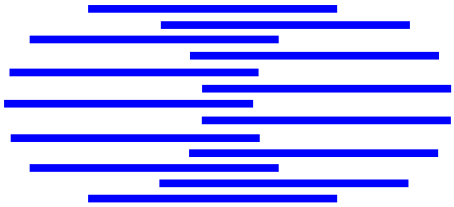
IDIAP COMMUNICATION

IDIAP

Martigny - Valais - Suisse



LATENT VARIABLE DECOMPOSITION FOR POSTERIORS OR LIKELIHOOD BASED SUBBAND ASR

Andrew C. Morris

IDIAP-Com 99-04

November 1999

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

IDIAP-Com 99-04

LATENT VARIABLE DECOMPOSITION FOR POSTERIORS OR LIKELIHOOD BASED SUBBAND ASR

Andrew C. Morris

November 1999

Abstract

Latent variable decomposition permits factorisation of posterior probability based, or likelihood based, speech unit discriminant functions into a composition of simpler functions which can be analysed separately and evaluated more accurately in the presence of band-limited noise, or other source of data mismatch. See [2,7] for a more self contained introduction to this subject. In this report we present the essential theoretical issues, and implementation details, for the key points concerning this approach to multiband ASR. In particular, we show that the posteriors and likelihood based multiband decompositions are very closely linked.

Acknowledgements: This work was supported by the EC/OFES RESPITE (REcognition of Speech by Partial Information TEchniques) project.

Contents

1. Introduction

When working with subband ASR (Automatic Speech Recognition) it is important to be aware of the core ideas underlying this approach to noise robustness. It is also important to make note of the different advantages of likelihood (HMM, Hidden Markov Model) and posteriors (HMM/ANN, HMM/Artificial Neural Network) based subband ASR models. In MB (Multi-Band) ASR it is required to combine the outputs from a number of separate subband experts into a single expert "opinion" or quantitative measure of some kind. In most cases the experts are combined for each data frame. With **posteriors based experts** the discriminant function to be evaluated for each data frame x^n, and phoneme q_k, is $P(q_k|x)$. These posteriors are usually modelled by an MLP neural network, training uses the LMSE, Least Mean Square Error, criterion, which has been shown to maximise the probability of correct classification and therefore makes use of state priors $P(q_k)$. With **likelihood based experts** the discriminant function to be evaluated for each frame is $p(x|q_k)$. In this case training is ML (Maximum Likelihood) based, and uses prior information only indirectly as state-transition probabilities.

As with the simple missing data (MD) based approach to noise robust ASR [6], one of the main aims of multiband ASR (MB) is to achieve robust recognition in the presence of band limited noise **by exploiting spectral data redundancy**. The key difference between the MB and MD approaches is that the MB approach uses a reliability weighted sum over **many (if not all possible) positions of unreliable data**, while the simple MD approach makes a hard decision about which data is reliable - data is either treated as 100% clean, or else it is ignored completely. Another very important advantage of MB over MD arises from the fact that competitive performance in clean speech requires that data features used in recognition are near-orthogonal. The data subset presented to each MB expert can be orthogonalised, because a separate expert is trained for every possible position of unreliable data. With MD orthogonalisation is not possible, because only one fullband expert is trained, and fullband orthogonalisation would mix all data components together, after which separation of reliable from unreliable data would no longer be possible.

Both the MD and MB approaches to robust ASR have the advantage that they **require no knowledge of the noise**. Models trained on clean speech can be used to recognise noisy speech, **provided only that some parts of the signal remain clean, and these parts can be identified**. Early experiments with MB combined one expert trained on each subband. Different combination functions were tested, some linear (linear weighted sum), and others nonlinear (geometrically weighted product, or MLP combination). However, these early models were disadvantaged by the fact that **each narrow-band expert has a greatly reduced performance compared to the fullband expert in clean speech**, so that the optimal combined narrow-band expert performance was very limited.

The aim of **latent variable decomposition** here is to theoretically factorise the posterior or likelihood based fullband discriminant function into a composition of simpler functions which can be analysed separately and evaluated more accurately in the presence of band-limited data mismatch. In the case of multiband ASR, the fullband data likelihood (or posterior) can be decomposed into a **weighted sum of clean-data likelihoods** (or posteriors) for each subset of data subbands, in which weights represent the probability that each data subset has negligible data mismatch.

In Section 2 we show how both the posteriors and likelihood based fullband discriminant functions can decomposed into a sum of terms, one for each subband combination, comprising: a subband-combination discriminant function, (which is conditioned on the data in this subband combination being free from data mismatch), and a corresponding mismatch sensitive weighting factor. In Section 3 we show how data mismatch and relative utility can be further decomposed into separate factors. In Section 4 we summarise a number of different approaches to weighting estimation. In Section 5 we expand on some of the technical considerations needed for implementing these methods.

2. Latent variable decomposition

Providing we consider all possible subband combinations (including the empty set) the events[1] c_i are exhaustive, so that $P(\bigcup_i c_i) = 1$. As only one subband combination can be that largest clean set, c_i are also mutually exclusive, so that $P(c_i \cap c_j) = 0$ when $i \neq j$, and $P(c_i \cup c_j) = P(c_i) + P(c_j)$ Note that when manipulating probabilities it is essential that the "events" for which probabilities are being calculated have a **precise formal definition**.

2.1 Posteriors decomposition

For a posteriors based system the function to be evaluated for each (x, q_k) is $P(q_k|x)$. As $\bigcup_i c_i$ is exhaustive:

$$P(q_k|x) = P(q_k \cap \bigcup_i c_i \Big| x) \tag{1}$$

and as c_i are mutually exclusive:

$$P(q_k \cap \bigcup_i c_i \Big| x) = \sum_i P(q_k \cap c_i|x) \tag{2}$$

Therefore, using Bayes' rule:

$$P(q_k|x) = \sum_i P(c_i|x) P(q_k|c_i \cap x) \tag{3}$$

The original quantity to be evaluated, $P(q_k|x)$, is now decomposed into weighting terms $P(c_i|x)$, and subband combination posteriors $P(q_k|c_i \cap x)$. Unlike the original fullband posterior, **the terms $P(q_k|c_i \cap x)$ are conditioned on the data in the subband subset being clean**, and the data not in this subset being not clean. Evaluation of the weighting terms is discussed in Section 4.

While it is "intuitively clear" that $P(q_k|c_i \cap x) \cong P(q_k|s_i)$, formulas based on intuition are not always correct, and in Section 2.1.1 we show under what conditions this is true.

2.1.1 Optimal estimate of incomplete data posterior

Here we borrow some terminology from MFT (missing feature theory) and refer to clean (no mismatch) data as "certain" data, and unclean data as "uncertain data". When the unclean data is to be ignored completely, we will refer it as "missing data". The "clean" or "certain" data components in the subset s_i of x will be denoted by x_c (dropping the subscript i), and uncertain components by x_u. Note here that x_c is a vector of **given scalar values**, while x_u is a vector of **random variables**, whose range of possible values is constrained only knowledge κ_u, which can be represented by a pdf.

The condition c_i gives us the partition of x into certain and uncertain parts (x_c, x_u). If **nothing** is known about the uncertain data, then this data is simply missing, or "not given", so that:

$$P(q_k|c_i \cap x) = P(q_k|x_c, \kappa_u) = P(q_k|x_c) = P(q_k|s_i), \text{ so that, from. Eq. 3,} \tag{4}$$

$$\hat{P}(q_k|x, \kappa) = \sum_i P(c_i|x) P(q_k|s_i) \tag{5}$$

1. *See Appendix A: "Notation", page 21 for definition of "x clean", c_i, s_i, etc.*

Clean subband posteriors, $P(q_k|x_c)$, can be obtained by training a separate posteriors based classifier on data from each subband combination. This classifier used is commonly an MLP, but could also be implemented by training a likelihood based classifier for each combination, then using Bayes' rule, plus class priors, to convert to posteriors[1].

So far we have assumed that the uncertain data is completely unconstrained, and therefore effectively missing. However, it is often the case that some knowledge constraining the uncertain data is available, and we should try to make as much use of this knowledge as possible. If the knowledge κ which we have about x is a partition into certain and uncertain parts (x_c, x_u), plus some further knowledge κ_u constraining, but not fully specifying, x_u, then what is the optimal estimate for $P(q_k|x, \kappa)$?

In the general field of function approximation, when the value of x is uncertain (i.e. when x is a random variable), the estimate for any function $\theta(x)$ of x which minimises the "expected quadratic error" $E[(\theta - x)^2]$ (the most common criterion used in the estimation of continuous statistics) is given by $\hat{\theta}(x) = E(\theta)$[2], or, in the case where the pdf for x is conditioned by knowledge κ, by: $\hat{\theta}(x) = E[\theta(x)|\kappa]$.

$P(q_k|x, \kappa)$ can be estimated as its conditional expected value as follows:

$$\hat{P}(q_k|x, \kappa) = \hat{P}(q_k|x_c, \kappa_u) = E[P(q_k|x_c, \kappa_u)|x_c, \kappa_u] = E\left[\frac{p(x_c, \kappa_u|q_k)P(q_k)}{p(x_c, \kappa_u)}\middle| x_c, \kappa_u\right] \quad (6)$$

$$= E\left[\frac{p(x_c|q_k)P(\kappa_u|x_c, q_k)P(q_k)}{p(x_c)P(\kappa_u|x_c)}\middle| x_c, \kappa_u\right] = P(q_k|x_c)\frac{P(\kappa_u|x_c, q_k)}{P(\kappa_u|x_c)} \quad (7)$$

If nothing is known about x_u then $P(\kappa_u|anything) = 1$, so the optimal estimate is the "marginal posterior" $P(q_k|x_c)$, as in Eq. 4. However, if the uncertain data is constrained in some way (see Section 5.3) then use of the second term in Eq. 7 can strongly improve recognition performance, especially when a large proportion of data is uncertain [Eq. 6].

2.2 Likelihood decomposition

For a likelihood based system the function to be evaluated for each (x, q_k) is $p(x|q_k)$. Decomposing as with posteriors, we obtain:

$$p(x|q_k) = \sum_i p(x \cap c_i|q_k) = \sum_i P(c_i|q_k)p(x|c_i \cap q_k) \quad (8)$$

This decomposition as it stands presents two problems:

2.2.1 Problem 1: Optimal estimate of incomplete data likelihood

It is **not** true that $p(x|c_i \cap q_k) \cong p(s_i|q_k)$. Intuitively speaking, this is because $p(x|q_k) \cong \prod_j p(x_j|q_k)$ and the reduced dimension of s_i with respect to that of x guarantees that $p(s_i|q_k)$ is much smaller or larger than $p(x|q_k)$, depending on whether each term in the product is less or greater than 1 (they are most often less than 1).

1. *This later system is sometimes known as a Gaussian Radial Basis Function classifier.*
2. *This is well known, but can be shown simply by differentiating the quadratic error w.r.t.* θ *and equating the result to zero.*

This problem can be solved without much difficulty using the same method as in Section 2.1.1 for posteriors. The knowledge we have in the present case about x_u is the knowledge we had in the posteriors case, plus the knowledge that x is from class q_k. The appropriate estimate for $p(x|c_i \cap q_k)$ is therefore obtained as follows:

$$\hat{p}(x|c_i \cap q_k) = E[p(x|c_i, q_k)\big|c_i, q_k] = E[p(x_c, x_u\big|\kappa_u, q_k)\big|x_c, \kappa_u, q_k]$$

$$= E[p(x_c\big|\kappa_u, q_k)p(x_u\big|x_c, \kappa_u, q_k)\big|x_c, \kappa_u, q_k] = p(x_c\big|q_k)E[p(x_u\big|x_c, \kappa_u, q_k)\big|x_c, \kappa_u, q_k]$$

$$= p(x_c\big|q_k)\int_{R_u} p(x_u\big|x_c, \kappa_u, q_k)p(x_u\big|x_c, \kappa_u, q_k)dx_u \tag{9}$$

For $p(x|q_k)$ diagonal covariance GM, both the marginal and the integral in Eq. 9 are evaluated in terms of available quantities in Section 5.

2.2.2 Problem 2: Obtaining weights which are sensitive to data mismatch

The decomposed likelihood terms $p(x|c_i \cap q_k)$ are conditioned on the data being clean, which is what we need for making use of experts trained on clean data. However, the weighting terms $P(c_i\big|q_k)$ in Eq. 8 are conditioned only on the speech unit identity. Different phonemes are likely to be affected differently by noise in different channels, so phoneme identity may carry a certain amount of information on whether the data in each subband combination should be considered as clean or not. However, weights conditioned on phoneme identity alone clearly have far less potential for selecting the most appropriate subband combination in the presence of strong band limited noise. There is no alternative decomposition of the data likelihood to give weights conditioned on x. **The only way out of this problem is therefore to make use of the posteriors decomposition, in which noise-conditioned weights arise naturally**. One way to do this would be to simply abandon likelihood based models in favour of posteriors based models, such as the HMM/ANN. However, likelihoods can have some advantages over posteriors (see Section 6), so a more constructive alternative, would be to use the link between “scaled likelihoods” and “scaled posteriors” given by Bayes’ rule, together with Eq. 5 (or Eq. 7, if it is required to apply the bounds constraint), to give:

$$\frac{p(x|q_k)}{p(x)} = \frac{P(q_k\big|x)}{P(q_k)} \cong \frac{\sum_i P(c_i\big|x)P(q_k\big|s_i)}{P(q_k)} = \sum_i P(c_i\big|x)\frac{p(s_i\big|q_k)}{p(s_i)} \tag{10}$$

Providing that $p(s_i\big|q_k)$ is evaluated for all k (this would preclude the use of beam search), $p(s_i)$ in Eq. 10 can be obtained directly using $p(s_i) = \sum_k P(q_k)p(s_i\big|q_k)$.

If required, $p(x)$ on the left side of Eq. 10 could also be obtained similarly after $p(x|q_k)$ has been evaluated for all k. However, as $p(x)$ is independent of q_k, and so is $p(X) = \prod_{n=1}^{N} p(x^n)$, so this factor can apparently be ignored However, if a language model is used, then the LM scaling factor would now be proportional to $\log p(x)$, which could vary greatly between utterances, so in this case it may be better if $p(x)$ is not ignored.

Note here that if Eq. 10 is used in place of Eq. 8 then the complex scaling factors given in Eq. 9 which are required for estimating the incomplete data likelihoods in Eq. 8 are no longer needed.

3. Decomposition with two latent variables

It is possible to perform decomposition using separate latent variables to specify which bands have reliable data and which bands are most effective for distinguishing speech units. This is desirable because (presence of data mismatch) and (utility of a particular set of clean subbands for distinguishing speech units) are different properties of the signal which it may be best to detect separately.

We start as for the decomposition with one latent variable, except here we have two independent latent variables, so we need a double sum to sum over all possible combinations. For posteriors we have:

$$P(q_k|x) \cong P(q_k \cap \bigcup_i b_i \cap \bigcup_j c_j | x) = \sum_i \sum_j P(q_k \cap b_i \cap c_j | x) \quad (11)$$

$$= \sum_i \sum_j P(c_j|x) P(b_i|c_j \cap x) P(q_k|b_i \cap c_j \cap x) \quad (12)$$

If we assume that x adds little new information about which combination is best when c_j is known, then $P(b_i|c_j \cap x) \cong P(b_i|c_j) = b_{ij}$. These fixed weights, are estimated as in Section 4.2. In the third term in Eq. 12 we can assume that the best set is a subset of the largest set of clean subbands, so $P(b_i|c_j) = 0$, and the last term in Eq. 12 can be ignored, unless $s_i \subseteq s_j$. In the remaining cases q_k is conditionally independent of c_j. Therefore:

$$P(q_k|x) \cong \sum_i P(q_k|s_i) \sum_j P(c_j|x) P(b_i|c_j) \quad (13)$$

For likelihoods, we have:

$$P(x|q_k) \cong P(x \cap \bigcup_i b_i \cap \bigcup_j c_j | q_k) = \sum_i \sum_j P(x \cap b_i \cap c_j | q_k) \quad (14)$$

$$= \sum_i \sum_j P(c_j|q_k) P(b_i|c_j \cap q_k) P(x|b_i \cap c_j \cap q_k) \quad (15)$$

In Eq. 15 c_j is unlikely to be dependent on q_k, so $P(c_j|q_k) \cong P(c_j)$. If nothing is known about the noise then $P(c_j)$ are unknown, but may be estimated using Eq. 24 with the assumption that each subband is equally likely to be noisy[1]. The second term is like the second term in Eq. 13, but is further conditioned on the speech unit whose likelihood we are estimating. This could be a distinct advantage, because different speech units are sure to have different distinguishing features in different subbands. In the third term in Eq. 15 we can drop the dependence on c_j for the same reason that it was dropped in Eq. 13. Therefore:

$$P(x|q_k) = \sum_i \sum_j P(c_j) P(b_i|c_j \cap q_k) P(x|b_i \cap q_k) \quad (16)$$

Eq. 16 presents the same problems that were discussed in Section 2.2, but the static weights in the two LV case will be more accurate. Alternatively, we could take the approach of Section 2.2.2 and use:

1. *Note that $P(c_j)$ will not be all equal in this case unless $d = 2$.*

$$\frac{p(x|q_k)}{p(x)} \cong \sum_i P(c_i|x) \frac{p(s_i|q_k)}{p(s_i)} \sum_j P(b_i|c_j) \tag{17}$$

4. Weight estimation

If the latent variable c in Eq. Eq. 5 is not an indicator for each subband combination being the largest clean combination, then we are no longer justified in assuming that $P(q_k|s_i) \cong \hat{P}(q_k|s_i;\Theta_i)$. On the other hand, if we can assume that the data is clean, and we are interested in exploiting the possibility that data in some subbands should carry more weight than data in others, then we can replace c by the latent variable b, which indicates whether each subband combination is "the best" or most useful. As b_i are also mutually exclusive and exhaustive, the same working used to derive Eq. Eq. 5 also gives us:

$$P(q_k|x) \cong \sum_i P(b_i|x) \hat{P}(q_k|s_i;\Theta_i) \tag{18}$$

Although the weights $P(b_i|x)$ in Eq. 18 depend on x, if we are interested in using fixed weights, then we must ignore x and assume that $\hat{P}(b_i|x) \cong \hat{P}(b_i)$. In the dual decomposition of Section 3, the fixed weighting factors are $P(b_i|c_j)$.

For a likelihood based system the same argument gives us:

$$p(x|q_k) \cong \sum_i P(b_i|q_k) \hat{p}(x|b_i \cap q_k) \tag{19}$$

or, in the case of the dual decomposition of Section 3, the fixed weighting factors are $P(b_i|c_j \cap q_k)$

If the data is clean then the weights are static and can be estimated from the training data set. We describe here a number of approaches to both fixed and adaptive weighting.

4.1 Fixed weight estimation using linear & non-linear LMSE

Fixed weights $w_i = P(b_i)$, or $w_{ik} = P(b_i|q_k)$, $i = 1 \ldots 2^d$, $k = 1 \ldots K$, can be estimated using the (supervised) LMSE (least mean square error) criterion:

$$w = \arg min_w \sum_{n=1}^{N} \sum_{k=1}^{2^d} (y_k^n(w) - t_k^n)^2 \tag{20}$$

where $y_k^n(w)$ is the estimated combined posterior for q_k at frame n, and $t_k^n = P(q_k|x^n)$ are the target posteriors, $= 1$ if target class for frame n is class q_k, else $= 0$. For **linear** LMSE:

$$y_k^n(w) = P(q_k|x^n;\Theta, w) = \sum_{i=1}^{2^d} w_{ik} P(q_k|s_i^n;\Theta_i)$$

In this case the resulting LMSE "normal equations" are linear and can be solved directly, but in this case the weights cannot be constrained to be positive or sum to 1 across all experts. For non linear LMSE, when an MLP is trained using "back error propagation" (a particular case of gradient descent), weights can be constrained to sum to 1, if required, by using the softmax activation function in the output layer.

4.2 Fixed weight estimation using maximum likelihood

The fixed weights $w_i = P(b_i)$, $w_{ij} = P(b_i|c_j)$ or $w_{ijk} = P(b_i|c_j \cap q_k)$ can be estimated using (supervised) relative-frequencies (relative frequency is the ML estimate for a Bernoulli probability) as follows:

$$P(b_i) \cong n_i/n,\ P(b_i|c_j) \cong n_{ij}/n,\ P(b_i|c_j \cap q_k) \cong n_{ijk}/n_k \tag{21}$$

where:

- n is the number of frames of training data.
- n_k is the number of frames for which the target phoneme q_k occurs.
- n_i is the number of frames of training data for which expert i has the largest posterior, across all experts for subsets of s_j, for the target phoneme - and therefore has the smallest KL distance from the target probability distribution.
- n_{ij} is the number of frames of training data for which expert i has the largest posterior, across all experts for subsets of s_j, for the target phoneme.
- n_{ijk} is the number of frames of training data for which expert i has the largest posterior, across all experts for subsets of s_j, when the target phoneme is q_k, .

Note that the number of occurrences of some q_k may be too small for the variance in this estimate to be acceptably small. In this case, variance can be reduced by using the m-estimate, $b_{ijk} = (n_{ijk} + m \cdot p_{0ijk})/(n_k + m)$, where p_{0ijk} is some reasonable prior estimate for this probability, and m is the minimum value of n_{ijk} which gives acceptable confidence in the probability estimate.

4.3 Adaptive weighting using maximum likelihood

It would be possible to estimate the weights $P(b_i|q_k)$ in Section 4.1 using unsupervised ML, with the same EM update equations as are usually used for estimating the Gaussian mixture weights in the context where all of the other Gaussian parameters are fixed, i.e. assign some suitable initial values to these weights, and then iterate as follows:

$$\hat{P}^{new}(b_i|q_k) = \frac{1}{N}\sum_n \hat{P}^{old}(b_i|q_k, x^n) \tag{22}$$

with $$\hat{P}^{old}(b_i|q_k, x^n) = \frac{\hat{P}^{old}(x^n|b_i, q_k)\hat{P}^{old}(b_i|q_k)}{\sum_j \hat{P}^{old}(x^n|b_j, q_k)\hat{P}^{old}(b_j|q_k)} \tag{23}$$

Notice here that $\hat{P}(x^n|b_i, q_k) \neq P(s_i^n|q_k)$. Evaluation of $\hat{P}(x^n|b_i, q_k)$ requires making use of the results in Section 5.2. Initial training of the combination weights in this way could make use of the full training data set, but as this training is unsupervised, these initial weights could be adapted to changing noise levels by combining them in a weighted sum (factor α) with similar weights which are estimated locally, over a window of data samples spanning a short time interval (factor β) of perhaps just a few hundred ms. To date this method has not been tested, but on-line ML based adaptation has been applied elsewhere with some success[1].

4.4 Adaptive weighting using estimated local SNR

We can obtain the adaptive weights, $P(c_i|x)$ in Eq. 5, Eq. 10 and Eq. 13 from direct measures of local data mismatch in each subband, such as SNR, or other measures of likeness to speech data, such as harmonicity [1]. The weight $P(c_i|x)$ is the probability that combination s_i is the largest set of clean subbands.

If we assume that for each subband x_i there is a fixed threshold SNR ε_i below which recognition should improve if subband x_j is ignored[1] (for the current data frame) [6], then $P(reliable(x_j)) = P(SNR_i > \varepsilon_i)$. SNR thresholds ε_i can be estimated by obtaining fullband recognition scores with different noise levels in band x_i, and comparing these with the recognition score when band x_i is ignored.

If we assume that subband combination reliability can reasonably be estimated from the reliability of each of its component subbands, and that subband reliabilities are independent, then, by definition of c_i:

$$c_i \Leftrightarrow (reliable(x_j) \forall x_j \in s_i) \cap (\neg reliable(x_j) \forall x_j \notin s_i)$$

$$\Rightarrow \quad P(c_i|x) = \prod_{x_j \in s_i} P(reliable(x_j)) \prod_{x_j \notin s_i} P(\neg reliable(x_j)) \tag{24}$$

4.5 Adaptive weighting using clean-data likelihood

While the likelihood based decomposition was shown to require access to class posteriors, for the purpose of estimation of the combination weights, $P(c_i|x)$ in Eq. 3 and Eq. 8, it is particularly interesting to have access to data likelihoods. This is because data likelihood carries considerable information about noise level. Intuitively, data which is not from the clean data population pdf should appear highly unlikely. Fig.1 shows that **this is true for spectral data, but not for data in which clean and noisy data has been mixed by orthogonalisation**. However, in both cases it can be seen that SNR level is strongly associated with a characteristic behaviour of data likelihood. If a pdf $p(x|\theta)$ is trained for clean data, and histograms $h(\log p(x))$ are obtained for clean and noisy data log likelihoods for each subband x_i, then reliability weights could be estimated on-line for x_i^n of as:

$$w_i^n = \frac{h_{clean}(\log p(x_i^n))}{h_{clean}(\log p(x_i^n)) + h_{snr0}(\log p(x_i^n))} \tag{25}$$

Example: If the subband histograms for x_i are as in Fig. Eq. 1 b (for MFCC data), and $\log p(x_i^n) = -85$, then the probability that this subband is locally free from data mismatch would be estimated as: $w_i^n = 0.03/(0.01 + 0.03) = 0.75$.

The tendency for orthogonalised data to have higher rather than lower likelihood may not be a bad thing. While there is a large overlap between the histograms in Fig Eq. 1 for clean and noisy orthogonalised data, as lower likelihood implies higher information capacity, information carried should be higher for data with lower likelihood, which is

1. *...though I can't think of any references right now!*

1. *This assumption is not accurate, because as the proportion of noisy subbands increases, the penalty for ignoring further subbands increases. A more accurate model would use a different SNR threshold for each number of other subbands which it has already been decided to ignore, with bands considered in order of increasing SNR.*

more likely to be identified here as clean. This form of weighting also gives a direct measure of data mismatch, while methods based on SNR estimation, for example, do not. If training data was noisy then high SNR does not necessarily imply large data mismatch. Also, SNR based weights for low noise levels in near silent periods will tend indicate very low SNR, while mismatch is small, so data reliability should be high.

To date this direct use of data likelihood in mismatch estimation has not been tested. However, several methods for noise robust ASR have recently been reported which are essentially based on this same idea [5,9].

5. Some implementation details

5.1 Within-stream orthogonalisation

For both practical and theoretical[1] reasons it is highly desirable that the data presented to each combination expert should be othogonalised in some way, rather than having neighbouring coefficients highly correlated, as they are with spectral data. As the motivation behind any “missing feature theory” based model is to separate clean from noisy (mismatching) data, it is not possible to apply an othogonalisation transform (such as the principal components, or discrete cosine transform) to the full data vector, because this would mix all the clean and noisy data together before it could be separated. Instead it is necessary to apply orthogonalisation separately within every subband combination.

This unfortunately means that subband combination likelihoods cannot be obtained by marginalisation of the full data pdfs, so **for both posteriors and likelihood based models, a separate expert must be trained on the locally orthogonalised data for each subband combination**. Furthermore, until a way is found over this problem, no information can be used which is based directly on spectral data. This means that **the spectral-bounds constraint discussed in Section 5.2 cannot be used in practical ASR.** And use of the harmonicity mismatch measure discussed in Section 4.4 would require access to the spectral data before it was orthogonalised.

5.2 Evaluation of incomplete data likelihood for diagonal covariance GMs

From Eq. 9 we have that the estimate for $p(x|c_i \cap q_k)$ which minimises the expected quadratic loss is:

$$\hat{p}(x|c_i \cap q_k) \;=\; p(x_c|q_k) \int_{R_u} p^2(x_u|x_c, \kappa_u, q_k) dx_u \tag{26}$$

In the present context we are mainly interested in Gaussian mixture (GM) models:

$$p(x|q_k) \;=\; \sum_j \alpha_{jk} N(x, \mu_{jk}, C_{jk}) \tag{27}$$

The left hand term $p(x_c|q_k)$ in Eq. 26 is simply the marginal with respect to the subcomponents of x in x_c. This can be evaluated directly from the GM for $p(x|q_k)$, using Eq. 38 with the class pdfs that were trained on clean data.

1. *The practical reason for data orthogonalisation is that full covariance matrices for high dimensional data vectors are very expensive in terms of storage and computation. The theoretical reason is that the large number of free parameters these require makes the amount of training data necessary to achieve satisfactory generalisation prohibitive.*

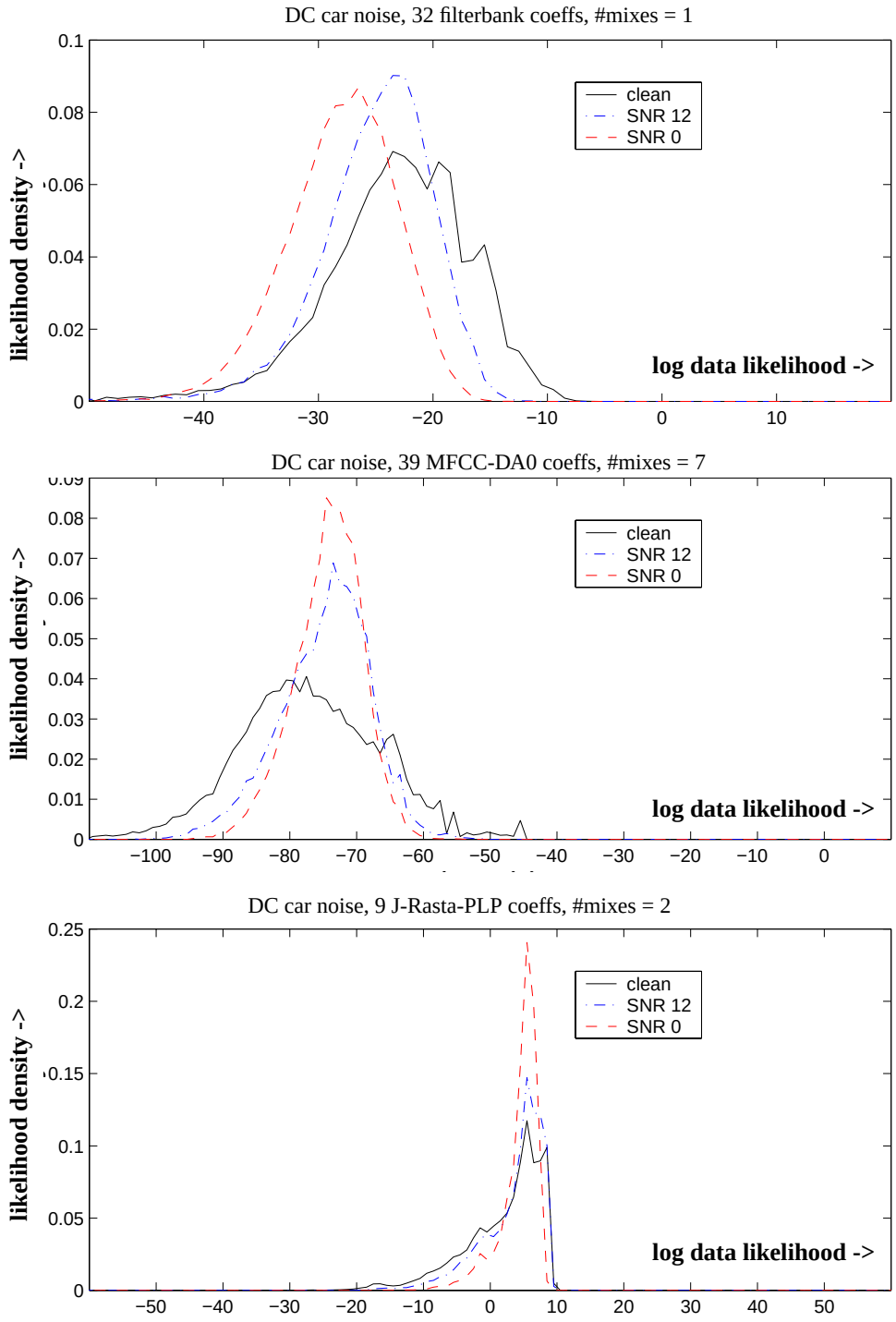


Figure 1: Plots show pdf histograms for [log p(x)] values over test sets with different levels of added car noise, where p(x) is the data pdf trained on the clean training set for the Number95 spoken numbers database. In the top figure (for spectral data) likelihood decreases as noise level increases. However, in the other figures, for orthogonalised data (MFCC & J-Rasta-PLP), likelihood increases with noise level.

The integral in Eq. 26 combines a scaling factor with the bounds constraint. Evaluation of this integral for GMs is always possible in closed form and little computational cost, except that the bounds constraint term (see Section 5.3) cannot be evaluated in closed form with full covariance. Unlike single Gaussians, diagonal covariance GMs can model dependence between data components. In the following working we will assume that covariance is diagonal. Dropping reference to the class index, k, the conditional pdf $p(x_u|x_c, \kappa_u, q)$ can be obtained as:

Knowledge, κ_u, of bounds on the uncertain data gives:

$$p(x_u|x_c, \kappa_u, q) = p(x_u|x_c, q) / \int_{R_u} p(x_u|x_c, q)dx_u = \gamma p(x_u|x_c, q) \text{ for } x_u \in R_u, \text{ else } = 0$$

If no bounds constraint is used, then $\gamma = 1$. Eq. 40 to Eq. 44 give $p(x_u|x_c, q)$ as:

$$p(x_u|x_c, q) = \sum_j \beta_j N(x_u, \mu_u^j, C_u^j), \text{ where}$$
$$\beta_j = \alpha_j N(x_c, \mu_c^j, C_c^j) / \sum_k \alpha_k N(x_c, \mu_c^k, C_u^k)$$

The integral in Eq. 26 can now be evaluated as follows:

$$I = \gamma^2 \int_{R_u} p^2(x_u|x_c, q)dx_u = \gamma^2 \int_{R_u} \left(\sum_j \beta_j N(x_u, \mu_u^j, C_u^j) \right)^2 dx_u \quad (28)$$

$$= \gamma^2 \int_{R_u} \sum_j \sum_k \beta_j \beta_k N(x_u, \mu_u^j, C_u^j) N(x_u, \mu_u^k, C_u^k) dx_u \quad (29)$$

The product of Gaussians is scaled Gaussian. If we write:

$$N(x_u, \mu_u^j, C_u^j) N(x_u, \mu_u^k, C_u^k) = \zeta^{jk} N(x_u, \mu_u^{jk}, C_u^{jk})$$

then $(\zeta^{jk}, \mu_u^{jk}, C_u^{jk})$ can be obtained from $(\mu_u^j, C_u^j, \mu_u^k, C_u^k)$ using Eq. 45 and Eq. 46, so that:

$$I = \gamma^2 \sum_j \sum_k \beta_j \beta_k \zeta^{jk} \int_{R_u} N(x_u, \mu_u^{jk}, C_u^{jk}) dx_u \quad (30)$$

If there is no bounds constraint (or, equivalently, if all constraints on the uncertain data are to be ignored) the integral in Eq. 30 is 1, so that: $I = \gamma^2 \sum_j \sum_k \beta_j \beta_k \zeta^{jk}$. Otherwise, if we wish to evaluate the integral in Eq. 30, and each GM component has diagonal covariance, then we can integrate over the interval $R_u = [0, \overline{x}_u]$ as follows[1,2]:

$$\int_{R_u} N(x_u, \mu_u^{jk}, C_u^{jk}) dx_u = \int_{R_u} \prod_i N(x_{u,i}, \mu_{u,i}^{jk}, (\sigma^2)_{u,i}^{jk}) dx_{u,i} \quad (31)$$

1. *In Eq. 32, Φ denotes the cdf for the standard univariate Gaussian pdf (mean 0 and variance 1). This can be evaluated (indirectly) using the standard C function erf().*
2. *In practice Gaussian probabilities are handled as log probabilities. Some care must be taken in evaluating log cdf values because the standard erf function is highly inaccurate for arguments with absolute value greater than about 6. This problem can easily be solved through use of an asymptotic limit for this cdf.*

$$= \prod_i \int_0^{\widehat{x}_{u,i}} N(x_{u,i}, \mu_{u,i}^{jk}, (\sigma^2)_{u,i}^{jk}) dx_{u,i} = \prod_i \left[\Phi(\frac{t - \mu_{u,i}^{jk}}{\sigma_{u,i}^{jk}}) \right]_{t=0}^{t=\widehat{x}_{u,i}} \quad (32)$$

With $|x_u| = n_u$, Eq. 46 gives:

$$\zeta^{jk} = \prod_i \zeta_i^{jk} = \prod_i N(\mu_{u,i}^j, \mu_{u,i}^k, (\sigma^2)_{u,i}^j + (\sigma^2)_{u,i}^k) \quad (33)$$

$$= \prod_{i=1}^{n_u} \left(\exp\left(-\frac{1}{2} \frac{(\mu_{u,i}^j - \mu_{u,i}^k)^2}{((\sigma^2)_{u,i}^j + (\sigma^2)_{u,i}^k)} \right) / (2\pi((\sigma^2)_{u,i}^j + (\sigma^2)_{u,i}^k))^{0.5} \right)$$

$$= (2\pi)^{-n_u/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n_u} \frac{(\mu_{u,i}^j - \mu_{u,i}^k)^2}{((\sigma^2)_{u,i}^j + (\sigma^2)_{u,i}^k)} \right) / \left(\prod_{i=1}^{n_u} ((\sigma^2)_{u,i}^j + (\sigma^2)_{u,i}^k) \right)^{0.5} \quad (34)$$

Example: In the case where all constraints on the uncertain data are ignored, and $p(x|q_k)$ is one mix spherical covariance Gaussian, with variance σ_k^2, Eq. 34 gives $I = (4\pi\sigma_k^2)^{-n_u/2}$, so that, from Eq. 26:

$$\hat{p}(x|c_i \cap q_k) = \hat{p}(x|x_c, q_k) = (4\pi\sigma_k^2)^{-n_u/2} \cdot p(x_c|q_k) \quad (35)$$

The scale of the optimal incomplete likelihood estimate is therefore **strongly dependent on** $p(x|q_k)$ **and** $|x_u|$. Even if the number of coefficients in each subband combination was the same, the variances in different subbands would differ. This factor is therefore far from being constant across subband combination experts, and **cannot be ignored**.

5.3 Optimal estimate of data posterior when uncertain data is constrained

In the present context we might assume that the main reason for data mismatch is interference from other sound sources. If the data vector consists of spectral parameters taken over a short term window of order 10 ms, then the effect of any interference will always be approximately additive. In this case the unknown clean value is bounded above by the noisy observed value. Furthermore, spectral energies are usually constrained to be positive (even after log compression, which can result in negative values). This means that the clean value is also bounded below by zero. It has been shown that this bounds constraint can considerably improve recognition performance, especially in the common situation that a large proportion of the original data is uncertain [3,4,6].

We can therefore obtain the optimal estimate for $P(q_k|x, \kappa)$ by first establishing what knowledge we have, and then evaluating the resulting conditional expectation. In the present case the knowledge we have[1] is the partition of x into (x_c, x_u), plus the bounds on each component of x_u, $\kappa_u = x_u \in (R_u = [0, \widehat{x}_u])$, where $\widehat{x}_u$ are the observed noisy upper bounds on x_u. We can now evaluate the optimum posteriors estimate given by Eq. 7 as follows. The term $P(\kappa_u|x_c)$ is independent of choice of class q_k, so can be ignored[2] if posteriors are subsequently normalised to sum to 1. The term $P(\kappa_u|x_c, q_k)$ can be evaluated as:

---

1. *It is important to note here that we do **not** know that x is from class q_k. If we did, then we would know that the posterior probability value was exactly 1.*
2. *Normalising to sum to 1 is equivalent to using $P(\kappa_u|x_c) = \sum_k P(q_k) P(\kappa_u|x_c, q_k)$.*

$$P(\kappa_u|x_c, q_k) = P(x_u \in R_u|x_c, q_k) = \int_{R_u} p(x_u|x_c, q_k)dx_u \quad (36)$$

where $p(x_u|x_c, q_k)$ is the conditional pdf for x_u, which can be obtained from the clean data pdf, $p(x|q_k)$. If likelihoods are available then this probability can be easily evaluated for GM pdfs provided that each mix component has diagonal covariance. For full details, see Section 5.2.

6. Conclusion

The rational for subband-combination decomposition, for both posteriors and likelihood based models, has been presented. It has been shown that decomposition for likelihood based models is not as straightforward as with posteriors based models, but is still computationally feasible. In particular, it has been shown that:

- **estimation of incomplete data likelihoods requires evaluation of an important scaling factor for each expert**, which involves an integral of the square of the conditional missing data likelihood (Eq. Eq. 9).
- **to obtain a likelihood decomposition in terms of functions trained on clean data, and with corresponding weights sensitive to noise, it is necessary to convert data likelihoods into posteriors** (Eqs. Eq. 10, Eq. 17). This does not introduce much extra computation, and avoids the need for the scaling factors mentioned above.

The decomposition of both posteriors and likelihood based discriminant functions was derived. In both cases the fullband discriminant function (or expert) was decomposed into a weighted sum of clean-data experts for each subband combination. It was also shown that decomposition with two latent variables can be used to separate terms involving substream data mismatch and substream data relevance to a particular speech unit. MFT [6] was used to show that when uncertain data can be constrained, the estimates for the incomplete data likelihood or posterior could theoretically be improved. However, the need to orthogonalise data within each subband subset (discussed in Section 5.1) means that it is not possible to make direct use of spectral data models in high performance ASR. For this reason:

- **subband combination likelihoods cannot be obtained by marginalisation of the full data likelihood** - as they were in [Eq. 4,Eq. 6]. For both posteriors and likelihood based models, **a separate expert must be trained on the locally orthogonalised data for each subband combination**.
- **the spectral-bounds constraint discussed in Section 5.3 cannot be used in practical ASR** (unless the spectral bounds are transmitted to the orthogonalised data, where they are diffused and unlikely to be effective).

Likelihood based models have a number of advantages over posteriors based models. One is that posteriors can always be obtained from likelihoods via Bayes' rule, but the reverse is not true. Another is that likelihood based ASR systems are more commonly used, so packages for building and tuning them (such as HTK) are more highly developed. With likelihood based systems unsupervised training with hidden states is possible, but with posteriors based models training must normally be supervised, with one model per phoneme. The theory for dealing with missing-features and on-line noise (and speaker) adaptation has also been developed primarily for likelihood based ASR. In Section 4.5 it was also shown that data likelihoods can be used directly for weighting subband-combination experts according to data reliability.

Latent variable decomposition of probabilities requires that the set of events used for decomposition are exhaustive, so that when applied to subband ASR this means that a separate expert must be used not only for each subband, but for every possible subband combination. As the number of subbands used increases, the number of subband

combination experts which it is necessary to train can soon become impractical. In some contexts it may be possible to use a-priori knowledge to set the weights for all but a small number of these combinations to zero. Otherwise an approximation to the 2^d expert functions can be obtained from the d subband experts alone [2].

Appendix A: Notation

$P(x)$	probability of "event x" occurring. The event x must be clearly defined.
$p(x)$	probability density at x of a continuous value $X: P(X \in [x, x + dx]) \cong p(x)dx$
$P(a \cap b)$	probability of events 'a' and 'b' occurring (same as $P(a, b)$)
$P(a \cup b)$	probability of events 'a' or 'b' occurring
q_k	speech unit whose presence at time t is being estimated. For a posteriors based system this will typically be a phoneme, while for a likelihood based system it is usually a hidden state for some phoneme or whole word model.
x	vector for a data window at time t. This will typically consist of a number of appended frames for a posteriors based system, or a single frame with appended smoothed first and second time differences for a likelihood based system.
x clean	x is from the same data population that was used in model training (i.e. x has no data mismatch)
d	number if spectral subbands
x_i	ith subband of x, for $i = 1 \ldots d$
(s_i, s_i')	partition of x into ith subband combination, s_i, and its complement, s_i', for $i = 1 \ldots 2^d$
(x_c, x_u)	partition of x into subbands which are clean (or "certain"), x_c, and unclean, x_u (or "uncertain")
$x_{c,i}$	ith component of x_c
c_i	event that subband combination s_i is the largest clean subset of x
b_i	event that s_i is the best subset of x for estimating speech unit posteriors
b_{ik}	event that s_i is the best subset of x for estimating speech unit posteriors when true unit is q_k
$\Phi(x)$	standard Gaussian cdf
$\phi(x)$	standard Gaussian pdf
$N(x, \mu, C)$	Gaussian pdf for x with mean vector μ and covariance matrix C
μ_u	subvector of uncertain components of μ corresponding to uncertain components of x
C_u	submatrix of uncertain components of C corresponding to uncertain components of x
$\mu^j_{u,i}$	ith component of μ_u for jth Gaussian mixture component
$(\sigma^2)^j_{u,i}$	ith variance in diagonal covariance matrix C_u for jth Gaussian mixture component
w	vector of sub-band expert weights

Appendix B: Properties of the Gaussian and Gaussian mixture pdf

Any joint pdf can be factored into its marginal and conditional pdfs: $p(a, b) = p(a)p(b|a)$. If certain and uncertain data components are collected together, then $p(x) = p(x_c, x_u) = p(x_c)p(x_c|x_u)$.

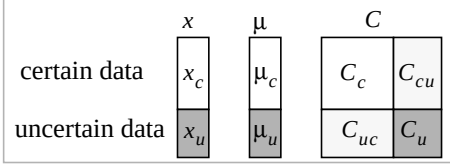


Figure 2: definition of various certain / uncertain sub-vectors and sub-matrices used in formulas for marginal and conditional pdfs.

Of multivariate pdfs, the Gaussian has a particularly convenient marginal and conditional form.

$$p(x|s) = N(x, \mu, C) = \exp(-0.5(x-\mu)'C^{-1}(x-\mu))/((2\pi)^{n_x}|C|)^{1/2} \quad (37)$$

$$p(x_c|s) = \int p(x|s)dx_m = \int N(x, \mu, C)dx_m = N_c(x_c) = N(x_c, \mu_c, C_c) \quad (38)$$

$$p(x_u|x_c) = N(x_u, \mu_{u|c}, C_{u|c}) = N_{u|c}(x_u) \quad (39)$$

$$\text{where } \mu_{u|c} = \mu_u + C_{uc}C_c^{-1}(x_c - \mu_c) \text{ and } C_{u|c} = C_u - C_{uc}C_c^{-1}C_{cu} \quad (40)$$

The marginal and conditional pdfs for the Gaussian mixture pdf are also of a convenient form:

$$p(x|s) = \sum_{i=1}^{m} a_i N_i(x) \quad (41)$$

$$p(x_c|s) = \sum_i a_i N(x_c, \mu_{ci}, C_{ci}) = \sum_i a_i N_{ci}(x_c) \quad (42)$$

$$p(x_u|x_c) = \sum_i b_i N(x_u, \mu_{u|ci}, C_{u|ci}) = \sum_i b_i N_{u|ci}(x_u) \quad (43)$$

$$\text{where } b_i = a_i N_{ci}(x_c) / \sum_i a_i N_{ci}(x_c) \quad (44)$$

The product of Gaussians is Gaussian: $N(x, \mu_1, C_1)N(x, \mu_2, C_2) = aN(x, \mu, C)$, where

$$C = (C_1^{-1} + C_2^{-1})^{-1}, \mu = C(C_1^{-1}\mu_1 + C_2^{-1}\mu_2), \text{ and} \quad (45)$$

$$a = \exp(-0.5[\mu_1'C_1^{-1}\mu_1 + \mu_2'C_2^{-1}\mu_2 - \mu'C^{-1}\mu])/((2\pi)^{n_x/2}|C_1 + C_2|^{1/2}) \quad (46)$$

When Gaussians are univariate, C, μ and a simplify as follows:

$$\sigma^2 = \sigma_1^2\sigma_2^2/(\sigma_1^2 + \sigma_2^2), \mu = (\mu_1\sigma_2^2 + \mu_2\sigma_1^2)/(\sigma_1^2 + \sigma_2^2), \text{ and} \quad (47)$$

$$a = \exp(-0.5[(\mu_1 - \mu_2)^2/(\sigma_1^2 + \sigma_2^2)])/(2\pi)^{1_x/2}(\sigma_1^2 + \sigma_2^2)^{1/2} = N(\mu_1, \mu_2, \sigma_1^2 + \sigma_2^2) \quad (48)$$

Appendix C: Accurate evaluation of the Gaussian cdf

The standard Gaussian cdf (cumulative distribution function):

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)dt$$

can be evaluated using the C standard erf function (whose definition differs from the common definition):

$$erf(x) = \int_{0}^{x} \frac{2}{\sqrt{\pi}} \exp(-t^2)dt$$

However, this accuracy of the C standard erf implementation (a polynomial approximation) falls off very rapidly as the absolute value of x increases beyond about 5. This instability can be avoided by making use of a result known as *Mill's ratio* concerning the asymptotic behaviour of the Gaussian cdf as x tends to infinity:

$$\lim_{x \to \infty} (1 - \Phi(x)) = \frac{\phi(x)}{x} \tag{49}$$

C code using this limit is given below:

```
#include <math.h>

#define MAXSDS 5
#define sqrtTPI 2.5066283
#define sqrt2 1.414235719
#define standard_gauss_pdf(x) (exp(-x*x/2e0)/sqrtTPI)

double erf(double x);

double standard_gauss_cdf(double x) {
     double   result;
     if (x < 0e0) result = (1e0 - standard_gauss_cdf(-x));
     else if (x < MAXSDS) result = (1e0 + erf(x/sqrt2))/2e0;
     else result = 1e0 - standard_gauss_pdf(x)/x;
     return result;
}
```

References

[1] Berthommier, F. & Glotin, H. (1999) “A new SNR-feature mapping for robust multistream speech recognition”, Proc. ICPhS’99.

[2] Hagen, A., Morris, A.C. & Bourlard, H. (1998) “Sub-band based speech recognition in noisy conditions: The Full-Combination approach”, Research Report IDIAP-RR 98-15.

[3] Kermorvant, C. & Morris, A. (1999) “A comparison of two strategies for ASR in additive noise: Missing Data and Spectral Subtraction”, Proc. Eurospeech’99, pp.2891-2844.

[4] 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.

[5] Ming, J. & Smith, F. J. (1999) “Union: a new approach for combining sub-band observations for noisy speech recognition”, Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp. 175-178.

[6] 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.

[7] Morris, A.C., Hagen, A. & Bourlard, H. (1999) “The full-combination subbands approach to noise robust HMM/ANN based ASR”, Proc. Eurospeech’99, pp. 599-602.

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

[9] de Veth, J., de Wet, F., Cranen, B. & Boves, L. (1999) “Missing feature theory in ASR: make sure you missing the right type of features”, Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp.231-234.