Proceedings of the Institute of Acoustics

A SEGMENTAL STATISTICAL MODEL FOR SPEECH PATTERN PROCESSING

Martin Russell

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

1. INTRODUCTION

At present the most successful automatic speech recognition systems, in terms of recognition accuracy, are those which use hidden Markov models (HMMs) to model speech at the acoustic level and dynamic programming based recognition algorithms which find the best interpretation of an unknown speech pattern in terms of the output of a sequence of HMMs. The most recent systems use HMMs to model speech at the phoneme level in order to address medium and large vocabularies and to avoid vocabulary-specific training.

This success is due to two factors. Firstly HMMs provide a formal statistical framework which is broadly appropriate for modelling speech patterns. This framework is able simultaneously to accomodate the time-varying nature of speech patterns, through the underlying Markov process, and the variable segmental structure of these patterns through the statistical processes which are identified with the states of the model. Secondly there exist computationally useful and rigorous mathematical methods for automatically optimising the parameters of a set of HMMs relative to training data, and for classifying an unknown speech pattern given a set of HMMs.

These two factors together constitute a powerful tool for speech recognition. However, from the perspective of speech modelling a number of the assumptions which the HMM framework makes are clearly incorrect. For example, the independence assumption states that the probability of an acoustic vector given a particular state depends 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, 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 these problems. The new model uses an underlying semi-Markov process [4, 7] 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 in sections 4 and 5, including the extension of the conventional Baum-Welch parameter estimation algorithm to this type of model. Finally, the relationships between gaussian segmental HMMs, variable frame-rate analysis, and HMMs with gaussian mixture densities are explored in sections 6 and 7.

A similar model has been studied by P Brown at IBM [9].

2. CONVENTIONAL HIDDEN MARKOV AND HIDDEN SEMI-MARKOV MODELS

In the HMM based approach to speech pattern modelling it is assumed that a sequence of acoustic observation vectors, $y = y_1, y_2, ..., y_t, ..., y_T$ corresponding to a given speech signal, is a probabilistic function of a hidden state sequence $x = x_1, x_2, ..., x_t, ..., x_T$ where each x_t is drawn from a finite set of states $\sigma = \{\sigma_1, .., \sigma_N\}$. The sequential and durational statistics of x are determined by a transition probability matrix

$A = [a_{ij}]_{i,j=1,...,N}$

where, $a_{ij} = Prob(x_t = \sigma_j | x_{t-1} = \sigma_i)$ is the probability of a transition from state σ_i to state σ_j, and an initial state probability vector

$\pi = [\pi_i]_{i=1,...,N}$

where $\pi_i = Prob(x_1 = \sigma_i)$. The pair $\mathcal{M} = (\pi, A)$ define an N state Markov process. The relationship between the observation vectors y_t and the hidden states x_t is defined by a set of probability density functions (PDFs) $\{b_i\}_{i=1,...,N}$, where $b_i(o) = Prob(y_t = o | x_t = \sigma_i)$ is the probability that the observation o is associated with state σ_i. The triple $\mathcal{H} = (\pi, A, \{b_i\})$ defines a hidden Markov process. The process is called hidden because it is not possible to unambiguously infer the state sequence which gave rise to a particular observation sequence. Intuitively the PDF associated with a particular state models variations in the acoustic vectors for the corresponding speech sound, and the sequence of states in an HMM models the sequence of sounds in the corresponding utterance.

One limitation of HMMs is the underlying geometric model of state (and hence speech segment) duration, which assigns maximum probability to a duration of 1 time-unit and progressively smaller probabilities to longer durations. A solution is to replace the underlying Markov process in an HMM with a semi-Markov process in which a state duration PDF $\mathcal{D}_i$ is associated with each state σ_i. For $d = 1, 2, 3, .., \mathcal{D}_i(d)$ is the probability of occupying state σ_i for precisely d time units.

A hidden semi-Markov process is a probabilistic function of a semi-Markov process. More precisely, an N state hidden semi-Markov model (HSMM) [4, 7], or Variable Duration HMM, $\mathcal{S} = (\pi, A, \{\mathcal{D}_i\}, \{b_i\})$ comprises an N-state Markov model $\mathcal{M} = (\pi, A)$, a set of N state duration PDFs $\mathcal{D}_1, ..., \mathcal{D}_N (\mathcal{D}_i : \mathbf{N} \rightarrow [0, 1])$, and a set of N state output PDFs $b_1, ..., b_N$ i, $b_i : \mathbf{R}^d \rightarrow [0, 1]$ (the symbols $\mathbf{N}$ and $\mathbf{R}^d$ denote the positive integers and real d dimensional space respectively). Intuitively one can visualise such a process as follows. At some time t state $x_m = \sigma_i$ is entered and a duration d_m is chosen randomly according to the state duration PDF $\mathcal{D}_i$. A sequence $y_t, ..., y_{t+d_m-1}$ of d_m acoustic vectors is then generated randomly and independently according to the PDF b_i. The process then moves to a new state σ_j according to A.

The principle of dynamic programming, and hence the standard dynamic programming based recognition algorithms, can be extended from HMMs to HSMMs. Also it has been shown that the Baum-Welch parameter estimation algorithm can be generalised to HSMMs with discrete, Poisson or Gamma state duration PDFs ([7, 4]). However, the need to explicitely consider times $t - \delta$ ($\delta = 1, 2, ..., d_{max}$) during HSMM based computations leads to an increase in computational load relative to HMMs.

To date, HSMMs have primarily been used to overcome the limitations of HMMs with respect to speech segment duration modelling. Consequently, because the resulting improvements in recognition accuracy are generally relatively modest and the increase in computational load is relatively high, there has been little recent work in this area. The objective of this paper is to show that the segment based formalism provided by HSMMs can provide a basis for addressing other limitations of HMMs.

3. SEGMENTAL HIDDEN MARKOV MODELS

A segmental HMM (SHMM) is a hidden Markov model in which the statistical process associated with a state models sequences of frames (segments) rather than individual frames. SHMMs are asynchronous, in the sense that whereas state transitions in a conventional HMM are synchronised with the frame-rate of the acoustic front-end, in a SHMM they are not. A HSMM is a SHMM in which a segment is a sequence of acoustic vectors drawn independently from a single PDF.

The SHMMs studied in this paper employ a more sophisticated segment model in which separate processes are used to cope with extra-segmental and intra-segmental sources of variability. Extra-segment variability associated with a state σ_i is characterised by a 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 $\mathcal{D}_i$. A state duration d_i is chosen according to the PDF $\mathcal{D}_i$ and a sequence of 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 d_i and a particular target v given state σ_i is given by:

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

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

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

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

$$\hat{P}(y, x|\mathcal{M}) = \prod_{i=1}^{N} a_{i-1,i} \hat{P}_{\sigma_i}(y_{t_{i-1}+1}^{t_i}), \tag{3}$$

$$\hat{P}(y|M) = \sum_{x} \hat{P}(y, x|\mathcal{M}) \tag{4}$$

These and similar expressions can be computed using the extensions of standard HMM algorithms to semi-Markov processes [4, 7].

4. 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 $\mathcal{P} = \mathbf{R}^n$ and a target $v = \mathcal{N}_{c,\tau_i}$ can be identified with its mean c. 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 GSHMM is only increased by the variance terms τ_i $(i = 1, ..., N)$ relative to a gaussian HMM.

It can be shown [8] that the target (mean) $\hat{c}$ which maximises $P_{\sigma_i}(y, c)$ is given by:

$$\hat{c} = \frac{\mu_i \tau_i + \sum_{t=t_{i-1}+1}^{t_i} y_t \gamma_i}{\tau_i + T\gamma_i} \tag{5}$$

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.

5. PARAMETER REESTIMATION FOR GAUSSIAN SEGMENTAL HMMS

A Baum-Welch type parameter reestimation process has been derived for GSHMMs [8]. 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:

$$\bar{\mu}_i = \frac{\sum_{x \in S_i} P(y, x|\mathcal{M}) \sum_{t=t_{i-1}+1}^{t_i} y_t}{\sum_{x \in S_i} P(y, x|\mathcal{M}) \mathcal{D}_i} \tag{6}$$

$$\bar{\gamma}_i = \frac{\sum_{x \in S_i} P(y, x|\mathcal{M})(\bar{\mu}_i - \bar{\hat{c}}_{x,i})^2}{\sum_{x \in S_i} P(y, x|\mathcal{M})} \tag{7}$$

[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

$$\bar{\tau}_i = \frac{\sum_{x \in S_i} P(y,x|\mathcal{M}) \sum_{t=t_{i-1}+1}^{t_i} (\hat{c}_{x,i} - y_t)^2}{\sum_{x \in S_i} P(y,x|\mathcal{M}) \mathcal{D}_i} \tag{8}$$

where $S_i = \{x : x_t = \sigma_i \text{ for some } t\}$ and $\hat{c}_{x,i} = \frac{\bar{\mu}_i \bar{\tau}_i + \sum_{t=t_{i-1}+1}^{t_i} y_t \bar{\tau}_i}{\bar{\tau}_i + \mathcal{D}_i \bar{\tau}_i}$.

If (i) $\bar{\tau}_i > \bar{\tau}_i$ 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, 5] and is presented in full in [8]. It is first shown that the auxillairy function $\hat{Q}(\mathcal{M}, \bar{\mathcal{M}})$ defined by:

$$\hat{Q}(\mathcal{M}, \bar{\mathcal{M}}) = \sum_x \hat{P}(y,x|\mathcal{M}) log \hat{P}(y,x|\bar{\mathcal{M}}) \tag{9}$$

has the property that if $\hat{Q}(\mathcal{M}, \bar{\mathcal{M}}) \geq \hat{Q}(\mathcal{M}, \mathcal{M})$ then $\hat{P}(y|\bar{\mathcal{M}}) \geq \hat{P}(y|\mathcal{M})$. Therefore, in order to increase $\hat{P}(y,x|\mathcal{M})$ it is sufficient to find a model $\bar{\mathcal{M}}$ which maximises $\hat{Q}(\mathcal{M}, \bar{\mathcal{M}})$, as a function of $\bar{\mathcal{M}}$. Equations (6), (7) and (8) occur as a critical point of $\hat{Q}(\mathcal{M}, .)$. 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, guaranteeing that the critical point is unique and is a maximum.

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 [6]. 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 in a number of ways:

(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 $y_1^{t_1}, ..., y_{t_{i-1}+1}^{t_i}, ..., y_{t_{M-1}+1}^{t_M}$ $(1 \leq t_1 \leq ... \leq t_M = T)$ such that some criterion

$$Dist(t_1, ..., t_i, ..., t_M) = \sum_{i=1}^{M} D(y_{t_{i-1}+1}^{t_i}) \tag{10}$$

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 Markov model 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^d. 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 $\bar{y}_{t_{i-1}+1}^{t_i}$ denotes an average over the sequence $y_{t_{i-1}+1}^{t_i}$.

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

$$P(\bar{y}, x|\mathcal{M}) = \prod_{i=1}^{M} a_{x_{i-1}, x_i} \mathcal{D}_{x_i}(d_i) b_{x_i}(\bar{y}_{t_{i-1}+1}^{t_i}) \tag{11}$$

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

Ideally the two equations (10) and (11) should be optimised jointly. Let

$$D(y_{t_{i-1}+1}^{t_i}) = \sum_{t=t_{i-1}+1}^{t_i} D_{EUC}(y_t, \bar{y}_{t_{i-1}+1}^{t_i}) \tag{12}$$

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

$$D_{EUC}(y_t, \bar{y}_{t_{i-1}+1}^{t_i}) = -K_1 log(\mathcal{N}_{\bar{y}_{t_{i-1}+1}^{t_i}, I}(y_t) + K_2 \tag{13}$$

where K_1 and K_2 are constants, minimising equation (10) is equivalent to maximising the quantity

$$P(t_1, ..., t_i, ..., t_M) = \prod_{i=1}^{M} \prod_{t=t_{i-1}+1}^{t_i} \mathcal{N}_{\bar{y}_{t_{i-1}+1}^{t_i}, I}(y_t) \tag{14}$$

Combining (11) and (14) gives an evaluation criterion for a VFR analysis scheme which satisfies (i), (ii) and (iii) above:

$$P(\bar{y}, x|\mathcal{H}) = \prod_{i=1}^{M} a_{x_{i-1},x_i} \mathcal{D}_{x_i}(d_i) b_{x_i}(\bar{y}_{t_{i-1}+1}^{t_i}) \prod_{t=t_{i-1}+1}^{t_i} \mathcal{N}_{\bar{y}_{t_{i-1}+1}^{t_i},1}(y_t) \tag{15}$$

But this has the same form as equation (3), with $\tau_i = 1$, for all i, and $\bar{y}_{t_{i-1}+1}^{t_i} = \hat{c}_{x,i}$.

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. Hence segmental HMMs can be regarded as an extension and integration of VFR and HMM-based analysis.

7. 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

$$b_i(o) = \sum_{j=1}^{J} w_j \mathcal{N}_{(\mu_j,\tau_j)}(o) \tag{16}$$

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

$$b_i(o) = \int_j w(j) \mathcal{N}_{(\mu_j,\tau_j)}(o) dj \tag{17}$$

where $\int_j w(j)dj = 1$. Parameter reestimation formulae for such models have been established in [5] and [3], and in [5] respectively.

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 exacerbate the problems associated with the independence assumption within a state.

The segment model proposed here is clearly related to (17), however in the new type of 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.

8. SUMMARY

This paper presents the basic theory of a new segmental HMM which addresses some of the limitations of conventional HMMs in the context of speech pattern modelling. The new

model is computationally useful in that it admits extensions of the conventional HMM classification and parameter estimation algorithms. Interesting relationships between segmental HMMs, conventional variable frame rate analysis, and continuous gaussian mixture HMMs have been described.

REFERENCES

[1] L E BAUM, T PETRIE, G SOULES & 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 & N C SEDGWICK, "A method for segmenting acoustic patterns with applications to automatic speech recognition", Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, pp 656-659, (1977)

[3] 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)

[4] S E LEVINSON, "Continuously variable duration hidden Markov models for automatic speech recognition", Computer Speech and Language, Vol 1, No 1, 29-46, (1986)

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

[6] S M PEELING & K M PONTING, "Variable frame rate analysis in the ARM continuous speech recognition system", Speech Communication 10, pp 155-162, (1991)

[7] M J RUSSELL, "Maximum likelihood hidden semi-Markov model parameter estimation for automatic speech recognition", RSRE Memorandum 3837, (1985)

[8] M J RUSSELL, "A segmental hidden Markov model for speech pattern processing", DRA Malvern memorandum 4599, July 1992.

[9] J S BRIDLE, Personal communication.