(C|x)$ must be modelled as random variables, with pdfs rather than exact values.

We assume that some prior process has partitioned each data vector x into *present* and *missing* (or *certain* and *uncertain*) parts, (x_p, x_m). The missing part is modelled as a random variable, whose observed values, if available at all, are uncertain and denoted x_u. Let κ denote all knowledge about x, including x_p, x_u, and any other available information, such as a noise model.

2.1 Missing Data Imputation

The simplest approach to recognition with missing data is to estimate x_m from its posterior pdf $f(x|\kappa)$, where the prior pdf for x is estimated from the training set. We can then proceed as with complete data, i.e. take the class decision:

$$d(x) = argmax_C P(C|x_p, \hat{x}_m) \quad (1)$$

We can obtain various estimates for x from $f(x|\kappa)$, such as its mean, median or mode. Of these the mean is the most accurate in that it has the minimum expected quadratic error:

$$E[x|\kappa] = argmin_\theta E[(x-\theta)^2|\kappa] \quad (2)$$

2.2 Class Probability Imputation

We can also try to estimate a value for the random variable $y = P(C|x)$ directly. While the pdf for y will in general be impractical to obtain, its expected value can be obtained in closed form without reference to its pdf because if $y \sim f_y(y)$ and $x \sim f_x(x)$, then:

$$E[y] = \int_y y f_y(y)dy = \int_x y f_x(x)dx \quad (3)$$

This gives us class probability conditional mean value imputation:

$$d(x,\kappa) = argmax_C E[P(C|x,\kappa)|\kappa] \quad (4)$$

where

$$E[P(C|x,\kappa)|\kappa] = \int_x P(C|x,\kappa) f(x|\kappa)dx \quad (5)$$

2.2.1 Certain-Uncertain Factorisation

Factorising $f(x|C) = f(x_p, x_m|C) = f(x_p|C) f(x_m|x_p, C)$

$$E[P(C|x,\kappa)|\kappa] = P(C|x_p,\kappa) E\left[\left.\frac{f(x_m|x_p,\kappa,C)}{f(x_m|x_p,\kappa)}\right|\kappa\right] \quad (6)$$

In the case when the classifier provides models f(x|C) for each class, and $\kappa = x_p$ only (x_u is not used), the denominator in Eq.6 cancels and the pdf integrates to 1, giving:

$$E[P(C|x,x_p)|x_p] = P(C|x_p) \quad (7)$$

This is the widely reported "marginal" technique [1,7,8,11,13].

1. P(x) is "probability of x" & f(x) is the "probability density at x".

2.2.2 *Adapting a Classifier to Operate with Missing Data*

For any classifier whose outputs approximate a posteriori class probabilities, $y_i(x) = P(C_i|x)$, the new output which is required to operate with MD x_m under constraints κ is:

$$y_i'(x) \approx E\left[P(C_i|\, x)\right|\, \kappa] \;=\; \int_{x_m} y_i(x) f(x_m|\, \kappa) dx_m \tag{8}$$

When $f(x_m|\kappa)$ and the above integral can be evaluated in closed form it is straightforward to adapt the classifier to operate with missing and uncertain data. As well as the HMM system described in section 4.1, this is also true of the RBF network described in [1].

3. BAYESIAN OPTIMAL CLASSIFICATION

The Bayesian approach [3] to deriving an optimum class decision C for given data (x, κ) is to first specify a *loss function*, $L(C, d(x, \kappa))$, for every true class C and decided class $d(x)$, and then minimise the overall expected loss or Bayes risk, $r(d(x, \kappa))$, with respect to this loss function and the posterior pdf, $P(C|\, x)$. In simple all or nothing classification, correct classification is usually assigned loss 0 and incorrect classification loss 1 (known as "zero-one loss"). The "Bayes risk" is the expected loss over all C and all x:

$$r(d(x, \kappa)) \;=\; E\left[L(C, d(x, \kappa))|\, \kappa\right] \tag{9}$$

$$= \sum_C \int_x L(C, d(x, \kappa)) f(C, x|\, \kappa)\, dx \tag{10}$$

$$= \int_x \left[\sum_C L(C, d(x, \kappa)) P(C|\, x, \kappa)\right] f(x|\, \kappa) dx \tag{11}$$

When $d(x, \kappa)$ selects correct class C°, then with zero-one loss:

$$\sum_C L(C, d(x, \kappa)) P(C|\, x, \kappa) \;=\; \sum_{C \neq C^\circ} P(C|\, x, \kappa) \tag{12}$$

$$= \;(1 - P(C^\circ|\, x, \kappa)) \tag{13}$$

This shows that with complete data the Bayes risk is simply the probability of misclassification, and this is minimised by the commonly used rule of maximising $P(C|\, x)$, i.e.:

$$d(x) \;=\; argmax_C P(C|\, x) \tag{14}$$

whereas with missing data the Bayes risk is given by:

$$r(d(x, \kappa)) \;=\; \int_x (1 - P(C^\circ|\, x, \kappa))\, f(x|\, \kappa) dx \tag{15}$$

which is minimised by the Bayes decision:

$$d(x, \kappa) \;=\; argmax_C \int_{x_m} P(C|\, x, \kappa) f(x|\, \kappa) dx_m \tag{16}$$

$$= \; argmax_C E\left[P(C|\, x, \kappa)|\, \kappa\right] \tag{17}$$

This shows us that the class probability mean value imputation technique, previously derived in section 2.2 (Eq.4) by another argument, is in fact Bayesian optimal (with respect to zero-one loss). *We now have a theoretical basis for optimal recognition with missing and uncertain data.*

4. HMM BASED CSR WITH MISSING DATA

4.1 Derivation of MD Adapted Recognition Formulae

CSR is presently dominated by HMM based systems and we have therefore focused on MD techniques as required within this framework. The continuous density HMM system [14,15] consists of an HMM model for each of a given set of speech units, with optional simple grammar model (word-word transition probabilities). The HMM model consists of a fixed number of hidden emitting states, each of which is modelled as a Gaussian mixture (GM) pdf (Eqs.27,28) and transition probabilities, $P(s_j|\, s_i)$, between each of these states. Speech units may be words or subword units such as monophones, diphones, or triphones. It does not concern us how the parameters for these models are estimated, because *our interest here is in recognition with MD by systems trained on complete data only.*

In recognition we must select the optimal state sequence $C = (s_1, s_2, \ldots s_T)$ given data matrix $M = (x_1, x_2, \ldots x_T)$. Under the usual simplifying assumptions of data independence between frames and Markovian dependence between states, for complete data the sequence probability is given by:

$$P(C|\, M) \;=\; \prod_i f(x_i|\, s_i)\, P(s_i|\, s_{i-1}) \tag{18}$$

In the case of the knowledge κ which concerns us here (Eq.23),

$$E\left[P(C|\, M, \kappa)\,|\, \kappa\right] \;=\; E\left[P(C|\, M)\,|\, \kappa\right] \tag{19}$$

and the same independence assumptions give the latter as:

$$\prod_i f(x_{pi}|\, s_i)\, P(s_i|\, s_{i-1})\, E\left[\left.\frac{f(x_{mi}|\, x_{pi}, s_i)}{f(x_{mi}|\, x_{pi})}\right|\, \kappa\right] \;=\; \prod_i Q_{ij} \tag{20}$$

Therefore, during the Viterbi recognition procedure with MD, for data frame x_i we must evaluate for each (active) state s_j:

$$Q_{ij} \;=\; f(x_{pi}|\, s_j)\, P(s_j|\, s_{i-1})\, E\left[\left.\frac{f(x_{mi}|\, x_{pi}, s_j)}{f(x_{mi}|\, x_{pi})}\right|\, \kappa\right] \tag{21}$$

where:

$$E\left[\left.\frac{f(x_m|\, x_p, s)}{f(x_m|\, x_p)}\right|\, \kappa\right] \;=\; \int_{x_m} \frac{f(x_m|\, x_p, s)}{f(x_m|\, x_p)} f(x_m|\, \kappa) dx_m \tag{22}$$

Besides the present data, which gives us certain values, further contributions to our knowledge (κ) of x include the following:

- the additivity of energy from different sound sources tells us that the uncertain data provides upper bounds for each of the missing components
- our preprocessing procedure ensures that missing values are bounded below by zero.

This gives us $\kappa = (x_p, x_m \in [0, x_u])$, *so* (23)

$$f(x_m|\, \kappa) \;=\; f(x_m|\, x_p) \Big/ \int_0^{x_u} f(x_m|\, x_p) dx_m \tag{24}$$

and

$$Q_{ij} \;=\; P(s_j|\, s_{i-1})\, f(x_p|\, s) \int_0^{x_u} f(x_m|\, x_p, s) dx_m \Big/ \int_0^{x_u} f(x_m|\, x_p) dx_m \tag{25}$$

If MD bounds are not used, or x_u is zero, the integrals in Eq.25 cancel out. Otherwise the denominator above can be ignored because it is independent of choice of state, so that:

$$Q_{ij} \propto P(s_j|\, s_{i-1})\, f(x_p|\, s) \int_0^{x_u} f(x_m|\, x_p, s) dx_m \tag{26}$$

4.2 Implementation Issues

4.2.1 *Properties of the Gaussian Mixture PDF*

Each HMM state is modelled as a Gaussian mixture pdf:

$$f(x|\, s_j) \;=\; \sum_i a_{ij} N(x, \mu_{ij}, C_{ij}) \tag{27}$$

This pdf is semiparametric (it can fit any pdf when given enough mixture components) and self conjugate (its posterior pdf is of the same family). Each mix component:

$$N(x, \mu, C) = [(2\pi)^n |C| e^{(x-\mu)^t C^{-1}(x-\mu)}]^{-0.5} \quad (28)$$

is specified by its mean vector μ and covariance matrix C. Let the present and missing components of μ and C corresponding to (x_p, x_m) be separated as:

$$\mu = (\mu_p, \mu_m) \qquad C = \begin{bmatrix} C_{pp} & C_{pm} \\ C_{pm} & C_{mm} \end{bmatrix} \quad (29)$$

then the marginal pdf of a GM is given by:

$$f(x_p|\, s) = \sum_i a_i N(x_p, \mu_{ip}, C_{ipp}) \quad (30)$$

and the conditional pdf by:

$$a_{im|\,p} = a_i N(x_p, \mu_{ip}, C_{ipp}) / f(x_p|\, s) \quad (31)$$

$$\mu_{m|\,p} = \mu_m + C_{pm}^t C_{pp}^{-1}(x_p - \mu_p) \quad (32)$$

$$C_{m|\,p} = C_{mm} - C_{pm}^t C_{pp}^{-1} C_{pm} \quad (33)$$

$$f(x_m|\, x_p, s) = \sum_i a_{im|\,p} N(x, \mu_{im|\,p}, C_{im|\,p}) \quad (34)$$

The product of GMs is GM. The quotient of a GM with a Gaussian is GM. The product of the marginal with the integral of the conditional (Eq.26) now takes on a simple form:

$$Q_{ij} \propto \sum_i a_i N(x_p, \mu_{ip}, C_{ipp}) \int_0^{x_u} N(x_m, \mu_{im|\,p}, C_{im|\,p}) dx_m \quad (35)$$

4.2.2 Evaluation of the Multivariate Gaussian Integral

Evaluation of the multivariate Gaussian integral in Eq.35 in closed form requires a change of variable which results in a diagonal covariance matrix. This can be achieved either:

1. exactly, by projecting x onto C's principal components
2. approximately, using a fixed discrete cosine transform
3. crudely, by simply treating all off diagonal values as zero

Option 3 gives unacceptably low performance. Options 1 or 2 can be applied using the following result [12]:

$$x \sim N(x, \mu, C) \Rightarrow y = A^t x \sim N(y, A^t\mu, A^t C A) \quad (36)$$

This still leaves a trapezoidal area of integration which must be approximated by a bounding rectangle. Most of the area of a high dimensional rectangle is close to its vertices [3], so this still leaves scope for inaccuracy. If this were not so the high cost of evaluating the MD conditional pdf in Eq.26 for every state for each data frame could be eliminated entirely, by forcing all values to have non zero uncertainty.

Univariate standard Gaussian pdf and cdf values ($\phi(x)$ and $\Phi(x)$ respectively) are typically so small that we must work with logarithms. When $abs(x) < 5$, $\log(\Phi(x))$ can be obtained using the C-standard erf() and log() functions. Otherwise logΦ(x) must be calculated directly, approximating $\Phi(x)$ by $-\phi(x)/x$ for $x < -5$ and $1 - \phi(x)/x$ for $x > 5$ [6].

5. EXPERIMENTATION WITH RM AND NOISEX

Our speech data is from the DARPA RM 1000 word speaker independent CSR corpus [15]. All 2880 sentences in the trn109 set are used for training; recognition tests uses every 1 in 5 sentences from the 500 sentence feb89 test set. HMM configuration and training follow the "RM Recipe" in [15]. This is a multi-stage procedure which starts with 1-mix monophone models and progresses through to 5-mix state-clustered triphones.

For Fig.1 and results *i-iv* in Fig.2 speech data is parametrised as a 16 channel mel scaled FFT filterbank (fbank16).

5.1 Clean Speech with Random Deletion

Our initial tests used uniform random deletions over frequency and time, and data without noise. Models used at this stage were 1-mix monophones with full covariance. The data pdf was estimated as Gaussian, $N(x, \mu, C)$. Figure 1 shows performance[1] against proportion of data deleted, for MD imputation techniques using increasing amounts of prior information (Eqs.1,2):

1. $\hat{x} = E[x|\, noinformation]$ (zero imputation)
2. $\hat{x} = E[x|\, \mu]$ (mean imputation)
3. $\hat{x} = E[x|\, \mu, C]$ (conditional mean imputation)

and for class probability imputation, with $\kappa = x_p$ (marginals):

4. $\langle P(C|\, x)\rangle = E[P(C|\, x, x_p)\,|\, x_p]$ (Eqs.4,7,30)

The characteristics shown in Fig.1 agree with those reported in [1,7,8,11], where the marginals technique holds out well with up to 60% random deletion.

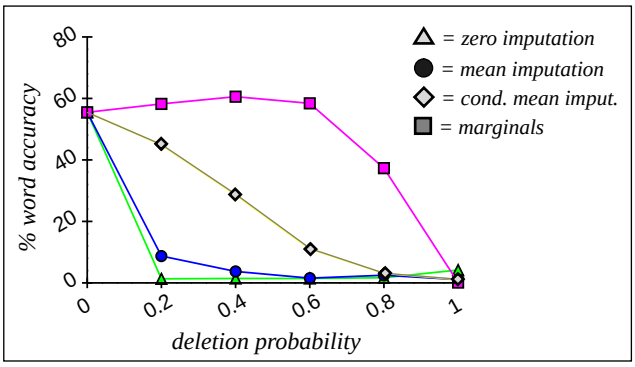


Figure 1: *ASR performance on clean data with random deletions over frequency and time, for three data imputation methods and for marginals Bayesian estimation.*

5.2 S/N Mixture with Local SNR Based Deletion

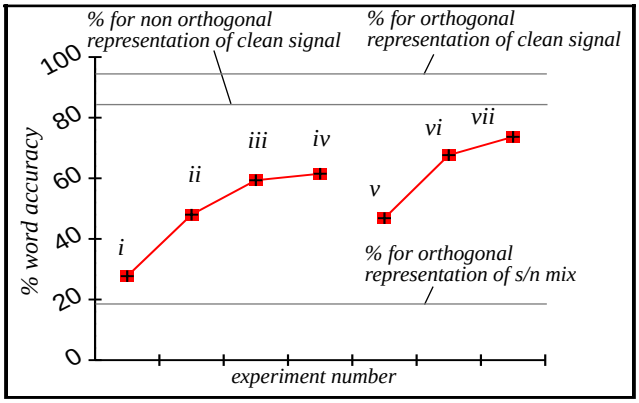
As spectro-temporally neigbouring data is highly correlated, the information overlap in a data sample is less the more uniformly this data is distributed. At a given deletion rate random deletion will therefore preserve considerably more information than the clustered time-frequency deletions which are more likely to occur in reality. For this reason we have also tested MD techniques using local snr based deletion, whereby values in a s/n mixture are deleted when the a priori local snr is below a given threshold. Under these conditions the 61% accuracy for marginals at 40% random deletion in Fig.1 falls to 32% for the same deletion rate.

In Fig.2 we are working with a s/n mixture at 18 dB global snr. Here 1-mix marginals accuracy falls to 28%. In the final stage of the RM Recipe (which reaches 95% absolute word accuracy when trained from a flat start) state models use 5-mix GMs, speech units are (state clustered) triphones, and data vectors have first and second difference components appended. Computational cost completely precludes the use of full covariance at this stage.

With complete data this problem is overcome by approximately orthogonalising the data vector using the DCT (or PCA). This has the effect that every data covariance matrix becomes approximately diagonal, thereby reducing the storage required for each covariance matrix from n^2 to n, and floating point operations involved in covariance matrix arithmetic from $O(n^2)$ to n, or, in the case of matrix inversion, from n^3 to n.

1. The "%word accuracy" usually quoted is 100.(H-I)/(H+S+D). Here we have used 100.H/(H+S+D+I), because this is a true percentage [10].

As combining a known with an unknown value results in an unknown value, neither of the linear preprocessing operations of data orthogonalisation or time differencing, frequently carried out during conventional ASR, are possible with MD. Result *iii* in Fig.2 shows that the disadvantage of going from full to diagonal covariance without orthogonalisation is more than offset by the advantage of using triphones and difference coefficients.



n	params	model	method	marg	bounds
i	fbank16	1-mix	marginals	snr	not snr
ii	fbank16	1-mix	marg+b	snr	not snr
iii	fbank16	5-mix	marg+b	snr	not snr
iv	fbank16	5-mix	marg+b	pk & snr	not snr
v	rate64	5-mix	marg+b	snr	not snr
vi	rate64	5-mix	marg+b	pk & snr	not snr
vii	rate64	5-mix	marg+b	pk & snr	pk & not snr

Figure 2: *shows performance of various strategies for the recognition of RM data mixed with helicopter noise from NOISEX, at 18 dB global snr. The table specifies the details of each experiment: parameterisation, model type, missing data method, and rules for inclusion of each value into the marginals or bounds factors (Eqs.6,26,35). Abbreviations: "1-mix" = "1-mix monophones, with full covariance", "5-mix" = "5-mix state-clustered triphones, with diagonal covariance and appended 1st and 2nd difference coefficients", "b" = "bounds", "snr" = "local snr > 18 dB", "pk" = "is spectral peak".*

5.3 Peaks Selection for Data Orthogonalisation

The only way a filterbank vector can be rendered orthogonal without projection is to select a subsample of points none of which are close neighbours. Spectral peaks are good candidates for this purpose [2] because they are usually well spaced, they are relatively unlikely to be dominated by noise, and formant centres are highly correlated with phoneme identity. Result *iv* shows that the additional peaks selection criterion does not give much advantage with the fbank16 data representation, in which peaks are very sparse. Results *vi-vii* in Fig.2 show that the advantage of peaks filtering is much more pronounced with the data representation from a 64 channel auditory nerve firing rate model [5] (rate64), while result *v* confirms that the degree of redundancy in high resolution representations can lead to a fall in performance when this redundancy is not exploited.

6. DISCUSSION

We have developed a Bayesian framework for classification with missing data.We have demonstrated that this can be applied to HMM based ASR with spectral data in which noise corrupted values have been tagged. In so doing it was noted that normal orthogonalisation techniques are not applicable with missing data and this problem was partly overcome using a high resolution spectral data representation from which only peaks were retained. Performance could possibly be improved by expanding the data vector to span a number of time frames, thereby exploiting temporal data correlation.

Acknowledgements: This work was supported by the EPSRC Communications Signal Processing & Coding Initiative (Research Grant GR/K18962).

REFERENCES

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

[2] Barker, J. & Cooke, M.P. (1997) "Modelling the recognition of spectrally reduced speech", *Proc. Eurospeech'97*, pp.2127-2130.

[3] Bishop, C. (1995) *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.

[4] Bourland, H., Dupont, S. & Ris, C. (1996), "Multi-stream speech recognition", *IDIAP-RR -96-07*.

[5] Cooke, M.P. (1993) *Modelling auditory processing and organisation*, Cambridge University Press.

[6] Dwight, H.B. (1967) *Tables of Integrals and Other Mathematical Data*, New York, MacMillan.

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

[8] Green, P.D., Cooke, M.P. & Crawford, M.D. (1995), "Auditory scene analysis and HMM recognition of speech in noise", *Proc. ICASSP'95*, pp.401-404.

[9] Morris, A.C., Green, P.D. & Cooke, M.P. (1998), "Bayesian techniques for recognition with missing data: application to the recognition of occluded speech by Hidden Markov Models", *Univ. of Sheffield Dept. of Computer Science Technical Report TR-98-02*.

[10] Hirsch, H.G. & Ehrlicher, C. (1995), "Noise estimation techniques for robust speech recognition", *Proc. ICASSP'95*, pp.153-156.

[11] Lippmann, R.P. & Carlson, B.A. (1997), "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise", *Proc. Eurospeech'97*, pp.37-40.

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

[13] NIPS'95 workshop, Denver (1996) "Missing data: Methods and Models", MIT Press.

[14] Rabiner, L.R. (1989), "A tutorial on HMMs and selected applications in speech recognition", *Proc. IEEE, 12(2)*, pp.267-296.

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