FROM MISSING DATA TO MAYBE USEFUL DATA: SOFT DATA MODELLING FOR NOISE ROBUST ASR

A. C. Morris Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), CH
J. Barker Dept. of Computer Science, Sheffield University, UK
H. Bourlard Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), CH
Swiss Federal Institute of Technology (EPFL), Lausanne, CH

ABSTRACT

Much research has been focused on the problem of achieving automatic speech recognition (ASR) which approaches human recognition performance in its level of robustness to noise and channel distortion. We present here a new approach to data modelling which has the potential to combine complementary existing state-of-the-art techniques for speech enhancement and noise adaptation into a single process.

In the “missing feature theory” (MFT) based approach to noise robust ASR, misinformative spectral data is detected and then ignored. Recent work has shown that MFT ASR greatly improves when the usual hard decision to exclude data features is softened by a continuous weighting between the likelihood contributions normally used for “good” and “bad” data. The new model presented here can be seen as arising from a generalisation of this “soft missing data” approach, in which the implicit good-bad mixture pdf is modelled explicitly as the data posterior pdf.

Initial “soft data” experiments compare the performance of different soft missing data models against baseline Gaussian mixture HMM performance. The test used is the Aurora 2.0 task for speaker independent continuous digits recognition.

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

1. INTRODUCTION

1.1 All Observation Data is Ultimately Probabilistic

According to the Bayesian theory of statistics, whenever the precision of an observation is finite, the information supplied by the observation is best described not by the value of the observation alone, but by its posterior pdf. The information supplied by an observation depends not only on the value of the observation, but also on its prior pdf. This leads us directly to the central idea presented here, which is to replace deterministic sensory data by probabilistic or “soft”’ data which more accurately represents the posterior data pdf. This model has the potential to more accurately represent our beliefs about the different possible hidden values which the true data may have.

One could make metaphysical analogies with the situation in quantum mechanics, where observable reality is most successfully modelled as being inherently probabilistic. But, metaphysics aside, the practical utility of internal or “hidden” probabilistic variables is already well established in Hidden Markov Models and “Bayesian networks”. The difference in the model proposed here is that the probabilistic values in question are not hidden internal variables. They are the raw input to any practical information processing system which needs to acknowledge limited observation precision.

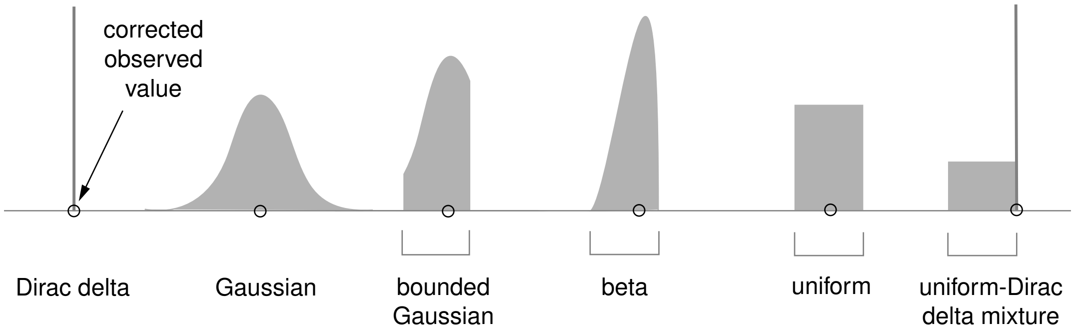

Fig.1 *A precise observation has a dirac pdf, but exact measurements do not exist. Beliefs about true values do exist, and can be expressed quantitatively by a pdf. Many common pdf functions, including the Gaussian, beta and uniform pdfs illustrated here, are defined by just two parameters. A common constraint is an interval in which the observation is known to lie. The pdf on the right is the uniform-delta weighted mixture pdf (4 parameters), which has so far given best results in ASR.*

1.2 The Data Mismatch Problem and Negative Data Utility

A parametric classifier assumes that the data during recognition comes from the same population as the data used in training. Because there are so many sources of variability in speech data, and even the largest of speech databases is finite, this assumption is often very inaccurate. Even low levels of mismatch between the data being recognised and the data used for training can lead to a large drop in recognition performance. Where distortion is predictable, one of the most effective strategies is to train a classifier using data representative of this range of distortion (see Fig.2). Mismatching data need not therefore be “corrupt” or contain less relevant information. It can even be “too clean”.

Mismatching data is simply less useful. In fact, if utility is quantified by the degree to which a particular data value increases classification performance, then data mismatch can rapidly lead to the utility of a large proportion of data becoming negative., i.e. performance increases when this data is ignored. When distortion is predictable, other methods beside training in noise, such as “parallel model combination”, can also be very effective. But for the majority of applications, the full range of distortions likely to be encountered is far too large for noise modelling of any kind to be successful.

If we can’t model noise then we need to model speech. Anything which is not speech-like can then be removed by suitable preprocessing, or detected and ignored or downweighted during recognition. Noise removal during preprocessing would fit into the present “soft data” approach as an integral part of observation pdf estimation, with existing noise robust spectral preprocessing techniques [13] supplying noise corrected spectral features as the pdf mode. It may be possible to persuade the same analysis to contribute towards estimating the shape of the pdf.

Training in noise is incompatible with the “soft data” approach. While it is often possible to estimate clean speech from noisy speech without knowing the noise type, it is not possible to estimate noisy speech from clean speech unless the noise is known.

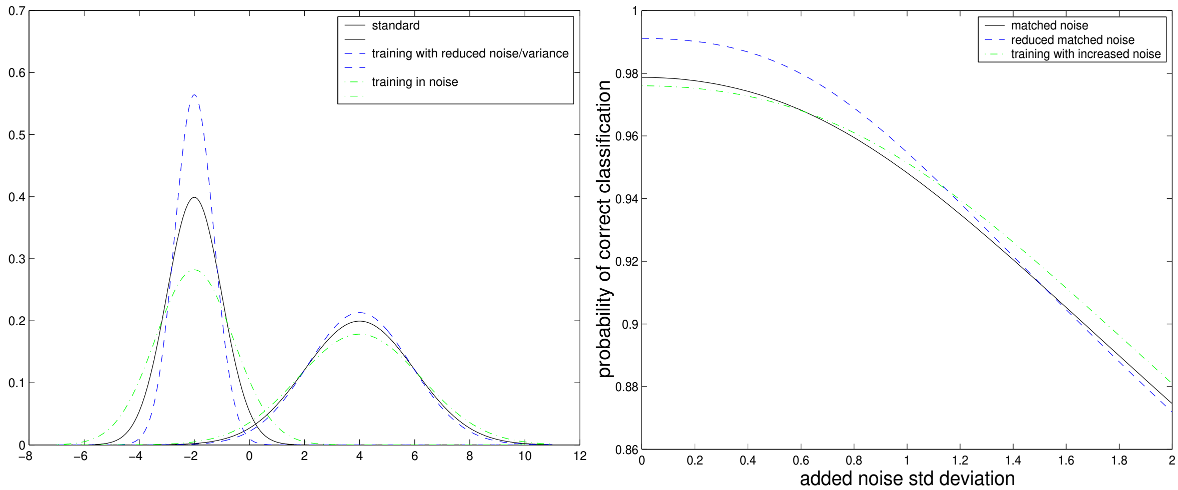

Fig.2a *Solid lines show Pdfs for two Gaussian data classes. Dash-dot line shows same with noise (non informative feature variance) added. Dashed line shows same with with noise removed.*

Fig.2b *P(correct classification) vs. noise level for two classes in Fig.2a (equal class priors). Solid line = matched noise. Dash-dot = training with added noise. Dashed = reduced matched noise.*

1.3 Data Pdf Estimation

As with any form of communication process, redundancy in the signal can sometimes offer the potential for lossless transmission. This is particularly the case in artificially engineered digital systems in which suitable coding can often proof against worst case distortion. Non speech interference can often be identified by a number of generic non speech characteristics (near stationarity, incoherence, etc.) and removed, without significant loss of speech information. Other sources of unmodelled variation, such as unfamiliar speakers, emotional states or speech rate, can also be reduced to some extent by “normalising” for specific factors. All of these proven techniques have their place in contributing towards observation pdf estimation.

In a natural system such as speech communication, complete removal of signal distortion, so that data mismatch is entirely eliminated, is not an option. Some parts of the spectro-temporal signal will invariably be masked by non speech interference to a level from which the clean speech signal is not recoverable. And some types of clean speech variation will always remain unmodelled. It follows that:

- preprocessing (observation pdf estimation) should correct the signal wherever possible (pdf mode = estimated clean value), but where not possible, mismatch should be detected but not “corrected”.
- the observation pdf should be minimally informative, i.e. have maximum possible entropy while accommodating all known constraints [6].
- no amount of robust preprocessing will ever be able to reduce the variance of more than a small proportion of observation data pdfs to near zero.

For the purpose of speech or image enhancement for listening to or viewing by humans, it is only the corrected signal (the observation pdf mode) which is relevant. But for feature enhancement prior to machine recognition, single data values leave no clue at all as to the precision of each value, or to the relative evidential weight which each value should carry, so a fuller representation of the data pdf needs to be retained.

In summary, the data pdf, which quantifies our relative degree of belief in each value which the feature could take, is normally represented extremely crudely by a single numerical value. The only pdf with just one parameter is the Dirac delta pdf ($\delta(x-a) = 0$ everywhere, except at $x = a$, and $\int\delta(x-a)dx = 1$), which describes deterministic data -while speech data is inherently indeterministic.

But what are the possibilities for both generating and making practical use of softer, more pdf-like, features? How difficult is it to estimate a complete pdf for each data feature, in place of the usual single value? The difficulty in estimating a complete pdf for each data feature lies in general somewhere between that of staying with the original single inaccurate value (no difficulty at all, but not very useful), and estimating the accurate value with 100% precision (extremely useful, but impossible).

Data reliability, or full pdf estimation, should be as accurate as possible, but strong benefits can often be gained even when the reliability measure is only very approximate. In situations like speech recognition, where feature data is highly redundant, the benefit of excluding a lot of mismatched data can easily outweigh the cost of losing a smaller amount of useful data (see Fig.3).

1.4 Simple Data Weighting

Various approaches to simple data weighting can be applied, according to the application area. One is to replace each single value by two values, where the old value specifies the pdf mean or mode, and the new value specifies its (Gaussian) variance. The other simple pdf which can often be used for this purpose is the (bounded) uniform distribution. However, while analytically tractable, white Gaussian or uniform noise do not well model every kind of data uncertainty.

2. MISSING FEATURE THEORY IN ASR

2.1 Motivation for Use of MFT in ASR

Interest in the “missing data” approach in speech recognition was initially motivated by studies of human speech perception. In the 1950’s the interest in this area arose from the development of speech coding techniques required in telecommunication [2,9], as did the whole basis for present day communication theory [22].

In particular it was observed that we are able to recognise many of the different speech sounds at extremely low SNRs, and this is largely due to the level of redundancy in the spectral data delivered into the auditory nerve by the peripheral auditory system. It was shown that we are able to recognise speech from a very small proportion of clean frequency channels at any one point in time. These experiments did not show how we are able to decide which channels are reliable. But they did prove that it must be possible to do so.

With the advent of computer speech recognition in the 1970’s, the need arose for techniques to improve the robustness to unpredictable signal variations. This resulted in several techniques for robust feature extraction, and for feature or model adaptation to noise and speaker variation. Computer speed and memory capacity then rapidly increased to the point where affordable real time speech processing was no longer a problem. However, the performance of ASR in most natural environments remained problematic, and the 1980’s and ‘90s saw the emergence of the “missing feature” [12,16,18] and “multiband” [5,14,17] models for noise robust ASR, which both aim to exploit and further develop our understanding of the human auditory system.

A considerable body of theory exists concerning the problem of how to go about pattern recognition, or data analysis in general, when some of the expected input data is missing, because this problem is very old. However, none of these techniques was developed specifically for dealing with the missing data problem in the form in which it arises in ASR. Missing data methods suitable for this purpose were (re?)invented in [1,18].

The understanding and performance of these missing feature theory (MFT) based systems has been steadily advancing ever since, but it would be fair to say that they are still a long way from achieving their potential, i.e. human recognition performance. Two main problems holding up MFT based ASR at the moment are a need for

- a stronger theoretical basis for the principled application of MFT to ASR
- a practical solution to accurate modelling with unorthogonalised spectral features

2.2 Bayesian Optimal Classification with Missing Data

For given data X, the Bayesian classification decision function $d(X)$ is defined as the decision which minimises the expected loss, or "Bayes risk", $r(d)$, for a given loss function, $L(C, d(X))$, over all possible classes, C, and values of X.

In the case where none of the observation data X is uncertain, the Bayes decision function $d(X)$ which minimises the expected zero-one loss (loss = 0 if classification is correct, else = 1) and thereby also maximises the probability of correct classification, is the usual MAP (Maximum A Posteriori probability) decision:

$$d(X) = argmax_C P(C|X) \tag{1}$$

If a part of the observation data is missing or uncertain, and this uncertainty is specified quantitatively by the knowledge that $X \sim s(X|X^{OBS})$, then (Appendix A and [18]) the Bayes decision function is given by

$$d(X) = argmax_C E[P(C|X)|(X \sim s(X|X^{OBS}))] \tag{2}$$

This means that if we have a trained parametric classifier $P(Q|X, \Theta)$, but we cannot apply this directly to X because some part of the data in X is missing or otherwise uncertain, then the value which we should maximise in its place is the expected value of this parametric function,

$$\hat{P}(Q|X, \Theta) = E[P(Q|X, \Theta)|X \sim s(X|X^{OBS})] \tag{3}$$

This simple rule provides us with a well defined procedure for obtaining optimal classification with any kind of missing or uncertain data.

2.3 Two Approaches to Expected Class Posterior Probability Estimation

The best approach to take in evaluating this expectation depends on the form in which constraints on the uncertain data are available. If good/bad data separation is 100% reliable, "good" data X^G is perfect, and an accurate noise model pdf $\hat{p}(X^B|X^T)$ is available for the bad data as a function of the unknown true values X^T of the bad data [1,7], then one can try to make use of the identity

$$E[P(Q|X, \Theta)|X^B \sim \hat{p}(X^B|X^T)] = P(Q|X^G, X^B) \tag{4}$$

$$= \frac{\int \hat{P}(Q|X^G, X^T, \Theta)\hat{p}(X^G, X^T)\hat{p}(X^B|X^T)dX^T}{\int \hat{p}(X^G, X^T)\hat{p}(X^B|X^T)dX^T} \tag{5}$$

Eq.5 is an identity under the very general condition that $p(X^B|X^T) = p(X^B|Q, X^G, X^T)$. If $\hat{p}(X)$ can be obtained from $\hat{P}(Q|X, \Theta)$, and the noise model has a simple form (such as Gaussian, or uniform), then Eq.5 may yield a closed form solution. However, in the more general case where missing data detection is probabilistic rather than deterministic, it is more instructive [18] to go back to the explicit form of the expectation integral,

$$E[P(Q|X, \Theta)|X \sim s(X|X^{OBS})] = \int P(Q|X, \Theta)s(X|X^{OBS})dX \tag{6}$$

2.4 Application of MFT in Viterbi Decoding with HMMs

In order to describe the procedure required for HMM decoding with soft data, It is instructive to first recall the usual procedure used for HMM decoding with deterministic data.

2.4.1 Normal Viterbi Decoding

Let Θ denote the set of HMM model parameters (Gaussian parameters, transition probabilities, state priors and mixture weights). For given features, X, Viterbi decoding uses the MAP objective

$$Q^{best} = \arg max_Q P(Q|X, \Theta) \tag{7}$$

As data models have the form $P(X|Q, \Theta)$, we use Bayes' rule

$$P(Q|X, \Theta) = \frac{P(X|Q, \Theta)P(Q|\Theta)}{P(X|\Theta)} \tag{8}$$

$P(X|\Theta)$ is the same for any choice of Q, so $P(Q|X, \Theta) \propto P(Q|\Theta)P(X|Q, \Theta)$

The Markovian independence assumptios gives $p(X|Q_a, \Theta) \cong \prod_n p(x_n|q_{a(n)}, \Theta)$.

$p(x_n|q_{a(n)}, \Theta)$ for each state is modelled by a mixture pdf

$$p(x|q_k, \Theta) = \sum_j P(m_j|q_k, \Theta)p(x|m_j, q_k, \Theta) \tag{9}$$

where each pdf component $p(x|m_j, q_k, \Theta)$ is usually a multivariate diagonal covariance Gaussian, specified by its mean vector μ_{jk} and variance vector σ^2_{jk}.

$P(Q)$ is modelled by transition probabilities: $P(Q_a|\Theta) = P(q_{a(1)})\prod_{n=2}^{N} P(q_{a(n)}|q_{a(n-1)})$

2.4.2 Viterbi Decoding with Soft Data

Eq.3 tells us that, with missing data, we should replace the usual posterior $P(Q|X)$ by its expected value, subject to the posterior observation pdf

$$\hat{P}(Q|X,\Theta) = E[P(Q|X,\Theta)|X \sim s(X|X^{OBS})]$$

Making use of Eq.6 for evaluating the expected value of $\hat{P}(Q|X,\Theta)$ gives us

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

$$= P(Q|\Theta)\int \frac{p(X|Q,\Theta)}{p(X|\Theta)}s(X|X^{OBS})dX \tag{10}$$

In the present context we will restrict our attention to the case of speech recognition with non negative spectral features, where each feature may or may not be dominated by additive noise.

Let $s'(X)$ denote the "evidence" pdf which is conditioned by the observed noisy data X^{OBS} together with the constraint (which applies to non negative features in additive noise) that the clean feature value cannot be negative, or greater than the observed noisy value, i.e. $X \in [0, X^{OBS}]$.

The posterior observation pdf $s(X)$ is conditioned by the evidence pdf, and the clean data prior, $p(X|\Theta)$, for which HMM models have been trained on "clean speech". Note here than mismatch would be minimised if the procedure used for training on clean speech also made use of the same soft data representation that is used during recognition.

Information in the evidence pdf is independent of information in the clean data prior, so the required posterior observation pdf, $s(X)$, can be obtained from the evidence and prior pdfs as follows:

$$s(X) = cs'(X)p(X|\Theta) \tag{11}$$

where c is just a normalising constant. Providing c is not permitted to depend on Q, and both c and $p(X|\Theta)$ are non zero, substitution of Eq.11 into Eq.10 gives us:

$$\hat{P}(Q|X,\Theta) \propto P(Q|\Theta)\int p(X|Q,\Theta)s'(X)dX \tag{12}$$

Transition probabilities are taken care of in recognition with missing data in exactly the same way as they are with normal recognition. The usual Markovian independence assumption, together with the assumption that $s'(X) \cong \prod_n s'(x_n)$, gives us

$$\int p(X|Q_a,\Theta)s'(X)dX = \int\prod_n p(x_n|q_{a(n)},\Theta)s'(x_n)dx_n \tag{13}$$

$$= \prod_n\int p(x_n|q_{a(n)},\Theta)s'(x_n)dx_n \tag{14}$$

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.9 is replaced by $\int p(x|m_j, q_k, \Theta)s'(x)dx$.

3. RECOGNITION EXPERIMENTS

3.1 Test Data

Tests were made on the Aurora 2.0 speaker independent connected digit recognition task [20], using the clean training set, and test set (a). Training HMMs used the full 8440 utterance training set, with no cross validation. HMMs used were the 12 standard 16 state whole word models specified for the Aurora task. Training started with 1-mix models from a flat start, then split from 1 to 2 to 3 to 5 to 7 mix components, using 4 Baum-Welch iterations with each number of mix components. The test set comprises 100 utterances for each of 4 noise types (subway, babble, car, exhibition) at 4 noise levels (clean, SNR 20dB, SNR 10dB, SNR 0dB). All results are shown averaged over the 4 noise types.

3.2 Evidence Pdf Models

Assuming $s'(x) = \prod_i s'(x_i)$, and modelling the evidence pdf by a two component mixture pdf,

$$s'(x_i) = \hat{p}_i f_1(x_i) + (1-\hat{p}_i) f_2(x_i)$$

$$\int p(x_i|m_{ij}, q_{ik}, \Theta) s'(x_i) dx_i = \int p(x_i|m_{ij}, q_{ik}, \Theta)(\hat{p}_i f_1(x_i) + (1-\hat{p}_i) f_2(x_i)) dx_i$$

$$= \hat{p}_i \int p(x_i|m_{ij}, q_{ik}, \Theta) f_1(x_i) dx_i + (1-\hat{p}_i) \int p(x_i|m_{ij}, q_{ik}, \Theta) f_2(x_i) dx_i \qquad (15)$$

Mixture component pdfs should have maximum entropy under the constraints which apply, and (for practical reasons) should provide closed form integrals in Eq.15. The four pdfs used here were

- Dirac-delta $\delta(x-a)$, maxent when observation has negligible variance
- uniform $u(a, b)$, maxent when observation has known bounds $[a, b]$
- Gaussian $g(x, \mu, \sigma^2)$, maxent when observation has known mean and variance (μ, σ^2)
- bounded Gaussian $g(x, \mu, \sigma^2) / \int_a^b g(x, \mu, \sigma^2)$, maxent for known mean, var. and bounds

3.3 Mixture Weight (Missing-Data Mask) Estimation

Mixture weights $\hat{p}_i$ are estimates of the probability that each spectral observation x_i is dominated by clean speech, i.e. that x_i is near clean. These probability estimates are based on the assumption that for the compression function φ used, $\varphi(speech + noise) \cong \varphi(max(speech, noise))$. Under this assumption it follows that when a value is dominated by speech, its observed value x_i^{OBS} will be close to its clean value. However, when a value is not dominated by speech, then it could either be dominated by noise, or it could be a near equal speech-noise sum. In either case, there is very little which can be said about the clean speech value, except that it is somewhere in $[0, x_i^{OBS}]$.

Both the estimated $P(clean)$ "mask" values and the best performing model (method 6) used here were taken from another recent study [3]. The method used to estimate these mask weights was based on a simple procedure sometimes used for speech enhancement by spectral subtraction, in which the noise spectrum is estimated as the average of the first 100 ms of spectral frames in the utterance. "Hard" 0/1 mask values for each spectral coefficient in the utterance are then set to1 if the observed value is greater than the corresponding estimated noise value, or else to 0. See [8,15] for more accurate (and more complicated) methods for noise estimation. See [3] for details of "soft" mask estimation once the noise estimate has been obtained.

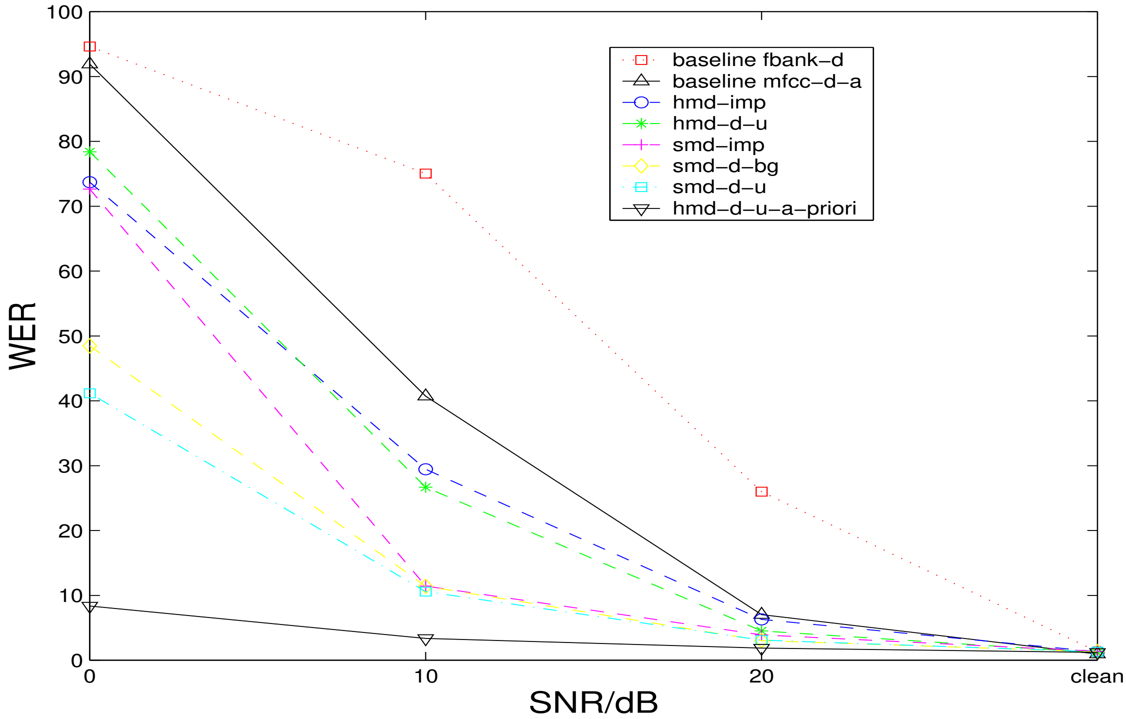

***Fig.3** WER vs. SNR for two baseline deterministic HMM based ASR systems and six missing data based recognition schemes as listed below (five using simple reliability mask estimation, the last using a-priori mask). Task is the Aurora 2.0 connected digits recognition task. Results are averaged over 4 noise types.*

Two baseline deterministic HMM based ASR systems, and six soft data based systems, were tested.

1. **baseline mfcc-d-a** commonly used orthogonalised acoustic features

2. **baseline fbank-d** unorthogonalised filterbank features used with all MD methods

3. **hmd-imp** hard missing data ($\hat{p}_i = 0$ or 1 only), with data imputation

   $f_1(x_i) = \delta(x - x_i^{OBS})$, $f_2(x_i) = \delta(x - \mu_{ijk})$

4. **hmd-d-u** hard missing data ($\hat{p}_i = 0$ or 1 only), with delta-uniform mix pdf

   $f_1(x_i) = \delta(x - x_i^{OBS})$, $f_2(x_i) = u(0, x_i^{OBS})$

5. **smd-imp** soft missing data (SMD) with data imputation (delta-delta mix pdf)

   $f_1(x_i) = \delta(x - x_i^{OBS})$, $f_2(x_i) = \delta(x - \mu_{ijk})$

6. **smd-d-u** SMD with delta-uniform mix pdf (from [3])

   $f_1(x_i) = \delta(x - x_i^{OBS})$, $f_2(x_i) = u(0, x_i^{OBS})$

7. **smd-d-bg** SMD with delta-bounded Gaussian mix pdf, $\mu = x/2, \sigma^2 = x^2$

   $f_1(x_i) = \delta(x - x_i^{OBS})$, $f_2(x_i) = g(x_i^{OBS}, \mu_{ijk}, \sigma^2_{ijk})$

8. **hmd-d-u-a-priori** hard missing data, with delta-uniform mix pdf, ($\hat{p}_i = 0$ or 1, a-priori)

   $f_1(x_i) = \delta(x - x_i^{OBS})$, $f_2(x_i) = u(0, x_i^{OBS})$

4. DISCUSSION

The tests reported here do not conform to all of the recommendations for recognition with soft data which were discussed in Section 1.3. In particular, they do not make use of available state of the art noise estimation techniques. Had they done so, the baseline techniques would have performed much better and the advantage of the soft-missing-data approach might have been less pronounced. However, the point to note is that the SMD approach has a theoretical advantage over the baseline methods, because it makes a more complete use of the available evidence - it uses observation accuracy values as well as the observations themselves.

The best results were obtained by method 6, using the delta-uniform mix pdf. The runner up is method 7, using the delta-bounded-Gaussian mix pdf. In the limit as the variance used with this Gaussian gets larger, this method becomes identical to method 6. The arbitrary imposition of a finite evidence pdf variance goes against the principle that the evidence pdf should have maximum entropy while accommodating all of the available evidence. Of the other methods, “data imputation” simply replaces “missing” values with their state means, thereby unjustifiably imposing a zero variance on the evidence pdf. Softening of the missing-data decision (i.e. replacing 0/1 $P(clean)$ mask values by values over $[0, 1]$) helps greatly in every case.

Our use of a 2-mixture component pdf is tied to the idea that each observation can be said to be clean (i.e. no significant mismatch) or not clean. When some kind of more sophisticated technique is used for generating clean spectral data, the resulting evidence pdf will no longer be possible to represent as a delta-uniform mix pdf. The mode will be at the estimated spectral value, and probability density will fall off smoothly with distance from the mode. Some potentially interesting pdfs (like the beta pdf which is often used in Bayesian estimation) do not provide closed form integrals in Eq.15. The evidence pdf may therefore have to be approximated by a bounded Gaussian, or more generally as a mixture pdf, using the same component pdfs that were used here.

Mismatch is not the only reason for downweighting data. Data can also be “cleaned” (prior to both training and recognition) by filtering out data which can be detected as less discriminative. Discriminative data is often correlated with easily identifiable spectro-temporal features [10,11,19,21] such as formant peaks and formant transitions. This would also fit within the present framework.

One problem with SMD based ASR is that it normally requires recognition using highly correlated spectral features, while for practical reasons it is not feasible to accurately model this correlation.

5. CONCLUSION

We have presented a case for extending the “missing data” approach to noise robust speech recognition to a general model for robust recognition which acknowledges the need to model data uncertainty and minimise data mismatch. A number of previously developed “hard-” and “soft-” missing-data based ASR techniques were interpreted, implemented and tested within this framework. The next tasks required to increase the efficiency of this approach are to implement maximum likelihood HMM training with soft data, and to incorporate an existing high performance noise removal procedure into the evidence pdf estimation procedure.

Acknowledgements

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). Web site: http://www.dcs.shef.ac.uk/research/groups/spandh/projects/respite/

Appendix A: Derivation of Bayes Decision Rule for Uncertain Data

Let s denote the knowledge that data $x \sim s(x|x^{OBS})$ rather than having a deterministic value.

The Bayesian approach [4,7] to deriving an optimum class decision $d(s)$ for a given s is to first specify a suitable quantitative *loss function*, $L(C, d(s))$, for every possible true class C and decided class $d(s)$, and then to minimise the overall expected loss or “Bayes risk”, $r(d(s))$, with respect to this loss function and the posterior pdf, $P(C|s)$.

In all or nothing classification, correct classification is assigned loss 0, and incorrect classification loss 1 (known as “zero-one loss”). The “Bayes risk” is then given by

$$r(d(s)) \;=\; E[L(C, d(s))|x \sim s] \tag{16}$$

$$= \sum_C \int_x L(C, d(s))p(C, x|s)dx \tag{17}$$

$$= \sum_C \int_x L(C, d(s))P(C|x, s)p(x|s)dx \tag{18}$$

$$= \int_x [\sum_C L(C, d(s))P(C|x, s)]s(x)dx \tag{19}$$

When $d(s)$ selects correct class $C°$, with zero-one loss,

$$\sum_C L(C, d(s))P(C|x, s) \;=\; \sum_{C \neq C°} P(C|x, s) \tag{20}$$

$$= (1 - P(C°|x, s)), \tag{21}$$

With deterministic data $s(x) = \delta(x - x^{OBS})$ Bayes risk is just the probability of misclassification,

$$r(d(s)) \;=\; \int_x (1 - P(C°|x))\delta(x - x^{OBS})dx \;=\; (1 - P(C°|x^{OBS})) \tag{22}$$

which is minimised by the usual MAP rule

$$d(x) \;=\; argmax_C P(C|x) \tag{23}$$

With uncertain data the Bayes risk is given by

$$r(d(s)) \;=\; \int_x (1 - P(C°|x, s))s(x)dx \;=\; 1 - \int_x P(C°|x, s)s(x)dx \tag{24}$$

which is minimised by the Bayes decision:

$$d(s) \;=\; argmax_C \int_x P(C|x, s)s(x)dx \tag{25}$$

$$= argmax_C \int_x P(C|x)s(x)dx \tag{26}$$

$$= argmax_C E[P(C|x)|s] \text{, QED} \tag{27}$$

This proof is adapted from [18].

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

[5] Bourlard, H. & Dupont, S. (1996) "A new ASR approach based on independent processing and recombination of partial frequency bands", Proc. ICSLP'96, Philadelphia, pp. 422-425.

[6] Deco, G. & Obradovic, D. (1996) An Information-Theoretic Approach to Neural Computing, Springer.

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

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

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

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

[11] 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.

[12] 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.

[13] Hermansky, H. and Morgan, N. (1994) "RASTA Processing of Speech", in IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, pp. 578-589.

[14] Hermansky, H., Tibrewela, S. & Pavel, M. (1996) "Towards ASR on partially corrupted speech", Proc ICSLP'96, pp. 462-465.

[15] Hirsch, H. G. and C. Ehrlicher (1995) "Noise estimation techniques for robust speech recognition", ICASSP95 , 1995, pp. 153-156.

[16] 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

[17] Morgan, N., Bourlard, H. & Hermansky, H. (1998) "Automatic speech recognition: an auditory perspective", Research Report IDIAP-RR 98-17.

[18] 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.

[19] 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.

[20] 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.

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

[22] Shannon, C. E. & Weaver, W. (1940) The mathematical theory of information, Univ. of Illinois Press, Urbana Ill.