LOW COST DURATION MODELLING FOR NOISE ROBUST SPEECH RECOGNITION

Andrew C. Morris[1], Simon Payne[1,2], Hervé Bourlard[1,3]

[1]Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
[2]Sheffield University, Dept. of Computer Science, UK
[3]Swiss Federal Institute of Technology (EPFL), Lausanne, CH

{morris,payne,bourlard}@idiap.ch

ABSTRACT

State transition matrices as used in standard HMM decoders have two widely perceived limitations. One is that the implicit Geometric state duration distributions which they model do not accurately reflect true duration distributions. The other is that they impose no hard limit on maximum duration with the result that state transition probabilities often have little influence when combined with acoustic probabilities, which are of a different order of magnitude. Explicit duration models were developed in the past to address the first problem. These were not widely taken up because their performance advantage in clean speech recognition was often not sufficiently great to offset the extra complexity which they introduced. However, duration models have much greater potential when applied to noisy speech recognition. In this paper we present a simple and generic form of explicit duration model and show that this leads to strong performance improvements when applied to connected digit recognition in noise.

Keywords. duration models, noise robust ASR, HMMs

1. INTRODUCTION

There are two widely perceived limitations of the implicit duration modelling inherent in standard HMM decoding.

1. The Geometric state duration distributions implicitly modelled in standard HMM decoding do not accurately reflect true state duration distributions.
2. This model imposes no hard limit on maximum duration, which limits the impact transition probabilities can have in combination with acoustic probabilities, which have a different order of magnitude.

This second limitation really has no theoretical basis, because the acoustic probabilities in Eq.(2) are implicitly divided by $P(X)$ and are therefore not dependent on the acoustic vector dimension. The literature on explicit duration modelling already shows that duration constraints can sometimes improve recognition performance [1, 2, 5, 11, 12]. The problem is that for recognition of clean speech this improvement is not always very great compared with the cost of implementation of what have previously been somewhat ad-hoc and/or computationally expensive duration models, and if care is not taken performance can even be reduced.

However, most applications require recognition in noise, and in this case the benefits of duration modelling can be much more pronounced. In this article we present a simple and theoretically consistent form of duration model implementation which considerably improves recognition performance in a wide range of noise conditions. Time normalised duration probabilities could also provide a useful measure of recognition confidence in other applications such as two-pass processing and keyword spotting.

2. DECODING WITH EXPLICIT DURATION MODELS

Let word, state and observation sequences be denoted $W = (w_1, \ldots, w_M)$, $Q = (q_1, \ldots, q_T)$ and $X = (x_1, \ldots, x_T)$. The equations which provide the basis for Viterbi MAP (Bayes optimal) decoding for whole word model HMMs are as follows.

$$W^\circ = \arg max_{W,Q} P(W, Q|X) \tag{1}$$

$$= \arg max_{W,Q} P(W)P(Q|W)p(X|Q, W) \tag{2}$$

2.1 Implicit state duration models

Q is equivalent to (Q, W) when the states for each word are modelled separately. Let q_t and q_k denote "the state at time t" and "the k^{th} of K state models" respectively. On applying the Markovian independence assumptions we get:

$$P(W) = P(w_1)\prod_{i=2}^{M} P(w_i|w_{i-1}) \tag{3}$$

$$P(Q|W) = P(Q) = P(q_1)\prod_{t=2}^{T} P(q_t|q_{t-1}) \tag{4}$$

$$p(X|Q, W) = p(X|Q) = \prod_{t=1}^{T} p(x_t|q_t) \tag{5}$$

Model parameters for standard HMM based ASR comprise parameters for the state pdfs $p(x_t|q_t)$; word and state priors $P(w_i)$, $P(q_k)$; word pair probabilities $P(w_i|w_j)$, and state transition probabilities $P(q_t|q_{t-1})$.

In applying the Markovian assumption to obtain Eq.(4)

$$P(q_t|q_1, q_2, \ldots q_{t-1}) \approx P(q_t|q_{t-1}) \tag{6}$$

it is assumed that state transition probabilities depend only on the previous state q_{t-1}, and not on the time d_{t-1} spent *so far* in that state. In this case the probability θ of a state change from a given state at any time step is constant, so that the *final* duration fd_k for each state q_k follows a Geometric distribution, with

$$P(fd_k \geq d) = (1-\theta)^{d-1} \tag{7}$$

$$P(fd_k = d) = \theta(1-\theta)^{d-1} \tag{8}$$

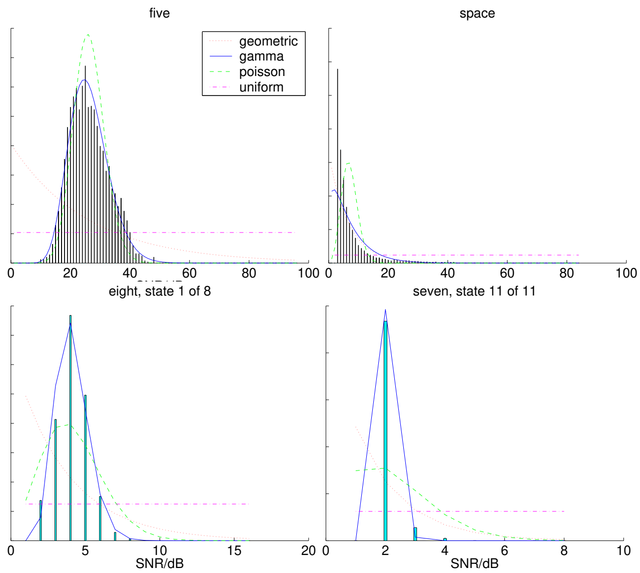

Fig. 1 Fitting duration model pdfs.
Duration histograms for a multi-state word (top left), and one state inter-word silence (top right), a typical within-word state (bottom left), and an end-word state (bottom right). Each case also shows fitted Geometric, Gamma, Poisson and Uniform pdfs.

In this model duration probability always decreases with d. A more accurate model would describe duration probability first increasing with d up to the average state duration, and only then decreasing.

2.2 Explicit state duration models

The state sequence Q implicitly specifies a duration for each state. In Eq.(6) this information is simply ignored. Explicit duration models replace Eq.(6) by

$$P(q_t|q_1, q_2, \ldots q_{t-1}) \approx P(q_t|q_{t-1}, d_{t-1}) \tag{9}$$

A histogram of duration counts $h_k(d)$ is first obtained for each state q_k from a suitable speech database together with its state level segmentation. Duration distributions $P_k(fd = d)$ are then obtained by first fitting a suitable parametric pdf (e.g. Poisson or Gamma) to the histogram data (see Section 3), truncating it, and then weighting it with the normalised histogram.

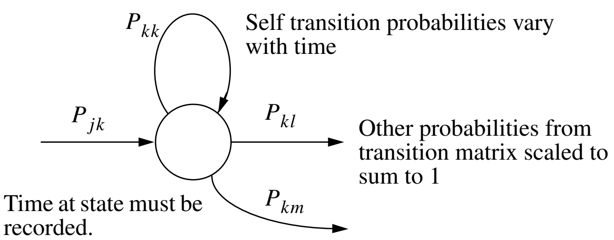

Fig.2 Explicit duration model
Word pair probabilities and state transition matrices are unchanged, but transition probabilities now depend on state duration.

If histogram probabilities were entirely replaced by their parametric fit, one could save storage by storing only the pdf parameters. However, except for pdfs where Eq.(17) has closed form (for the Geometric pdf these are the constant transition probabilities) it is better to store (probability final duration greater than or equal to) values, $Pge(d)$, for each d.

$$Pge_k(d) \equiv P(fd_k \geq d) = \sum_{i \geq d} P(fd_k = d) \tag{10}$$

2.2.1 *State sequence probability evaluation*

Probabilities $P(fd_k = d)$ are not needed in practice, but could be recovered using

$$P(fd_k = d) = Pge_k(d) - Pge_k(d+1) \tag{11}$$

Duration dependent self transition probabilities

$$P_{kk} = P(q_t = q_k|q_{t-1} = q_k, fd_{t-1} \geq d) \tag{12}$$

$$= P(fd_t > d|fd_t \geq d) \tag{13}$$

can be obtained from $Pge(d)$ values as follows:

$$P(fd > d) = P(fd > d \wedge fd \geq d) \tag{14}$$

$$= P(fd > d| fd \geq d)P(fd \geq d) \tag{15}$$

$$\Rightarrow P(fd > d| fd \geq d) = \frac{P(fd > d)}{P(fd \geq d)} = \frac{Pge(d+1)}{Pge(d)} \tag{16}$$

$$\text{i.e. } P_{kk} = \frac{Pge_k(d+1)}{Pge_k(d)} \tag{17}$$

Non self transition probabilities

$$P_{jk} = P(q_t = q_k|q_{t-1} = q_{j \neq k}, fd_{t-1} \geq d) \tag{18}$$

can then be obtained by normalising the usual transition probabilities $P(q_k|q_j)$ from any state to sum to one:

$$P_{jk} = \alpha P(q_k|q_j) \text{, such that} \tag{19}$$

$$P_{kk} + \alpha \sum_{j \neq k} P(q_k|q_j) = 1 \text{, but} \tag{20}$$

$$\sum_{j \neq k} P_{jk} = 1 - P(q_k|q_k) \text{, so } \alpha = \frac{(1 - P_{kk})}{(1 - P(q_k|q_k))} \tag{21}$$

$$\text{i.e. } P_{jk} = \frac{(1 - P_{kk})P(q_k|q_j)}{(1 - P(q_k|q_k))} \tag{22}$$

2.2.2 *Simplicity of implementation*

An important point about this form of explicit duration model is its simplicity. Once these Pge probabilities are available for each state, there is no change required at all to the standard decoder which uses implicit duration models, except to propagate the current state duration d along each Viterbi path, and to provide a routine which provides duration dependent transition probabilities given d and the static transition probabilities using Eqs (17, 22).

2.2.3 *Word duration models*

One can obtain word duration models in exactly the same way as for state duration models, and propagate the current word duration as well as, or instead of, the state duration. Word durations are much longer than state durations and are more bell shaped, so that candidate word duration pdfs would include the Gaussian.

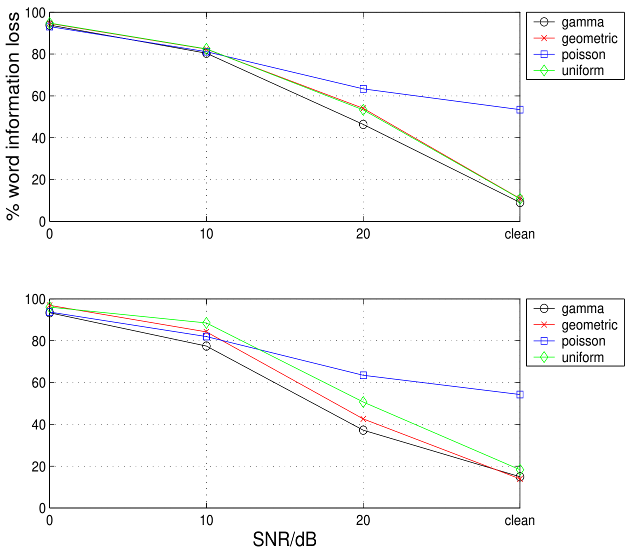

Fig.3 Performance of different parametric pdfs

Top Fig. compares WER for Gamma, Geometric, Poisson and Uniform pdfs with subway noise when duration model scaling factor is optimised for clean data. Bottom Fig. shows same when scaling factor is optimised for SNR 20dB.*

However, it would not be correct to multiply together word and state sequence duration probabilities, because they are two ways of modelling the same sequence. At the end of each word one could replace the contribution of the state duration probability for this word by the word duration probability -but we tried this for the tests described in Section 4 and performance was decreased.

3. DURATION MODEL ESTIMATION

Having obtained duration histograms $h[i]$ for each word and/or state, these values must be smoothed before they can be used reliably as probabilities. One could convolve the histogram with some arbitrary smoothing window, but it is more common to make use of fitted parametric pdfs. To fit a parametric pdf we first estimate its mean μ and variance σ^2 as the sample mean $\bar{x}$ and sample variance s^2.

$$N = \sum_i i \,,\ \bar{x} = \sum_i i h[i]/N \,,\ s^2 = \sum_i i^2 h[i]/N - \bar{x}^2 \quad (23)$$

We have tested the Gamma, Poisson, Geometric and Uniform distributions [4]. The Gamma pdf has two parameters, $\alpha = \mu^2/\sigma^2$ and $\lambda = \mu/\sigma^2$. This allows independent control of the pdf shape and scale.

Gamma: $f_{\alpha,\lambda}(x) = x^{(\alpha-1)} e^{-\lambda x}/\lambda^{-\alpha}\Gamma(\alpha)$

Poisson: one parameter, $\alpha = \mu$: $f_{\alpha}(x) = e^{-\alpha}\alpha^{x}/x!$

Geometric: one parameter, $\alpha = 1/\mu$: $f_{\alpha}(x) = \alpha(1-\alpha)^{(x-1)}$

In every case, the pdf is forced to zero for durations less than the minimum number of steps through each state model, or greater than F times the observed range of durations (where F is a given tuning factor) then normalised. The histogram is then smoothed by first normalising it to have area 1, $h[i] \leftarrow h[i]/N$, then forming a weighted average with the fitted pdf

$$P(fd = d) = Pe[d] \leftarrow \theta \cdot h[d] + (1-\theta) \cdot f(d) \quad (24)$$

Stepping F in steps of 0.5 and θ in steps of 0.1, tests in Section 4 gave best results with $F = 2$ and $\theta = 0$. Figure 3 compares the performance of these four pdfs with increasing levels of subway noise (speech data as in Section 4). The Gamma distribution gives the best results at all intermediate (i.e. realistic) noise levels, and so was used throughout for subsequent tests. From figure 1 it is not surprising that the Gamma pdf performs best.

4. EXPERIMENTS

4.1 Recognition system

Duration models were installed in the decoder for an HMM/ANN recognition system in which scaled state likelihoods $P(q_k|x)/P(q_k)$ are estimated by an MLP [3]. The MLP used here had 64 input units, 400 hidden units, and 118 output units. The decoder used is the same as for an HMM/GMM, except that the transition probabilities are not trained but fixed. HMMs were whole word models with straight-through topology, with the same transition probability used throughout. This probability was tuned for clean speech to the value 0.6.

4.2 Test database

The test database was Aurora 2.0 [10] which is based on TIDigits (speaker independent connected digits, 0..9, oh, silence), downsampled to 8 kHz. Training used the full 8440 clean training set. Tests used the four noise types in test set (a) (each test comprising 1001 utterances at SNR 0, 10, 20dB, and clean. Features were 32 ms 32 channel mel scaled log filter-bank coefficients, together with first time differences, at 10 ms centres (the missing-data recognition approach, with which the present duration models are intended to cooperate, can only use filterbank coefficients).

4.3 Test results

Following previous tests with different pdfs and pdf/histogram smoothing factors (see Fig.3), the tests made here all use a truncated Gamma pdf to fully replace the original duration histograms, with a duration range factor of 2.0. The top figure in Fig. 4 shows WER* scores for explicit and implicit duration models at four noise levels for just two of the four noise types tested (subway and babble). In this figure the duration model scaling factor ω $[P(W,Q|X) \propto P(W) \cdot P(Q|W)^{\omega} \cdot p(X|Q,W)^{(1-\omega)}]$ is tuned for the clean condition. The middle figure shows the same when the scaling factor is tuned for SNR 20dB. The bottom figure shows scores averaged over all of the four noise types tested. These results show that explicit duration modelling results in between 8% and 15% absolute error rate reduction for SNRs 0 to 20 dB. For recognition in noise we need an accurate measure of recognition performance. The usual word error rate measure

$$WER = 100(S + D + I)/(H + S + D) \quad (25)$$

is neither a true percentage nor a reliable measure of recognition performance because it is not symmetric in D and I. For both duration probability scaling factor tuning and recognition scoring we therefore used the following D/I symmetric word error rate

$$WER^* = 100[1 - H/(H + S + D)(H + S + I)] \quad (26)$$

WER* is a true percentage. It is also proportional to the Pearson statistic for testing independence between a true and predicted sequence, so it is a direct measure of information loss [8].

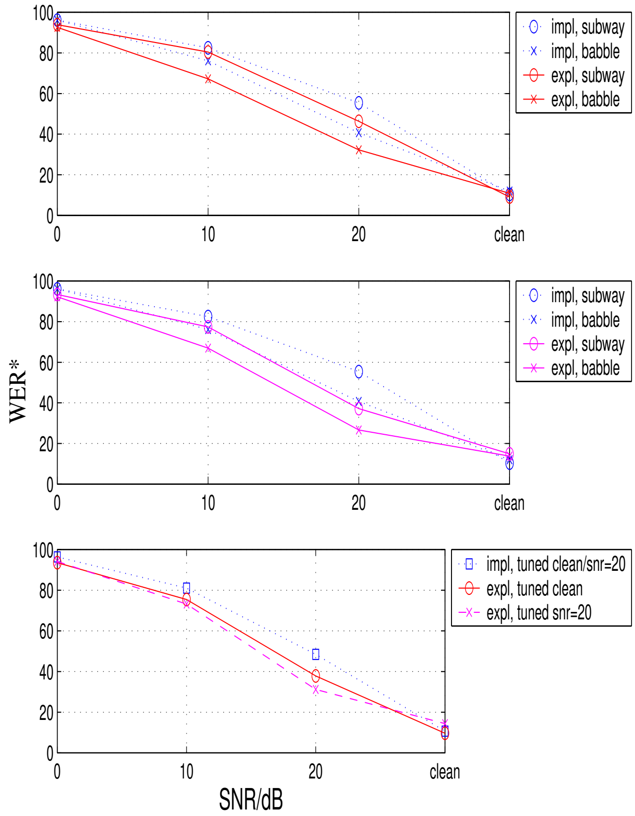

Fig.4 Performance of different parametric pdfs

Top Fig. compares WER for subway and babble noises using implicit and explicit Gamma duration models when duration prob. scaling factor is optimised for clean data. Middle Fig. shows same when scaling factor is optimised for SNR 20dB. Bottom Fig. shows results averaged over all four noise types tested*

5. DISCUSSION AND CONCLUSION

When transition probabilities depend on state duration their values at Viterbi step at t+1 are dependent on the choice of highest scoring path to each state at step t. This leaves open the possibility that a subpath of the optimal path may be eliminated at some $t < T$, in which case the solution found would be suboptimal. However, this has not prevented us from achieving a considerable performance advantage in noise. The durations cost at step t used by the model described here was: the probability that all the states so far have the identity they have; so far terminated states have the duration they have, and the currently open state will have a final duration >= its current duration. Though not guaranteeing optimality, this makes maximum use of available causal information. Algorithms using semi-hidden Markov models have previously been developed to find globally optimal solutions in conjunction with duration models [12], but these were computationally expensive. With whole word models it is not possible to enforce minimum duration constraints by increasing the number of states, as this would result in an excessively long minimum word duration.

One cannot tune the duration scaling factor to the SNR (unless an SNR estimate is available), but even when no duration probability scaling factor is used, the explicit duration models tested here still gave a very significant improvement in noise on the system tested.

One factor limiting the effectiveness of duration models is that durations vary greatly between speakers, and with rate of speech. Duration constraints would be more effective if they were combined with a reliable estimate of the current speech rate [7].

6. FUTURE DIRECTIONS

An important feature of duration models is that their performance advantage is likely to be additive when combined with other techniques for robust recognition which do not make use of duration information. Such techniques include speech enhancement prior to recognition, multistream ASR [9], and "missing data" methods [6]. This was the principal reason we decided to look again at duration modelling. After the positive results from these tests we will next be experimenting with the way in which these duration models combine with other such techniques.

Acknowledgements

This work was carried out within the EC/OFES RESPITE (REcognition of Speech by Partial Information TEchniques) project. Thanks to Jon Barker for discussions on decoder optimality.

REFERENCES

[1] Bilmes, J., N. Morgan, N., Wu, S.-L. & Bourlard, H. (1996) "Stochastic Perceptual Speech Models With Durational Dependence", Proc. ICSLP'96, pp. 1301-1304.

[2] Birshtein, D. (1995) "Robust parametric modelling of durations in hidden Markov models", Proc. ICASSP'95, pp. 548-551.

[3] Bourlard, H. (1997) "State-of-the-Art and Recent Progress in Hybrid HMM/ANN Speech Recognition", Proc. ICANN'97.

[4] Evans, M., Hastings, N. & Peacock, B. (2000) **Statistical Distributions**, Wiley series in probability and statistics.

[5] Levinson, S.E. (1986) "Continuously variable duration hidden Markov models for automatic speech recognition", Computer Speech and Language, Vol.1, pp.29-45.

[6] McCowan, I.A., Morris, A.C. & Bourlard, B. (2002) "Improving speech recognition performance of small microsphone arrays using missing data techniques", Proc ICSLP'02 (in press).

[7] Morgan, N., Fosler, E., and Mirghafori, N. (1997) "Speech recognition using on-line estimation of speaking rate", Proc. Eurospeech'97, pp. 2079-2082.

[8] Morris, A.C. (2002) "An information theoretic measure of sequence recognition performance", IDIAP TR com02-03.

[9] Morris, A.C., Hagen, A. & Bourlard, H. (2001) "MAP Combination of Multi-Stream HMM or HMM/ANN Experts", Proc. Eurospeech 2001, pp. 225-228.

[10] 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.

[11] Power, K. (1996) "Duration Modelling for improved connected digit recognition", Proc. ICSLP'96, pp. 885-888.

[12] Russel, M.J. and Moore, R.K. (1985) "Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition", Proc. ICASSP'85, pp. 5-8.