IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

# DATA UTILITY MODELLING FOR MISMATCH REDUCTION

Andrew C. Morris

IDIAP-RR 01-30

October 2001

ACCEPTED FOR PUBLICATION IN
Proc. CRAC (Consistent & Reliable Acoustic Cues for sound analysis) workshop
Satellite event at Eurospeech 2001, Aalborg, Denmark

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 - 27 - 721 77 11
fax +41 - 27 - 721 77 12
email secretariat@idiap.ch
internet http://www.idiap.ch

# DATA UTILITY MODELLING FOR MISMATCH REDUCTION

Andrew Morris, Jon Barker, Hervé Bourlard

October 2001

**Abstract**

In the “missing data” (MD) approach to noise robust automatic speech recognition (ASR), speech models are trained on clean data, and during recognition sections of spectral data dominated by noise are detected and treated as “missing”. However, this all-or-nothing hard decision about which data is missing does not accurately reflect the probabilistic nature of missing data detection. Recent work has shown greatly improved performance by the “soft missing data” (SMD) approach, in which the “missing” status of each data value is represented by a continuous probability rather than a 0/1 value. This probability is then used to weight between the different likelihood contributions which the MD model normally assigns to each spectral observation according to its “missing” status. This article presents an analysis which shows that the SMD approach effectively implements a Maximum A-Posteriori (MAP) decoding strategy with missing or uncertain data, subject to the interpretation that the missing/not-missing probabilities are weights for a mixture pdf which models the pdf for each hidden clean data input, after conditioning by the noisy data input, a local noise estimate, and any information which may be available. An important feature of this “soft data” model is that control over this “evidence pdf” provides a principled framework not only for ignoring unreliable data, but also for focusing attention on more discriminative features, and for data enhancement.

**Keywords**: Bayesian recognition, missing data, data utility, HMMs, robust ASR

**Acknowledgements:** The Soft MD approach, which the SMD model presented here was designed to fit, was introduced in [3]. This work was supported by the EC/OFES (European Community / Swiss Federal Office for Education and Science) RESPITE project (REcognition of Speech by Partial Information TEchniques).

# Contents

# 1. Introduction

The aim of this paper is to explain how the theory of recognition with missing data was recently generalised to cover the “soft missing data” model [3,12], and to show that while the strategy of the previous missing-data approach was simply to detect and ignore missing data, in the new soft-data approach the central task becomes the *estimation of a clean data pdf to replace each noisy data value*. The degree to which this pdf affects the recognition process is controlled by its peakedness. This means that there is now scope for the detection not only of misinformative data (as before) but also of uninformative data, which is also known to reduce model discrimination (or conversely of data *utility)*. Furthermore, as the mode of this pdf represents the most likely clean data value, pdf estimation also implicitly includes data enhancement.

In Section 2 we introduce the core theory behind the “missing data” (MD) approach to noise robust ASR, and describe briefly how data to be treated as missing is detected. Section 3 then explains how the “soft missing data” (SMD) generalisation of missing-data ASR was introduced. Section 4 discusses results obtained so far with the SMD model, and the implications of this model which remain to be tested.

# 2. Missing feature theory (MFT) in ASR

The “missing data” approach in ASR was initially motivated by studies within the telecommunications industry on human speech perception [2,6] which showed that we are able to recognise most different speech sounds when a very large proportion of the auditory nerve image is masked by noise. The aim of MD ASR is to exploit this redundancy in the auditory image as a means of reducing the effect of data mismatch, in which a very large drop in recognition performance often occurs when even a small part of the input data does not match the data used in model training. In MD ASR [3,5,9,11,12,13] models are trained on clean speech only, and during recognition sections of (compressed) spectral data dominated by noise are detected and treated as “missing”.

## 2.1 Missing-data detection

Missing feature detection is most commonly based on spectro-temporarily localised signal-to-noise ratio (SNR) estimation.

The method used for local SNR estimation for the results reported here was based on a simple procedure originally used for speech enhancement by spectral subtraction. While the noise spectrum is often estimated in periods of non-speech, here we simply used the average of the first 100 ms of spectral data in each utterance. See [5,8,10] for more accurate (and more complex) methods for noise estimation.

On the basis of the “maximum assumption” (see Section 3.2), 0/1 MD mask values are set to 1 if the estimated SNR is greater than zero, or else to 0, while SMD uses a sigmoid function to squash the local SNR estimate to obtain P(not missing). See [3] for more details of “soft” mask estimation.

## 2.2 Bayesian Optimal Classification with Missing Data

It is common knowledge that when none of the observation data $X$ is uncertain, the class decision which maximises the probability of correct classification is the MAP decision $Q_{MAP} = argmax_Q P(Q|X)$, and for a trained classifier

$$Q_{MAP} = argmax_Q P(Q|X, \Theta) \tag{1}$$

If clean data $X$ is *partly missing or uncertain*, and has pdf $s(X)$, then the optimum decision function (by the same optimality criterion) is given [13] by

$$Q_{MAP} = argmax_Q E[P(Q|X, \Theta)|X \sim s(X)] \tag{2}$$

## 2.3 Normal Viterbi decoding with HMMs

We summarise here the equations used for decoding with normal HMMs. The Viterbi dynamic programming procedure provides a rapid approximation of the MAP objective in Eq.1. As HMMs model $P(X|Q)$ rather than $P(Q|X)$, Bayes' rule is used, together with the fact that $P(X|\Theta)$ is the same for any choice of Q, to give

$$Q_{MAP} = \arg max_Q P(Q|\Theta)P(X|Q, \Theta) \tag{3}$$

$P(Q|\Theta)$ is not further considered here because it is not affected by missing data - though it is worth noting that *pronunciation and duration modelling become progressively more important as noise level increases*.

For a particular state sequence $Q_a$, having state $q_{a(t)}$ at time $t$, the usual Markovian independence assumption gives

$$p(X|Q_a, \Theta) \cong \prod_t p(x_t|q_{a(t)}, \Theta) \tag{4}$$

where $p(x_t|q_k, \Theta)$ is usually modelled by a mixture pdf

$$p(x|q_k, \Theta) = \sum_j P(m_j|q_k, \Theta)p(x|m_j, q_k, \Theta) \tag{5}$$

in which each pdf component $p(x|m_j, q_k, \Theta)$ is a multivariate diagonal covariance Gaussian, for which

$$p(x|m_j, q_k, \Theta) = \prod_i p(x_i|m_j, q_k, \Theta) \tag{6}$$

## 2.4 Viterbi decoding with hard missing data

For 0/1 mask values, the only effect of replacing the usual MAP objective in Eq.1 with the MD MAP objective in Eq.2 is to replace "missing" factors $p(x_i|m_j, q_k, \Theta)$ in Eq.6 by

$$\int_0^{x_{i,obs}} p(x_i|m_j, q_k, \Theta)dx_i \tag{7}$$

Apart from a constant factor which is independent of the choice of $Q$ (though which helps numerical stability when the integral in Eq.7 tends to zero with $x_{i,obs}$), Eq.7 results as a special case (for $\varphi_i$ always = 0 or 1) from the SMD generalisation which is presented later in Eq.17.

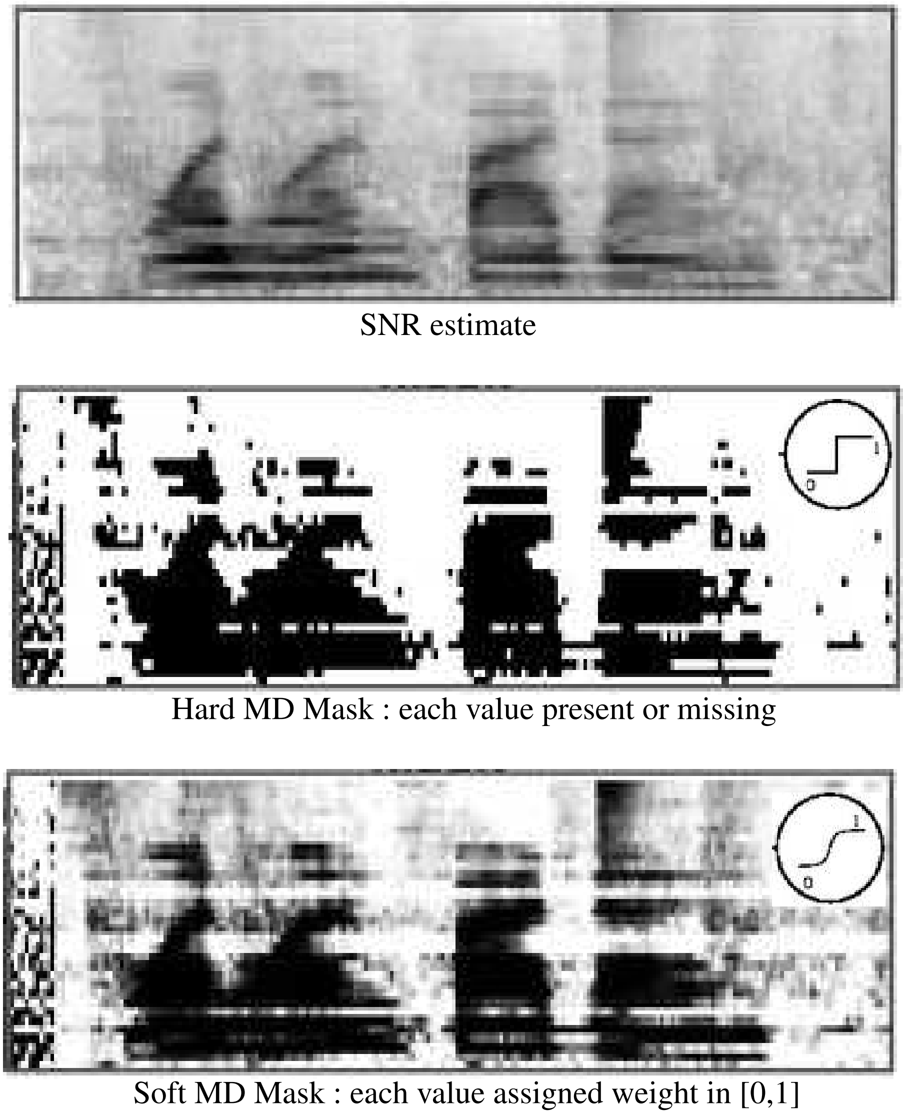


*Figure 1. After SNR estimation, "hard MD" ASR applies threshold at SNR=0 to obtain hard "missing" decision. "Soft MD" model squashes SNR estimate to give soft P(not missing) weight.*

## 3. Generalisation of MFT to soft missing data

While MD ASR has shown promising results in the past [9,13], the process of missing data detection is inherently approximate, and recent work [3] has shown that the "soft missing data" approach, in which the "missing" status of each data value is represented by a continuous probability rather than a simple 0/1 value, can lead to greatly improved performance for insignificant extra cost (see Fig.3). In this section we show how continuous MD probabilities are used in the SMD model to weight between the separate contributions to the data likelihood which were previously used for "present" and "missing" data. We then discuss how the resulting model permits modelling of a more general kind of data utility.

### 3.1 Viterbi decoding with soft missing data

For MAP decoding with soft data we must evaluate the expected posterior probability in Eq.2. The only approach we have found in the literature for this purpose [1,4] cannot be applied here, because (1) MD detection here is probabilistic, and (2) MD ASR does not assume any fixed noise model. Instead we can make direct use of the expectation integral [13],

$$E[P(Q|X,\Theta)|X \sim s(X)] = \int P(Q|X,\Theta)s(X)dX \tag{8}$$

$$= P(Q|\Theta)\int \frac{p(X|Q,\Theta)}{p(X|\Theta)} s(X)dX \tag{9}$$

We now consider the form of the clean data pdf $s(X)$. Let knowledge on which the clean data pdf depends be divided into three parts: the clean training data set (modelled by the clean data prior $p(X|X_{tr}) = p(X|\Theta)$), the observed noisy utterance data $X_{obs}$, and any other knowledge $(\kappa)$ (such as estimates of the local noise level at each point in $X$, bounds constraining each observed value, estimated observation precision, data utility, and so on). Using Bayes' rule, and the independence assumption $P(X_{tr}, X_{obs}) = P(X_{tr})P(X_{obs})$,

$$p(X|X_{tr}, X_{obs}, \kappa) = p(X|\Theta)p(X|X_{obs}, \kappa)/p(X) \tag{10}$$

Providing the pre-evidence prior $p(X)$ is very flat, it can be taken into the normalising constant $(c)$, so that

$$s(X) = cp(X|\Theta)p(X|X_{obs}, \kappa) = cp(X|\Theta)s'(X) \tag{11}$$

where we will call $s'(X) = p(X|X_{obs}, \kappa)$ the evidence pdf. Combining Eqs. 3, 9 & 11 now gives the key SMD equation,

$$Q_{MAP} = \arg max_Q P(Q|\Theta)\int p(X|Q,\Theta)s'(X)dX \tag{12}$$

An important feature of (hard- and) soft MD is its ease of implementation, particularly within commonly used diagonal covariance Gaussian mixture model based HMMs. With the assumption that $s'(X) \cong \prod_t s'(x_t)$,

$$\int p(X|Q_a,\Theta)s'(X)dX$$

$$= \int \prod_t \ldots = \prod_t \int p(x_t|q_{a(t)},\Theta)s'(x_t)dx_t \tag{13}$$

Therefore, as the integral of the mixture component sum is the sum of the integrals, the only difference between HMM decoding with deterministic and probabilistic data is that the mixture component contribution $p(x|m_j, q_k, \Theta)$ in Eq.5 is replaced by $\int p(x|m_j, q_k, \Theta)s'(x)dx$.

We will next present the simple present/missing components mixture pdf that was (implicitly) used for modelling $s'(x)$ in [3].

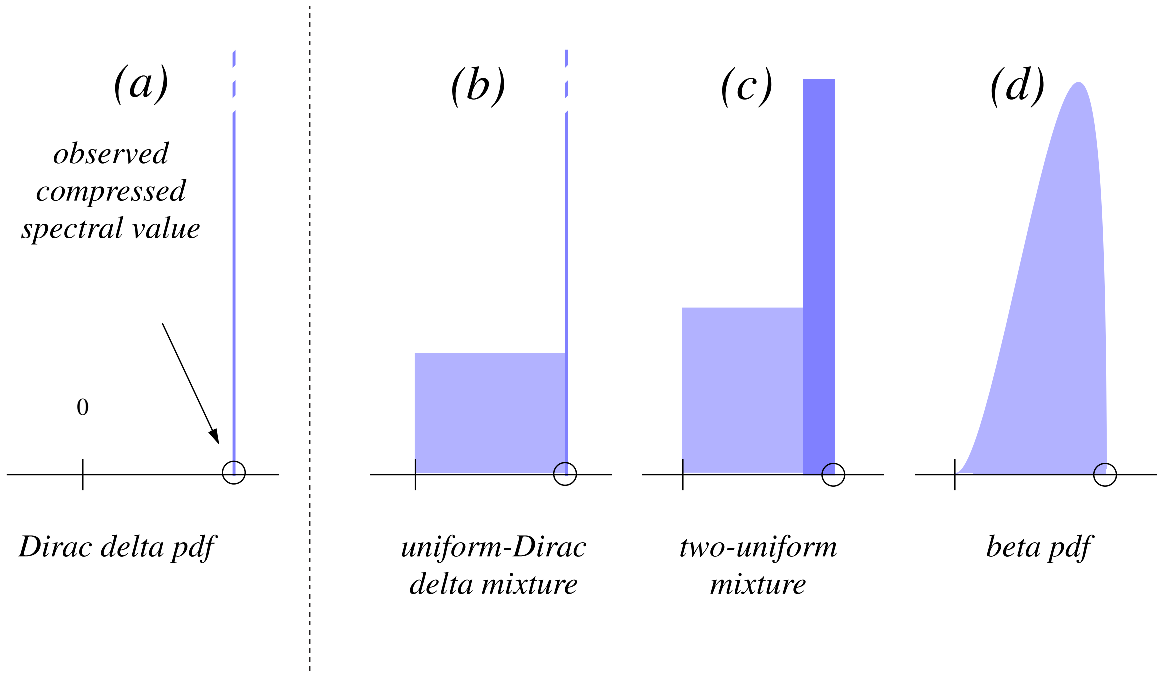


*Figure 2.* **Evidence PDF Model.** *Baseline HMM uses Dirac delta evidence pdf at obs. value (a). Hard MD uses* ***either*** *uniform* ***or*** *Dirac pdf. SMD here uses a uniform-Dirac mixture pdf (b). Alternative soft data pdfs include a two-uniform mixture (c), and a beta pdf (d).*

## 3.2 Evidence pdf model for soft missing data

Missing-data mask probabilities here are based on local SNR estimates, which make use of the "maximum assumption" that for compressed data (here this is cube-root compression),

$$P(clean) = P(SNR > 0) = \varphi \tag{14}$$

In the "soft missing data" model for which results are reported in Fig.3, the SMD mask $P(clean)$ values are used to weight between the likelihood contributions used in the hard MD approach for "present" and "missing" components. In other words, for clean data the true data value is simply equal to the observed value, while missing data is subject only to the bounds constraint $x \in [0, x_{obs}]$. In this case the evidence pdf is modelled by a mixture pdf, for each component $x_i$, as

$$s'(x_i) = P(cln)p(x_i|cln) + P(\neg cln)p(x_i|\neg cln)$$

$$= \varphi_i\delta(x_i - x_{i,obs}) + (1-\varphi_i)u(0, x_{i,obs}) \tag{15}$$

where $\delta(x_i - x_{i,obs})$ is the Dirac delta function about $x_{obs}$, and $u(0, x_{i,obs})$ is the uniform distribution over $[0, x_{i,obs}]$. Assuming that $s'(x_t) \cong \prod_i s'(x_{ti})$ and substituting into the integral in Eq.13

$$\int p(x|m_j, q_k, \Theta)s'(x)dx = \prod_i \int p(x_i|m_j, q_k, \Theta)s'(x_i)dx_i \tag{16}$$

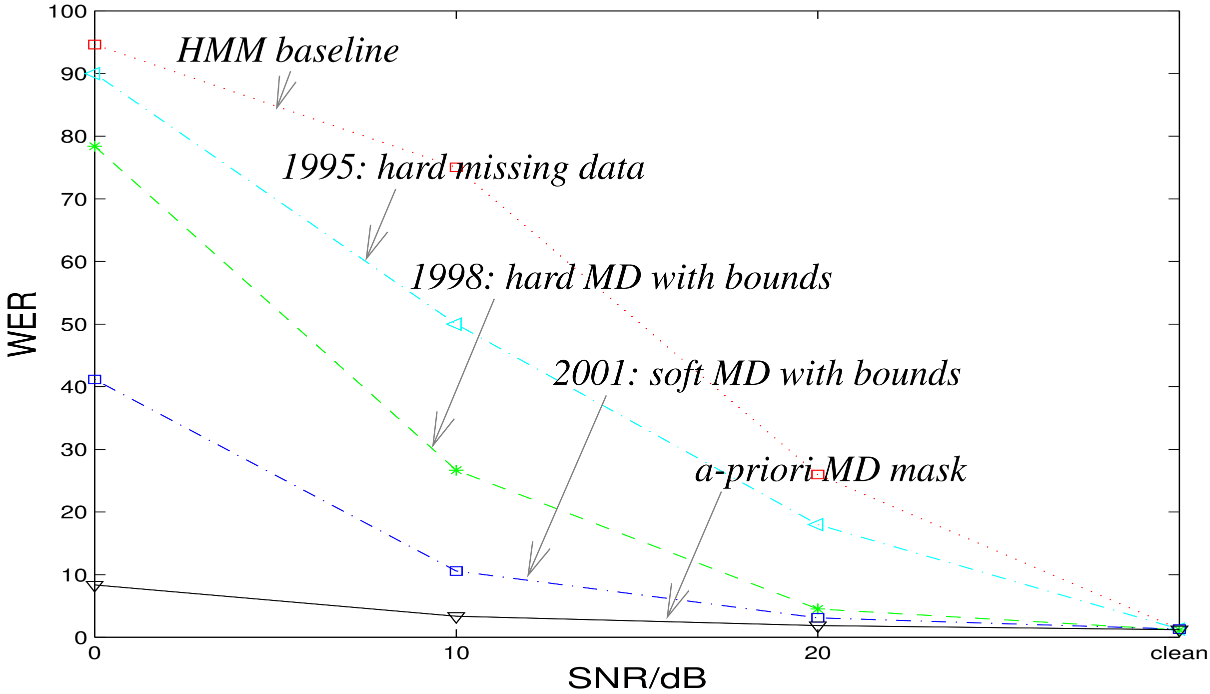


*Figure 3. Performance of different MD models against baseline Gaussian mixture HMM performance. Task is Aurora 2.0 connected digits. SNR estimation uses noise spectrum = first 100 ms of signal. Results averaged over 4 noise types. (No tests yet with utility estimation).*

$$\int p(x_i|m_j, q_k, \Theta)s'(x_i)dx_i = \varphi_i p(x_{i,obs}|m_j, q_k, \Theta) + \frac{(1-\varphi_i)}{x_{i,obs}}\int_0^{x_{i,obs}} p(x_i|m_j, q_k, \Theta)dx_i \tag{17}$$

Here it could be argued that the "max" assumption in Eq.14 implies an alternative evidence pdf as follows:

$$s'(x_i) = p(x_i|x_{i,obs}, \kappa) = P(snr > 0)p(x_i|snr > 0) + P(snr \le 0)p(x_i|snr \le 0) \tag{18}$$

This is a two-uniform pdf mixture (see Fig.2(c)),

$$s'(x_i) = \varphi_i u(x_{i,snr0}, x_{i,obs}) + (1-\varphi_i)u(0, x_{i,snr0}) \tag{19}$$

where for compression $\gamma$, $x_{i,snr0} = \gamma(\gamma^{-1}(x_{i,obs})/2)$. Though plausible, this model would tend to soften the evidence provided by all clean data values, and has not yet been tested.

## 4. Discussion

**Test results.** The initial "soft data" experiments in Fig.3 [3] compare the performance of different missing data models against baseline Gaussian mixture HMM performance. The test used is the Aurora 2.0 task for speaker independent continuous digits recognition [15]. Results shown here for the “soft data” model do not yet improve over results for Aurora multi-condition training (not shown), but these initial tests used only a very simplistic form of SNR estimation for MD mask estimation.

**Advantages of MD approach in general.** Although multi-condition training has produced good results, the great potential for mismatch reduction by noise removal can only ever benefit systems which train on clean speech alone. Yet speech restoration is not always possible, and it is better to have the ability to detect and ignore uncorrectable data than to replace it with false data (each soft data pdf should be minimally informative). Strong benefits are often gained with MD ASR even when mask estimation is very approximate. The benefit of excluding mismatched data easily outweighs the cost of also losing a smaller amount of useful data.

**Advantages of the SMD model.** The clean data pdf estimation at the heart of the SMD model provides a natural framework within which to incorporate techniques not only for noise detection, but also for noise removal (where possible) and data utility estimation (for increased discrimination).

**Attention modelling.** The auditory system contains many mechanisms which could have a role in the focusing of (conscious or pre-conscious) attention, which is known to play a crucial role in acoustic perception. For example, neurons in the first stages of central auditory processing detect various acoustic features, such as phoneme transitions [7], while frequency selectivity in the basilar membrane is actively controlled by feedback connections to the outer hair cells. Acoustic event detection models [14,16] could possibly be used to downweight uninformative data, and language models to focus the range of expected phonetic characteristics.

**Noise removal.** Many techniques are available for noise removal. SNR estimation should not be confined to missing-data mask estimation.

**Need for improved language modelling.** A known weakness of HMM modelling is the inability of state transition probabilities to sufficiently model temporal invariants. The relative importance of the duration/ pronunciation/ language model increases with noise level, so the $P(Q|\Theta)$ term in Eq.12 holds a major potential for improving robustness.

**Need for improved data reliability estimation.** MD ASR shows a strong advantage even with very approximate MD mask estimation, but performance could easily be increased considerably in future by using more advanced methods for missing data detection, such as those reviewed in [5].

**Limitation to diagonal covariance models.** While the MD ASR techniques described here offer a practical solution to noise robust ASR, they can be applied only with diagonal covariance models. However, diagonal covariance mixture models do permit some degree of covariance modelling.

## 5. Summary and conclusion

The “missing data” approach to robust speech recognition was recently improved by replacing discrete 0/1 P(missing) values by a continuous value in $[0, 1]$. In this paper it was shown how a generalisation of “missing data” theory which was introduced to account for this “soft missing data” approach has resulted in a model where each deterministic input data value is replaced by a hidden variable whose probabilistic value is represented by a pdf. This “data utility” pdf is conditioned by all knowledge concerning the observation value, so this model provides a natural framework for mismatch robust recognition, in which not only reliability estimation, but also noise removal and data salience estimation is incorporated into the clean data pdf estimation procedure. For optimal performance, this analysis should be applied to training as well as to recognition.

## Acknowledgements

The Soft MD approach, which the SMD model presented here was designed to fit, was introduced in [3]. This work was supported by the EC/OFES (European Community / Swiss Federal Office for Education and Science) RESPITE project (REcognition of Speech by Partial Information TEchniques).

## References

[1] Ahmed, S. & Tresp, V. (1993) “Some solutions to the missing feature problem in vision”, in Advances in Neural Information Processing Systems 5, Morgan Kauffman, San Mateo, pp. 393-400.

[2] Allen, J. B. (1994) “How do humans process and recognise speech?”, IEEE Trans. on Speech and Signal Processing, Vol.2, No.4, pp.567-576.

[3] Barker, J., Josifovski, L., Cooke, M.P. & Green, P.D. (2000) “Soft decisions in missing data techniques for robust automatic speech recognition”, Proc. ICSLP-2000, pp.373-376.

[4] Duda, R. O., Hart, P. E. & Stork, D. G. (2000) Pattern classification, 2nd Ed., John Wiley & Sons, Inc.

[5] El-Maliki, M. (2000) “Speaker verification with missing features in noisy environments”, PhD thesis, Dept. d’Electricité, Ecole Polytechnique Fédérate de Lausanne.

[6] Fletcher, H. (1922) “The nature of speech and its interpretation”, J. Franklin Inst., 193(6), pp.729-747.

[7] Furui, S. (1986) “On the role of spectral transition for speech perception”, J. Acoust. Soc. Am., 80(4), pp.1016-1025.

[8] Gaillard, F., Berthommier, F., Feng, G., & Schwartz, J.-L. (1999) “A reliability criterion for time-frequency labelling based on periodicity in an auditory scene”, Proc. Eurospeech’99, pp.2603-2606.

[9] 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.

[10] Hirsch, H. G. and C. Ehrlicher (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] Morris, A. C., Barker, J. & Bourlard, H. (2001) From missing data to maybe useful data: Soft data modelling for noise robust ASR”, Proc. IoA Workshop on Innovative methods in Speech Processing (WISP 2001), pp.153-164.

[13] Morris, A.C., Cooke, M. & Green, P. (1998) “Some solutions to the missing feature problem in data classification, with application to noise robust ASR”, Proc. ICASSP'98, pp.737-740.

[14] Morris, A.C., Pardo, J.M. (1995) “Phoneme transition detection and broad classification using a simple model based on the function of onset detector cells found in the cochlear nucleus”, Proc. Eurospeech’95, pp.115-118.

[15] Pearce, D. & Hirsch, H.-G. (2000) “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions”, Proc. ICSLP’00, Vol.4, pp.29-32.

[16] Salomon, A. & Espy-Wilson, C. (2000) “Detection of speech landmarks using temporal cues”, Proc. ICSLP 2000, vol.3., pp.762-765.