A SEGMENTAL HMM FOR SPEECH PATTERN MODELLING

Martin Russell

Speech Research Unit
DRA Malvern, Malvern, Worcs WR14 3PS, UK

ABSTRACT

A simple segmental hidden Markov model (HMM) is proposed which addresses some of the limitations of conventional HMM based methods. The important features of this approach are the use of an underlying semi-Markov process, in which state transitions are segment-synchronous rather than frame-synchronous and state duration is modelled explicitly, and a state segment model in which separate statistical processes are used to characterise "extra-segmental" and "intra-segmental" variability. A basic mathematical analysis of gaussian segmental HMMs is presented and model parameter reestimation equations are derived. The relationship between the new type of model and variable frame rate analysis and conventional gaussian mixture based hidden Markov models is explained.

1 INTRODUCTION

In the context of speech pattern modelling a number of the assumptions which the standard hidden Markov model (HMM) framework makes are clearly incorrect. For example, the independence assumption states that the probability of an acoustic vector given a particular state depends directly on the vector and the state but is otherwise independent of other vectors in the sequence. Problems associated with this assumption are compounded by the nature of the state model, in which "extra-segmental" variations (such as speaker, or choice of "target" for a particular utterance), and "intra-segmental" variations (which occur once the state target has been chosen) are characterised by a single model. Hence, in principle, the model allows extra-state factors such as identity of speaker to change in synchrony with the frame rate of the acoustic patterns.

This paper proposes a simple segmental HMM which addresses this problem. The segmental model uses an underlying semi-Markov process [5, 8] to model speech at the segment level and, at the state level, employs separate models for extra-segmental and intra-segmental sources of variability. This enables extra-segmental factors to be fixed throughout a state occupancy. The basic theory of gaussian segmental HMMs is presented, including the extension of the conventional Baum-Welch parameter estimation algorithm to this type of model. Finally, some relationships between gaussian segmental HMMs, conventional variable frame-rate analysis, and conventional HMMs with gaussian mixture densities are described.

A similar model has been studied by Peter Brown [3].

2 SEGMENTAL HMMS

A segmental HMM $\mathcal{M}$ is a hidden semi-Markov model in which extra-segmental variability associated with a state σ_i is characterised by a probability density function (PDF) b_i called the *state target PDF*. On arrival at state σ_i a *target* is chosen according to this PDF. This target is a PDF v which, intuitively, models legitimate within-segment variation once all sources of extra-segment variation have been fixed. Formally, the statistical process associated with state σ_i is defined by a PDF $b_i : \mathcal{P} \rightarrow [0, 1]$, where $\mathcal{P}$ denotes a set of PDFs defined on the set of acoustic vectors, and a state duration PDF d_i. A state duration $\mathcal{D}_i$ is chosen according to the PDF d_i and a sequence of $\mathcal{D}_i$ vectors is then generated randomly and independently according to the target v.

Given a sequence of observation vectors $y = y_1, ..., y_T$, the joint probability of a subsequence $y_{t_{i-1}+1}^{t_i} = y_{t_{i-1}+1}, ..., y_{t_i}$ of length $\mathcal{D}_i$ and a particular target v given state σ_i is given by:

$$P_{\sigma_i}(y_{t_{i-1}+1}^{t_i}, v) = d_i(\mathcal{D}_i) b_i(v) \prod_{t=t_{i-1}+1}^{t_i} v(y_t), \qquad (1)$$

and the probability of $y_{t_{i-1}+1}^{t_i}$ given σ_i is the integral $P_{\sigma_i}(y_{t_{i-1}+1}^{t_i}) = \int_v P_{\sigma_i}(y_{t_{i-1}+1}^{t_i}, v)$.

This paper presents an analysis of the alternative probability function

$$\hat{P}_{\sigma_i}(y_{t_{i-1}+1}^{t_i}) = P_{\sigma_i}(y_{t_{i-1}+1}^{t_i}, \hat{v}), \qquad (2)$$

0-7803-0946-4/93 $3.00 © 1993 BCC

where $\hat{v}$ is the target which maximises $P_{\sigma_i}(y^{t_i}_{t_{i-1}+1}, v)$.

$$\hat{v} = \arg\max_{v} P_{\sigma_i}(y^{t_i}_{t_{i-1}+1}, v) \tag{3}$$

Given a state sequence $x = x_1, ..., x_T$, such that a transition from state σ_i to state σ_{i+1} occurs at time t_i, the "joint probability" $\hat{P}(y, x|\mathcal{M})$ of y and x given $\mathcal{M}$, and the "probability" $\hat{P}(y|M)$ of y given $\mathcal{M}$ are given by[1]:

$$\begin{aligned} \hat{P}(y, x|\mathcal{M}) &= \prod_{i=1}^{N} a_{i-1,i} P_{\sigma_i}(y^{t_i}_{t_{i-1}+1}), & (4) \\ \hat{P}(y|M) &= \sum_{x} \hat{P}(y, x|\mathcal{M}) & (5) \end{aligned}$$

These and similar expressions can be computed using the extensions of standard HMM algorithms to semi-Markov processes [5, 8].

3 GAUSSIAN SEGMENTAL HMMS

Now consider the case where acoustic vectors are drawn from n-dimensional space $\mathbf{R}^n$, and for each state σ_i a target is any gaussian PDF defined on $\mathbf{R}^n$ with fixed variance τ_i. Then a target $v = \mathcal{N}_{c,\tau_i}$ can be identified with its mean c, and $\mathcal{P} = \mathbf{R}^n$. If the state target PDF b_i is a gaussian PDF defined on $\mathbf{R}^n$, with mean μ_i and variance γ_i, then the resulting model $\mathcal{M}$ will be called a gaussian segmental HMM (GSHMM). The number of parameters in a GHSMM is only increased by the variance terms τ_i ($i = 1, ..., N$) relative to a gaussian HMM.

In this case it can be shown that the target (mean) $\hat{c}$ which maximises $P_{\sigma_i}(y, c)$ is given by:

$$\hat{c} = \frac{\mu_i \tau_i + \sum_{t=t_{i-1}+1}^{t_i} y_t \gamma_i}{\tau_i + T\gamma_i} \tag{6}$$

Thus the "best target" is a linear combination of the expected target for state σ_i and the actual observations. If τ_i is large, so that the observations are not expected to be tightly constrained by the target process, then $\hat{c}$ is biased towards the state mean μ_i. But if γ_i is large and τ_i is small, $\hat{c}$ is biased towards the actual observations.

4 PARAMETER REESTIMATION FOR GAUSSIAN SEGMENTAL HMMS

A Baum-Welch type parameter reestimation process can be derived for GSHMMs. As above, let $\mathcal{M}$ be an N-state GSHMM with parameters μ_i, γ_i and τ_i, and let y be a sequence of observations vectors. Let $\bar{\mathcal{M}}$ be the GSHMM with parameters $\bar{\mu}_i$. $\bar{\gamma}_i$ and $\bar{\tau}_i$, defined by:

$$\begin{aligned} \bar{\mu}_i &= \frac{\sum_{x \in S_i} P(y, x|\mathcal{M}) K_{x,i} \sum_{t=t_{i-1}+1}^{t_i} y_t}{\sum_{x \in S_i} P(y, x|\mathcal{M}) K_{x,i} \mathcal{D}_i} & (7) \\ \bar{\gamma}_i &= \frac{\sum_{x \in S_i} P(y, x|\mathcal{M})(\bar{\mu}_i - \hat{\bar{c}}_{x,i})^2}{\sum_{x \in S_i} P(y, x|\mathcal{M})} & (8) \\ \bar{\tau}_i &= \frac{\sum_{x \in S_i} P(y, x|\mathcal{M}) \sum_{t=t_{i-1}+1}^{t_i} (\hat{\bar{c}}_{x,i} - y_t)^2}{\sum_{x \in S_i} P(y, x|\mathcal{M}) \mathcal{D}_i} & (9) \end{aligned}$$

where $S_i = \{x : x_t = \sigma_i \text{ for some } t\}$, $K_{x,i} = (\bar{\tau}_i + \mathcal{D}_i \bar{\gamma}_i)$, and $\hat{\bar{c}}_{x,i} = (\bar{\mu}_i \bar{\tau}_i + \sum_{t=t_{i-1+1}}^{t_i} y_t \bar{\gamma}_i) K_{x,i}^{-1}$. If (i) $\bar{\gamma}_i > \bar{\tau}_i$ for $i = 1, ..., N$, and (ii) $y = y_1, ..., y_T$ is not constant, then $\hat{P}(y|\bar{\mathcal{M}}) \geq \hat{P}(y|\mathcal{M})$.

The proof follows [1, 6]. Equations (7), (8) and (9) occur as the unique critical point of an auxilliary function $\hat{Q}(\mathcal{M}, \bar{\mathcal{M}})$, defined by:

$$\hat{Q}(\mathcal{M}, \bar{\mathcal{M}}) = \sum_{x} \hat{P}(y, x|\mathcal{M}) log \hat{P}(y, x|\bar{\mathcal{M}}) \tag{10}$$

Properties (i) and (ii) are used to show that this function is concave and tends to $-\infty$ as $\bar{\mathcal{M}}$ approaches the boundary of the parameter space. The derivation of equation (7) is included for completeness. Differentiating equation (10) with respect to $\bar{\mu}_i$ leads to,

$$\begin{aligned} \frac{\partial}{\partial \bar{\mu}_i} \hat{Q}(\mathcal{M}, \bar{\mathcal{M}}) &= \sum_{x} \hat{P}(y, x|\mathcal{M}) \frac{\partial}{\partial \bar{\mu}_i} log \hat{P}(y, x|\bar{\mathcal{M}}) \\ &= \sum_{x \in S_i} \hat{P}(y, x|\mathcal{M}) (\frac{\partial}{\partial \bar{\mu}_i} log \mathcal{N}_{(\bar{\mu}_i, \bar{\gamma}_i)}(\hat{\bar{c}}_{x,i}) \\ &+ \sum_{t=t_{i-1}+1}^{t_i} \frac{\partial}{\partial \bar{\mu}_i} log \mathcal{N}_{(\hat{\bar{c}}_{x,i}, \bar{\tau}_i)}(y_t)) & (11) \end{aligned}$$

Taking into account the expression for $\hat{\bar{c}}_{x,i}$ above

$$\frac{\partial}{\partial \bar{\mu}_i} log \mathcal{N}_{(\bar{\mu}_i, \bar{\gamma}_i)}(\hat{\bar{c}}_{x,i}) = \frac{\mathcal{D}_i(\hat{\bar{c}}_{x,i} - \bar{\mu}_i)}{K_{x,i}} \tag{12}$$

and,

$$\frac{\partial}{\partial \bar{\mu}_i} log \mathcal{N}_{(\hat{\bar{c}}_{x,i}, \bar{\tau}_i)}(y_t) = \frac{(y_t - \hat{\bar{c}}_{x,i})}{K_{x,i}} \tag{13}$$

where $O = \sum_{t=t_{i-1}+1}^{t_i} y_t$. Therefore,

$$\begin{aligned} \frac{\partial}{\partial \bar{\mu}_i} \hat{Q}(\mathcal{M}, \bar{\mathcal{M}}) &= \sum_{x \in S_i} \hat{P}(y, x|\mathcal{M}) (\frac{\mathcal{D}_i(\hat{\bar{c}}_{x,i} - \bar{\mu}_i)}{K_{x,i}} \\ &+ \sum_{t=t_{i-1}+1}^{t_i} \frac{(y_t - \hat{\bar{c}}_{x,i})}{K_{x,i}}) & (14) \end{aligned}$$

[1] to simplify notation it will be assumed that (i) the underlying Markov process is strictly left-right, and (ii) all observations are scalar. Neither of these assumptions are necessary

Setting this partial derivative to zero leads directly to equation (7).

Note that equations (7), (8) and (9) do not constitute valid Baum-Welch type reestimation formulae since the parameters of $\bar{\mathcal{M}}$ appear on the right-hand sides of the equations. Intuitively correct reestimation formulae are obtained by replacing $\bar{\mu}_i$, $\bar{\gamma}_i$ and $\bar{\tau}_i$ with μ_i, γ_i and τ_i on ther right-hand-sides of the equations, but the usefulness of the resulting formulae for parameter reestimation needs to be tested experimentally.

5 RELATIONSHIP WITH GAUSSIAN MIXTURE DENSITIES

A class of state output PDFs which is commonly used with conventional HMMs is the class of gaussian mixture densities. In such an HMM the state output PDF b_i associated with the ith state has the form

$$b_i(o) = \sum_{j=1}^{J} w_j \mathcal{N}_{(\mu_j, \gamma_j)}(o) \qquad (15)$$

for any observation o, where $\sum_{j=1}^{J} w_j = 1$. There is also a continuous version:

$$b_i(o) = \int_j w(j) \mathcal{N}_{(\mu_j, \gamma_j)}(o) dj \qquad (16)$$

where $\int_j w(j)dj = 1$. Parameter reestimation formulae for such models have been established in [6] and [4].

Gaussian mixtures are used to compensate for the fact that the observations associated with a particular state will not in general conform with a single gaussian PDF. This is particularly true if the models are used to characterise speech from a number of speakers. Thus, gaussian mixtures are typically used to model broad sources of extra-segmental variablity and hence, from the viewpoint of this paper, they may exacerbate the problems associated with the independence assumption within a state.

The segment model proposed here is clearly related to (16), however in the segmental model a single component of the continuous mixture is chosen on entering a state and all observations emitted during a particular state occupancy are drawn from that component.

6 RELATIONSHIP WITH VARIABLE FRAME RATE ANALYSIS

The gaussian segmental HMM based analysis proposed here can be interpreted as a natural extension and integration of conventional Variable Frame Rate (VFR) analysis and hidden Markov modelling.

VFR analysis is a method for data-rate reduction which has been shown to give improved performance over fixed frame rate analysis for automatic speech recognition [7]. In its simplest form VFR is used to remove vectors from an observation sequence. A distance is computed between the current observation vector and the most recently retained vector, and the current vector is discarded if this distance falls below a threshold T. When a new observation vector causes the distance to exceed the threshold, the new vector is kept and becomes the most recently retained vector. VFR analysis replaces sequences of similar vectors with a single vector, and hence reduces the amount of computation required for recognition.

This basic VFR algorithm can be improved:

(i) Rather than replacing a sequence of acoustic vectors $y_s, ..., y_t$ with y_s, the first vector in the sequence, it should replaced with an average $\bar{y}_s^t$ over the sequence.

(ii) For a finite sequence $y = y_1, ..., y_T$ the "left-right" threshold based segmentation used in the basic VFR algorithm should be replaced with a "global" dynamic programming based segmentation algorithm ([2]) which partitions the sequence y into M subsequences $y_1^{t_1}, ..., y_{ti-1+1}^{t_i}, ..., y_{t_{M-1}+1}^{t_M}$ $(1 \leq t_1 \leq ... \leq t_M = T)$ such that some criterion

$$Dist(t_1, ..., t_i, ..., t_M) = \sum_{i=1}^{M} D(y_{t_{i-1}+1}^{t_i}) \qquad (17)$$

is minimised. $D(y_{t_{i-1}+1}^{t_i})$ is typically a distortion measure on the sequence $y_{t_{i-1}+1}^{t_i}$, for example the sum of euclidean distances between vectors in the sequence and the sequence mean.

(iii) In HMM based speech pattern processing it is clearly sub-optimal to segment the sequence of acoustic observation vectors and discard information during VFR analysis, and then to perform a second state-level segmentation. The segmentation of the observation sequence during VFR analysis should be integrated with the state-level segmentation performed in the model based analysis.

Extending the basic VFR algorithm in these ways leads naturally to a segmental HMM based analysis. Suppose that $\mathcal{M} = (\pi, A, \{b_i\})$ is a HMM, with $b_i = \mathcal{N}_{(\mu_i, \gamma_i)}$, and that $y = y_1, ..., y_t, ..., y_T$ is a sequence of acoustic vectors in R^n. In a dynamic programming based VFR scheme of the type alluded to above, after VFR analysis the sequence y is represented by the sequence $\bar{y} = \bar{y}_1^{t_1}, ..., \bar{y}_{t_{i-1}+1}^{t_i}, ..., \bar{y}_{t_{M-1}+1}^{t_M}$, where $\bar{y}_{t_{i-1}+1}^{y_t}$ denotes an average over the sequence $y_{t_{i-1}+1}^{y_t}$.

During subsequent HMM based processing, dynamic programming is used again to find a state sequence $x = x_1, ..., x_M$ relative to the HMM $\mathcal{M}$, such that the probability

$$P(\bar{y}, x|\mathcal{M}) = \prod_{i=1}^{M} a_{x_{i-1},x_i} d_{x_i}(\mathcal{D}_i) b_{x_i}(\bar{y}^{t_i}_{t_{i-1}+1}) \quad (18)$$

is maximised. d_{x_i} is a state dependent duration PDF which is applied to the VFR count $\mathcal{D}_i$.

Ideally the two equations (17) and (18) should be optimised jointly. Let

$$D(y^{t_i}_{t_{i-1}+1}) = \sum_{t=t_{i-1}+1}^{t_i} D_{EUC}(y_t, \bar{y}^{t_i}_{t_{i-1}+1}) \quad (19)$$

where D_{EUC} denotes the squared euclidean metric. Then, since

$$D_{EUC}(y_t, \bar{y}^{t_i}_{t_{i-1}+1}) = -K_1 log(\mathcal{N}_{\bar{y}^{t_i}_{t_{i-1}+1},1}(y_t) + K_2 \quad (20)$$

where K_1 and K_2 are constants, minimising equation (17) is equivalent to maximising the quantity

$$P(t_1, ..., t_i, ..., t_M) = \prod_{i=1}^{M} \prod_{t=t_{i-1}+1}^{t_i} \mathcal{N}_{\bar{y}^{t_i}_{t_{i-1}+1},1}(y_t) \quad (21)$$

Combining (18) and (21) gives an evaluation criterion $Q((\bar{y}, x|\mathcal{M})$ for a joint VFR-HMM analysis scheme which satisfies (i), (ii) and (iii) above:

$$Q(\bar{y}, x|\mathcal{M}) = \prod_{i=1}^{M} a_{x_{i-1},x_i} d_{x_i}(\mathcal{D}_i) b_{x_i}(\bar{y}^{t_i}_{t_{i-1}+1}) \prod_{t=t_{i-1}+1}^{t_i} \mathcal{N}_{\bar{y}^{t_i}_{t_{i-1}+1},1}(y_t) \quad (22)$$

But this has the same form as equation (4), with $\tau_i = 1$, for all i, and $\bar{y}^{t_i}_{t_{i-1}+1} = \hat{c}_{x,i}$.

In other words, replacing the basic VFR analysis procedure described above with a dynamic programming based method and integrating this with the higher-level HMM based processing leads naturally to the type of gaussian segmental HMM based analysis proposed in this paper.

7 SUMMARY

This paper presents a segmental HMM which addresses some of the limitations of conventional HMMs in the context of speech pattern modelling. Specifically the segmental model alleviates some of the problems caused by the (time) independence assumption and its interaction with the single process state model. A basic mathematical analysis is presented and Baum-Welch type parameter reestimation formulae are presented for the special case of Gaussian segmental HMMs. Interesting relationships between segmental HMMs, conventional variable frame rate analysis, and continuous gaussian mixture HMMs are exposed.

8 ACKNOWLEDGEMENTS

Thanks are due to my colleagues at SRU, to John Bridle at Dragon Systems UK, and to Mark Gales at Cambridge University Engineering Department for their helpful comments on this work.

References

[1] L E Baum, T Petrie, G Soules and N Weiss, "A maximisation technique occuring in the statistical analysis of probabilistic functions of Markov chains", The Annals of Mathematical Statistics, Vol. 41, No. 1, 164-171, 1970.

[2] J S Bridle and N C Sedgwick, "A method for segmenting acoustic patterns with applications to automatic speech recognition", Proc ICASSP'77, 656-659, 1977.

[3] P Brown, Personal communication.

[4] B-H Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains, AT&T Tech. J., vol 64, no. 4, pp 1235-1249, 1985.

[5] S E Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition", Computer Speech and Language, Vol 1, No 1, 29-46, March 1986.

[6] L Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources", IEEE Trans. Information Theory, vol IT-28, 5, 1982.

[7] S M Peeling and K M Ponting, "Variable frame rate analysis in the ARM continuous speech recognition system", Speech Communication 10, pp 155-162, 1991.

[8] M J Russell, "Maximum likelihood hidden semi-Markov model parameter estimation for automatic speech recognition", RSRE Memorandum 3837, July 1985.