MAP Combination of Multi-Stream HMM or HMM/ANN Experts

*Andrew Morris, Astrid Hagen, Hervé Bourlard**

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
P.O. Box 592, 4 Rue du Simplon, CH-1920, Martigny, Switzerland

`morris,hagen,bourlard@idiap.ch`
`http://www.idiap.ch/`

Abstract

Automatic speech recognition (ASR) performance falls dramatically with the level of mismatch between training and test data. The human ability to recognise speech when a large proportion of frequencies are dominated by noise has inspired the "missing data" and "multi-band" approaches to noise robust ASR. "Missing data" ASR identifies low SNR spectral data in each data frame and then ignores it. Multi-band ASR trains a separate model for each position of missing data, estimates a reliability weight for each model, then combines model outputs in a weighted sum. A problem with both approaches is that local data reliability estimation is inherently inaccurate and also assumes that all of the training data was clean. In this article we present a model in which adaptive multi-band expert weighting is incorporated naturally into the maximum a posteriori (MAP) decoding process.

1. Introduction

One of the main factors limiting the take up of ASR in practical applications is the rapid degradation in recognition performance which occurs with mismatch between training and test data. Inspired by the robustness of human recognition to band-limited noise [1,10,12], various ASR models have been proposed in which performance is greatly improved by identifying and then treating as "missing data" parts of the spectral signal which are dominated by noise [4,10,14].

In Section 2 we briefly introduce the "missing data" and "multi-band" approaches which cover the necessary background for the present model. In Section 3 we present the proposed new model for MAP combination of sub-band experts. This model is tested in Section 4 and discussed in Section 5.

2. Missing data and multi-band ASR

Let $X = (x_1, \ldots, x_T)$ denote the acoustic features for an utterance to be recognised, and $X = (X_p, X_m)$ denote the partition of X into clean and noisy data (also referred to as present/missing data).

2.1 Missing data ASR

In the "missing data" approach [5,10,14], a model

$$\hat{P}(Q|X, Xisclean) = P(Q|X, \Theta) \tag{1}$$

is trained on clean data to give model parameters Θ. Noise dominated data X_m is detected by some kind of local SNR estimation technique [2,3], followed by MAP decoding for the given position of missing data.

$$\hat{Q} = \arg max_Q P(Q|X_p, \Theta) \tag{2}$$

$$= \arg max_Q P(Q|\Theta)P(X_p|Q, \Theta)/P(X_p) \tag{3}$$

$$= \arg max_Q P(Q|\Theta)P(X_p|Q, \Theta)\text{, as } X_p \perp Q \tag{4}$$

2.2 Multi-band ASR

In the first multi-band HMM/ANN hybrid models [4,9] one MLP expert was trained for each data sub-band. Outputs from each expert were combined as a weighted sum and passed on for Viterbi decoding as scaled likelihoods. However, independent processing of sub-bands can result in loss of joint information and reduced performance with clean speech. The "full combination" (FC) multi-band model avoids this problem.

2.3 Full combination multi-band ASR

Let the $M = 2^d$ different combinations of $0 \ldots d$ sub-bands from a set of d sub-bands if data vector x be denoted $x^{(i)}$, where $i = 1 \ldots M$. Let $b^{(i)}$ denote that $x^{(i)}$ is clean and its complement is missing, giving

$$P(q_k|b^{(i)}, x) = P(q_k|x^{(i)}, \Theta) \tag{5}$$

Under the "maximum assumption" that all data values represent either 100% clean speech or 100% noise[1], the events b_i are exhaustive and mutually exclusive. In this case the full-band phoneme posterior probability for each class (phoneme or hidden state) q_k can be decomposed into a weighted sum of M sub-band combination posteriors [7]:

$$\hat{P}_w(q_k|x) = \sum_{i=1}^{M} P(q_k, b^{(i)}|x) \tag{6}$$

$$= \sum_i P(b_i|x)P(q_k|b^{(i)}, x) \tag{7}$$

$$= \sum_i w_i P(q_k|x^{(i)}, \Theta) \tag{8}$$

* Also professor at Swiss Federal Institute of Technology (EPFL), Lausanne

1. For logarithmically compressed speech and noise spectral energy values a and b, if $a > b$ then $\log(a+b) <$ $\log(a) + 1$, so if $a \gg 1$, $\log(a+b) \cong \log(a)$.

Probabilities $P(q_k|x^{(i)})$ for each state q_k are estimated by an MLP expert trained for each sub-band combination.

2.4 Outstanding problems

Missing-data and multi-band models have shown steadily improving results, but they have a number of limitations:

1. both local SNR estimation and combination expert reliability weighting are inherently inaccurate.
2. data is often corrupted enough to reduce recognition performance, while retaining considerable speech information, which should not simply be discarded.
3. in the missing-data approach, the need to avoid mixing clean and noisy data precludes data orthogonalisation, resulting in low performance in clean speech.
4. none of the methods tested for estimating expert weights [3,8] have shown any strong advantage over using equal weights, except in narrow band noise.

Some method is required to "hide" reliability estimation.

3. MAP combination

It is not possible to use the maximum likelihood (ML) objective for parameter adaptation with missing data unless a model is available for noise dominated data. Multicondition training has given good results on the Aurora 2.0 task, but this depends on the range of noise conditions being severely limited. The model we present here trains with clean data only, and uses the MAP objective for expert combination weight adaptation. Unlike ML, the MAP objective is discriminatory, so any increase in non speech like data variation will decrease the MAP objective.

This model presented is a reformulation of Eq.7 to cover the entire utterance and not just a single time frame[1].

3.1 MAP combination leads to 0/1 weights

Let $B^{(i)}$ denote the event that $X^{(i)}$ is present and its complement is missing, i.e. the same components of x_t are missing for all t. Now substitute B for b, X for x and Q for q in Eq.7 to obtain

$$\hat{P}_w(Q|X) = \sum_i w_i P(Q|X^{(i)}, \Theta) \tag{9}$$

Eq.9 has the form $A = \sum w_i a_i$, and during Viterbi MAP decoding a_i are given values. This means that the weight corresponding to $max_i a_i$ must be one, and all other weights are zero (see Appendix A). Therefore

$$max_w \hat{P}_w(Q|X) = max_i P(Q|X^{(i)}, \Theta) \tag{10}$$

The events B_i are exclusive, but they are certainly not exhaustive, because varying speech and noise energy guarantee that the components of x_t which are dominated by noise or "missing" are *not* the same for all t. However, it follows from the Principal of Optimality that the optimal weight sequence over time, for a given Q, will always consist only of 0/1 weight values.

3.2 MAP evaluation for HMMs and HMM/ANNs

An added advantage of MAP combination is that it can be applied simply to both HMM and HMM/ANN models. If we make the same Markovian independence assumptions that are used with HMMs

$$P(Q) \cong P(q_1)\prod_{t=2} P(q_t|q_{t-1}) \tag{11}$$

$$p(X^{(i)}|Q) \cong \prod_t p(x_t^{(i)}|q_t) \tag{12}$$

and the further (more contentious) assumption

$$p(X^{(i)}) \cong \prod_t p(x_t^{(i)}) \tag{13}$$

then we can directly express Eq.9 in terms of the quantities $p(x|q_k)$ and $P(q_k|x)$, as modelled by HMMs or HMM/ANNs respectively.

$$\hat{P}_w(Q|X) = \sum_i w_i P(X^{(i)}|Q)P(Q)/P(X^{(i)}) \tag{14}$$

$$\cong P(Q)\sum_i w_i \prod_t p(x_t^{(i)}|q_t)/p(x_t^{(i)}) \tag{15}$$

$$= P(Q)\sum_i w_i \prod_t p(q_t|x_t^{(i)})/p(q_t) \tag{16}$$

For MLP based HMM/ANNs it is necessary to train a separate MLP expert for each sub-band combination. While the number of possible sub-band (or sub-stream) combinations can be very high, it is not always necessary to train for all possible combinations. For sub-band combination with d bands, combinations containing << d bands may be omitted. For combination of streams of features from different time scales, or different data modalities, it is not necessary to train for combinations in which different streams is independent.

For HMMs it is necessary to train separate experts for different combinations only if it is required to orthogonalise within each combination. In the case where the data in each combination consists only of features concatenated from different sub-streams, it is only necessary to train a single expert for the full-band combination. For the Gaussian mixture models normally used in CDHMMs, the marginal density for each mix component $p(x_t^{(i)}|m_j, q_t)$ can be evaluated directly from full-band densities $p(x_t|m_j, q_t)$.

3.3 Decoder implementation

Model 1 (MAPMB1): If we assume that Eq.10 is true,

$$Q_{MAP} = \arg max_Q max_i P(Q|X^{(i)}, \Theta) \tag{17}$$

This solution can be obtained using a normal Viterbi decoder, by noting the MAP solution from each sub-band combination expert,

$$Q^{(i)} = \arg max_Q P(Q|X^{(i)}, \Theta) \tag{18}$$

together with its associated MAP probability,

1. This model represents a consistent reformulation of the model presented in [6].

$$\varphi_i = max_Q P(Q|X^{(i)}, \Theta) \qquad (19)$$

then selecting

$$Q_{MAP} = Q^{(\arg max_i \varphi_i)} \qquad (20)$$

Model 2 (MAPMB2): If we replace $B^{(i)}$ by $b^{(i)}$ and write

$$\beta_t = (b_t^{(1)}, \ldots, b_t^{(M)}) \text{ and } \beta = (\beta_1, \ldots, \beta_T)$$

then to obtain

$$Q_{MAP} = \arg max_{Q,\beta} P(Q, \beta | X^{(i)}, \Theta) \qquad (21)$$

it is necessary to use 2D Viterbi decoding.

4. ASR tests

We have so far tested MAP combination only under the assumption that the same components of x_t are missing for all t (Model 1, Section 3.3).

4.1 Data preparation

Tests were based on the Aurora 2.0 connected digits database [15]. HMM models were trained on the full clean training set. While the standard Aurora models use 13 MFCC features with three mixture components, in these initial tests we used unorthogonalised 32 channel auditory model filterbank data at 10 ms centres (so that results would be comparable with recent missing-data ASR tests [2,13]). For this data seven mixture components were needed to better model data covariance.

4.2 Recognition tests

Recognition tests were made to compare baseline HMM against an HMM using MAPMB multi-band, Model 1 (Section 3.3, Eqs.15 & 17-20). Both systems used the same HMMs trained on clean fullband data, and the same input data (32 channel fbank, with first differences). For MAPMB the data was divided into just two sub-bands, one for static and one for difference features (division into 4 sub-bands was also tested, but computation increased and results did not improve).

Test data was a 200 example cross section selected from each of the 1001 example noise conditions for test set (a) (subway, babble, car and exhibition noise, at SNR 0, 10 and 20 dB, and clean). Results are summarised in Figures 1a,b,c. Figure 1c also compares results with those recently obtained with an advanced missing-data model [2,13].

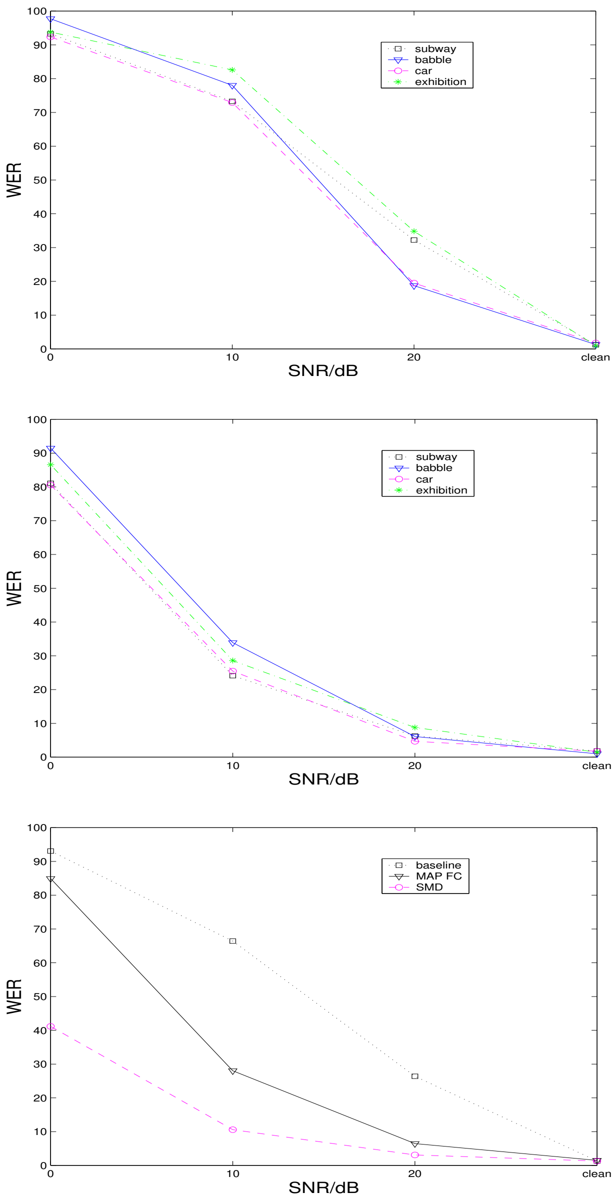

Figure 1.

Figure 1a. (top) shows WER scores for the HMM baseline, for the four noise types in Aurora 2.0 test set (a), against SNR.

Figure 1b. (middle) shows scores for HMM based MAPMB1 model (MAP fixed combination over time).

Figure 1c. (bottom) shows scores averaged over all four noise conditions, for the baseline HMM, MAP full-combination, and for the "soft missing data" model.(SNR based missing data).

5. Discussion

Results reported here are for Model 1 only (missing components of x_t assumed same for all t). Performance increases are clearly to be expected from Model 2, in which missing sub-bands can vary in time. Further improvements should also result from training a separate set of HMMs with orthogonalised features from each sub-band combination.

5.1 Improved duration modelling

We have observed that adding noise to speech data often results in previously distinct sounds coming to resemble a subset of clean but noise-like sounds, rather than in easily detectable outlier data. The importance of duration modelling therefore increases with noise level. Recognition output such as the following occurs frequently (exhibition noise, SNR 0dB).

```
True:MDJ_2252590="2 2 5 2 5 9 zero", 2.79s
Guess="nine"

True:MFP_2868="2 8 6 8", 1.58s
Guess="oh"
```

That an utterance 2.8s long can be recognised as a single word shows that the Markovian state sequence model, given by Eq.11, is inadequate and some form of improved duration modelling is required, especially for recognition in noise.

5.2 Second order Markov models

As spectral data usually changes continuously in time, neighbouring feature vectors are clearly correlated and the assumption in Eq.13 is highly inaccurate. For HMMs this could be corrected by a second order Markov model

$p(x_t|q_k)$ in Eq.15 becomes $p(x_{t-1}, x_t|q_k)$

For HMM/ANNs (Eq.16) this should not be a problem, because the input vector spans several data frames.

6. Conclusion

We have shown how the discriminative MAP objective can be applied in a computationally feasible way to select optimal combination weights for full-combination multi-stream ASR. Experimentation is still at an early stage and the non standard set up tested here does not permit direct comparison with standard test results. However, the results reported have served as a proof of concept for MAP combination. It is now worth proceeding with some of the ideas discussed above, including 2D Viterbi decoding and improved duration modelling.

Acknowledgements

This work was carried out in the framework of both the EC/OFES SPHEAR (Speech, Hearing and Recognition) and RESPITE (REcognition of Speech by Partial Information TEchniques) projects.

Appendix A: Optimum weights select max posterior

Eq.14 has the form

$$A = \sum_i w_i P(Q|X^{(i)}) = \sum_i w_i a_i \qquad (22)$$

where a_i are fixed values. We can find w to maximise this, subject to the constraints $\sum_i w_i = 1$ and $w_i \geq 0$, as follows. First, without loss of generality, label a_i (which are all positive) in order of decreasing magnitude.

$$A = w_1 a_{max} + w_2 a_2 + \ldots + (1 - w_1 - \ldots) a_{min} \qquad (23)$$

Differentiating with respect to each free parameter w_j $j = 1\ldots(M-1)$, gives

$$\frac{dA}{dw_j} = a_j - a_{min} \qquad (24)$$

But $a_j - a_{min} \geq 0$, so A is always increasing, and increases fastest with increase in w_1. From this it follows that A is maximised when $w_1 = 1$ and all other $w_i = 0$. Therefore

$$max_w A = max_w \sum_i w_i a_i \qquad (25)$$

$$= a_{max} = max_i P(Q|X^{(i)}) \qquad (26)$$

References

[1] 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.

[2] 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.

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

[4] Bourlard, H., Dupont, S. & Ris, C. (1996) “Multi-stream speech recognition”, Research Report IDIAP-RR-96-07.

[5] 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.

[6] Hagen, A., Bourlard, H. & Morris, A.C (2001) “Adaptive ML weighting in multi-band recombination of Gaussian mixture ASR”, Proc. ICASSP 2001.

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

[8] Hagen, A., Morris, A.C. & Bourlard, H. (1999) “Different weighting schemes in the full combination sub-bands approach for noise robust ASR”, Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions.

[9] Hermansky, H., Tibrewela, S. & Pavel, M. (1996) “Towards ASR on partially corrupted speech”, Proc ICSLP’96, pp. 462-465.

[10] 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

[11] Ming, J. & Smith, F.J. (2000) “A probabilistic union model for sub-band noisy speech recognition”, Proc. ICASSP 2000, pp.1787-1790.

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

[13] Morris, A.C., Barker, J. & Bourlard, H. (2001) “From missing data to maybe useful data: soft data modelling for noise robust ASR”, Proc. WISP 2001, workshop on Innovative methods in Speech Processing, pp.153-164.

[14] 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.

[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] 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.