A General Strategy for Hidden Markov Chain Parameterisation in Composite Feature-Spaces

David Windridge, Richard Bowden, and Josef Kittler

Centre for Vision, Speech and Signal Processing
Dept. of Electronic & Electrical Engineering, University of Surrey,
Guildford, GU2 5XH Surrey, UK
Tel.: +44 1483 876043
d.windridge@eim.surrey.ac.uk

Abstract. A general technique for the construction of hidden Markov models (HMMs) from multiple-variable time-series observations in noisy experimental environments is set out. The proposed methodology provides an ICA-based feature-selection technique for determining the number, and the transition sequence, of underlying hidden states, along with the statistics of the observed-state emission characteristics. In retaining correlation information between features, the method is potentially far more general than Gaussian mixture model HMM parameterisation methods such as Baum-Welch re-estimation, to which we demonstrate our method reduces when an arbitrary separation of features, or an experimentally-limited feature-space is imposed.

1 Introduction

1.1 Objective

We shall set out a generic methodology for the automated generation of maximally generalising hidden Markov models given an initial time-sequence (or possibly Markovian) model of observed feature-space transitions within a noisy experimental environment. That is, for any experimental situation where multiple measurements are associated with individual temporally-ordered observation states, and where the observed state-transition mechanisms are governed to any extent by stochastic processes detrimental to classification, we propose to derive a method for automated hidden Markov model (HMM) generation that will provide maximal generalisation under the stated experimental assumptions.

In order to qualify as more general than existing HMM parameterisation techniques, the elaborated methodology might be expected to incorporate existing methods as a limiting subset: when measurement modalities are artificially constrained, we will therefore, in section 2.21, demonstrate an equivalence between the results of our technique and that of the Gaussian mixture-model estimation methods (such as Baum-Welch re-estimation [1]).

Other issues arise in the application of this method, particularly the possibility of sub- and super-sampling of hidden states, with their attendant consequences for classification performance; these, however, fall beyond the scope

of the current paper, and will be dealt with in a future publication. Also reserved for future publication is experimental evidence for the utility of this method, demonstrating an approximately 50 percent performance improvement over ‘naïve’ Markovian modelling of observed transitions in typical experimental spaces: indeed, the method may be considered as a technique for transforming naturally between Markovian and hidden-Markovian models of a particular feature-space.

1.2 Overview of Technique

The mechanics of the proposed methodology will centre on determining both the number of, and the transition sequence of, a set of putative hidden states underlying experimental scenarios characterised by a certain common type of noise model, in addition to determining the emission characteristics of the corresponding observation states. More specifically, the technique for generating the HMM structure derives from the many-to-one mapping that spontaneously arises between the space of observed feature-space transitions and the corresponding ‘hidden’ state-transitions occurring in a subspace defined by the underlying ‘operational’ manifold specific to the nature of the experimental system under consideration. Most importantly, this mapping will be of an essentially stochastic nature, as required by the relation of hidden and observed states in hidden Markov modelling [cf eg. 2]. It is this stochasticity, in particular, that accounts for the failure of purely Markovian descriptions of observed feature-space transitions to generalise within such experimental environments, requiring the utilisation of hidden Markov techniques for optimal classification.

Since there will, in general, be several distinct noise sources represented within an observed feature-space, characterised by their collective stochastic independence, the technique adopted to bring about the distinction of the signal and noise components consists in an initial independent component analysis (ICA) [enacted via coordinate transformation], followed by a secondary feature-selection (FS) procedure, this process being sufficient to completely eliminate all of the independent noise features. It is, in particular, only the prior ICA-transformation stage that permits feature-section to remove the independent noise sources. Without this step, since the original features contain both information and noise signals, feature-selection would, in removing both simultaneously, invariably act to degrade overall classification performance.

Once this ICA+FS procedure has taken place, there then exists an implicit many-to-one mapping between the observed feature-space and the optimally-compactly described signal-space arising from the procedure. However, since the observed state-space into which the signal subspace is mapped is defined by the noise features removed by feature-selection, these mappings are dictated purely by chance, and not by the *specific* configuration of the noise vectors in the training set. Thus it becomes possible to ignore the particular transition structure of the mapped states and consider rather a stochastic model of the behaviour of observed states with respect to the ‘hidden’ states in the ICA+FS space. We have thus all the ingredients required, given an observed set of feature-

space transition sequences, to arrive at a maximally generalising hidden Markov model in which a naturally occurring distinction between hidden states and observed states is modelled by some emission probability. Moreover, once the ICA+FS procedure has taken place, it becomes possible to eliminate it entirely from the classifying process, working only with the HMM model so derived. Setting out precisely how this process takes place will be the concern of the next section of this paper.

2 Details of Methodology

2.1 ICA Description

The first requirement of our technique is the ICA transformation of the observed pattern space: we shall give a very brief overview of this process as follows (with more detailed and general treatments being available in, for instance, [3] and [4]): Let $\mathbf{t}$ be a vector describing the state of the f_n individual feature channels over the complete temporal range, t . We wish to describe this vector in terms of the fully independent feature-set vector, $\mathbf{t}'$, of presently unknown dimensionality. The two vector quantities are related by a mixing matrix, M , via $\mathbf{t}' = M\mathbf{t}$. The determination of the matrix M is hence the objective of our calculation, the first stage of which is the decorrelating (or ‘whitening’, ‘sphering’) of the original input space. That is, we shall require a linear transformation $\mathbf{t}_w = W\mathbf{t}'$ such that the expectation $E(\mathbf{t}_w\mathbf{t}_w^T)$ is equal to I (I being the identity matrix). A simple solution to this constraint exists via the expansion:

$$I = E(\mathbf{t}_w\mathbf{t}_w^T) = E(W\mathbf{t}'[W\mathbf{t}']^T) \quad (1)$$

$$= E(W\mathbf{t}'\mathbf{t}'^TW^T) = E(W[\mathbf{t}'\mathbf{t}'^T]W^T) \quad (2)$$

Setting $S = E(\mathbf{t}'\mathbf{t}'^T)$, we observe that $W = S^{\frac{1}{2}}$ fulfils the terms of the constraint, the last term in the equation becoming $S^{\frac{1}{2}}SS^{\frac{1}{2}} (= I)$. Having found a suitable W , it only remains to perform a rotation of the whitened pattern-space axes such that the non-Gaussianity of the probability distribution of the individual variates of the transformed space is maximised (linear mixtures of variates being invariably more Gaussian than their components via the central limit theorem). This is usually achieved via an appropriate statistical, information theoretic or morphological measure of non-Gaussianity, and any one of a number of algorithms for finding global maxima/minima.

2.2 Experimental Noise in the Context of Independent Component Analysis

A necessary preliminary for the detailing of our technique is to consider the manner in which experimental noise relates to the above processes. Perhaps the paradigmatic illustration of the application of ICA techniques in this context is the ‘blind’ discrimination of spatially-distinct sound sources by multiple

randomly-placed microphones. Within the feature-space (technically a Hilbert space) defined by the simultaneous composition of the microphone outputs, each of the various signal sources has associated with it a characteristic vector, the orientation of which is determined by their placement relative to the various microphones. This is not an exact replication of the physical geometry of the situation, since each sound source's amplitude is attenuated by the inverse-square of the distance between itself and the microphone in question; however, it is a linear transformation thereof. Critically, when more than one source is considered, the individual fixed-orientation vectors are combined via linear addition of the components (it is this quality, along with the presence of an inner product term [to be introduced later], that defines the feature-space as a Hilbert space). That is:

$$\mathbf{S} = S^1 \hat{S}_1 + S^2 \hat{S}_2 + \dots + S^s \hat{S}_s \quad (3)$$

denotes the total source feature vector, with the $\hat{S}_i$ unit vectors denoting individual signal source 'orientations'. The coefficients S^i vary over the range $[0 : S_{\max}^i]$, defined by the maximum signal output and source/microphone geometry.

Since the s sound sources are completely independent, the *noiseless* signal vector consequently describes a uniform Euclidean subspace $\mathcal{S}$ within the original feature-space, defined by tensor multiplication of the various source vectors:

$$\mathcal{S} = S_{\max}^1 \hat{S}_1 \otimes S_{\max}^2 \hat{S}_2 \otimes \dots \otimes S_{\max}^s \hat{S}_s \quad (4)$$

(written henceforth as: $\mathcal{S} = \bigotimes_{i=1}^s S_i \hat{S}_i$). In our example, the origin of this subspace embedding is straightforwardly inferred from the geometry already inherent in the nature of the experimental situation, with the signal/microphone positioning characterising the unit vectors composing the feature subspace. This behaviour is very much more general, however, and precisely describes *any* situation in which features are composed of differing linear signal combinations. Crucially, it also approximates the very much wider class of situations in which the feature subspace *topology* is identical to that of a projective hyperplane.

We wish now to establish exactly how stochastic noise modifies this signal geometry. This can be encompassed within the above regime by considering white-noise emitting speakers placed at random locations within the experimental geometry. Each of the noise sources has again associated with it a characteristic orientation vector determined by its relative geometry, written $N^i \hat{N}_i$, where $\hat{N}_i$ is the noise source unit vector and N^i a uniform random variate distributed over the range $[0 : N_{\max}^i]$. These orientation vectors combine via tensor multiplication in the same manner as the signal vectors, increasing the dimensionality of the signal manifold $\mathcal{S}^{\mathcal{N}}$ by the number of independent noise sources, n :

$$\mathcal{S}^{\mathcal{N}} = \left(\bigotimes_{i=1}^s S_i \hat{S}_i \right) \otimes \left(\bigotimes_{i=1}^n N_i \hat{N}_i \right) \quad (5)$$

The actual, measured feature vector $\mathbf{V}$ is thus composed of the vector sum; $\mathbf{S} + \mathbf{N}$, the latter term denoting the pure noise vector.

Applying ICA-transformation to the *distribution* over $\mathcal{S}^{\mathcal{N}}$ performs an optimal coordinatisation of the manifold, in the sense of generating a hyper-space over which state vectors are uniformly distributed, and entirely eliminating the vector magnitude coefficients:

$$\mathcal{S}^{\mathcal{N}'} = \left(\bigotimes_{i=1}^s \hat{S}_i' \right) \otimes \left(\bigotimes_{i=1}^n \hat{N}_i' \right) \quad (6)$$

Here the $\hat{S}_i', \hat{N}_i'$ are *orthogonal* unit vectors, which is to say they obey the constraints $\hat{S}_i \cdot \hat{S}_j = 0$, $\hat{S}_i \cdot \hat{N}_j = 0$, $\hat{N}_i \cdot \hat{N}_j = 0 \ \forall i \neq j$: dashed terms will in general indicate ICA-transformed quantities throughout the paper.

Hence we see that signal and noise sources, being independent at both the collective and individual levels, are orthogonalised with respect to each other via ICA-transformation. Under these circumstances it is possible to *selectively* eliminate individual signal and noise features from the space. In particular, it becomes possible to perform feature-selection on the transformed space via a classification-performance based criterion such as forward searching [5]. Since the noise dimensions can only degrade classification performance, we would naturally expect the terminating output of the feature selection procedure to be the subspace $\mathcal{S}' = \hat{S}_1' \otimes \hat{S}_2' \otimes \dots \otimes \hat{S}_s'$, representing perfect noise elimination. This immediately implies a projective mapping of observational states:

$$\bigotimes_{i=1}^s \hat{N}_i' \hat{N}_1' \rightarrow \mathbf{1} \quad (7)$$

In fact there is a very much more inclusive projection than this (in the sense of a mapping of states beyond those present in the operational manifold) implied by the feature-selection from f_n -dimensions to s -dimensions, given by the orthogonal complement space mapping:

$$\left(\bigotimes_{i=1}^s S_{\max}^i \hat{S}_i \right)^{\perp} \rightarrow \bigotimes_{i=1}^s S_{\max}^i \hat{S}_i \quad (8)$$

This does not represent a difference in state mappings from equation 7 for the training data, but does so for test-data with noise characteristics differing from the training set (caused, perhaps, by insufficiently large training sets). From the point of view of the construction of hidden Markov models, however, we shall consider only the identified noise features in defining the emission characteristics of the hidden states in order to give precise statistical descriptions.

It is thus our central contention that the many-to-one mapping of the observed states to signal states represented by equation 7, without which observational state transitions would be partially stochastic, represents (in reverse) the emission spectrum for the 'hidden' state transitions that take place within the embedded signal space, $S_{\max}^1 \hat{S}_1 \otimes S_{\max}^2 \hat{S}_2 \otimes \dots \otimes S_{\max}^s \hat{S}_s$. That is, critically, the various transitions that take place between the noise vector component of the total observed feature vector $\mathbf{S} + \mathbf{N}$, ie $\mathbf{N}^1 \rightarrow \mathbf{N}^2 \rightarrow \dots$ for times $T = 1, 2, 3$, are

of a purely random nature, and may be modelled by an emission probability density over states determined by the signal vector $\mathbf{S}$. In essence, by removing ICA noise *components* in the feature-selection, we have thus replaced the particular mapping of observed-to-signal-subspace states present in the training-set by a more general mapping that can only be interpreted statistically, which is to say via a *hidden* Markov process of random emission of observation states.

In more conventional Markovian terminology [cf eg 2] a Markov chain of observed state transition coefficients, $a_{(\mathbf{S}+\mathbf{N})^t(\mathbf{S}+\mathbf{N})^{t+1}}$, is here transformed to an HMM with *hidden* state transition coefficients;

$$a'_{(\mathbf{S})^t(\mathbf{S})^{t+1}} = \int_{\forall \mathbf{N}_1} \int_{\forall \mathbf{N}_2} a_{(\mathbf{S}^t+\mathbf{N}_1)(\mathbf{S}^{t+1}+\mathbf{N}_2)} d\mathbf{N}_1 d\mathbf{N}_2 \quad (9)$$

and state emission probabilities:

$$\begin{cases} b_{\mathbf{S}^t}(\mathbf{V}) = 1/|\bigotimes_{i=1}^n N_{\max}^i \hat{N}_i| \\ \text{if } \mathbf{V} - \mathbf{S}^t \in \{\bigotimes_{i=1}^n N_{\max}^i \hat{N}_i\}; N^i \in \mathbb{R} \end{cases}$$

$$\begin{cases} b_{\mathbf{S}^t}(\mathbf{V}) = 0 \\ \text{if } \mathbf{V} - \mathbf{S}^t \notin \{\bigotimes_{i=1}^n N_{\max}^i \hat{N}_i\}; N^i \in \mathbb{R} \end{cases}$$

We will demonstrate in the next section that when the observed features are considered as separate data streams (which is to say, when the vector decomposition $\mathbf{V} \rightarrow \{\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3 \dots\}$ takes place), the above emission spectrum defaults exactly to the Gaussian emission spectra typically assumed by conventional hidden Markov model parametrisation techniques, and hence that the above method serves as a Gaussian mixture model parameteriser under these circumstances.

We reiterate that while the above analysis deals with an idealised noise model, to the extent that the underlying conditions of additive linear noise hold, substantial benefit may still be gained by application of the technique in the fully general case. Indeed, empirical testing on data for which no *a priori* guarantee of the validity of the model assumptions can be made (to be published at a later date), nonetheless yields very substantial performance improvement on application of the ICA+FS technique.

2.3 Relation to Gaussian Mixture-Model Parameterisation Methods

It is frequently found in the experimental domain that the total feature-space is naturally divided into separate subspaces (for instance, division of the physical configuration-space into position and velocity vectors). We will aim to demonstrate in this subsection that, under these circumstances, our methodology approximates that of the Gaussian mixture-model parameterisation methods (such as Baum-Welch re-estimation). In carrying out this exercise, it should be emphasised that the generality of our method means that optimal results are always to be obtained by generating a composite feature-space (ie, one in which the contributing subspaces are appended) prior to application of the ICA+FS method:

however it is reassuring to note that imposing this limitation reproduces familiar, empirically-validated methodologies.

We shall initially consider only a single dimensional marginal projection for simplicity of demonstration, later providing an argument for its generalisation to higher dimensionality feature space projections (such as those required to specify independent position and velocity vectors). A more general indication of the tendency of oblique projections of uniform multivariate distributions to approximate Gaussian distributions is given explicit proof in [6].

Firstly, let the chosen marginal feature, D_c , be characterised by its unit vector, $\hat{D}_c$, in the composite observed space. Measurements of observed states (defined by the projection $\hat{D}_c \cdot \mathbf{V}$) are thus characterised, for a given signal vector $\mathbf{S}_g$, by the sum of the individual random variates $\hat{D}_c \cdot N^i \hat{N}_i$ and the signal vector marginal projection, $\hat{D}_c \cdot \mathbf{S}_g$. That is to say:

$$\hat{D}_c \cdot \mathbf{V} = \sum_{i=1}^{i=n} \hat{D}_c \cdot N^i \hat{N}_i + \hat{D}_c \cdot \mathbf{S}_g$$

These random noise variates are uniformly distributed over the range $[\hat{D}_c \cdot \mathbf{S}_g : \hat{D}_c \cdot \mathbf{S}_g + \hat{D}_c \cdot \hat{N}_i N_{\max}^i]$, and hence have a mean $\hat{D}_c \cdot \mathbf{S}_g + \hat{D}_c \cdot \hat{N}_i N_{\max}^i / 2$ and variance $(\hat{D}_c \cdot \hat{N}_i N_{\max}^i)^2 / 12$. Now, according to the central limit theorem, the composite variate:

$$X = \frac{\sum_{i=1}^n x_i - \sum_{i=1}^n \mu_i}{\sum_{i=1}^n \sigma_i^2} \quad (10)$$

with *individual* means, μ_i , and variances, σ_i , obeys a Gaussian distribution with mean 0 and variance 1 in the limit $n > \infty$, on the proviso that a number of constraints on the individual variate probability density functions are observed, all of which are fulfilled by the uniform distribution.

Comparing this to the variate sum $\hat{D}_c \cdot \mathbf{V}$ above, it is clear that we only require a coordinate transform of:

$$\hat{D}_c \cdot \mathbf{V} \rightarrow \frac{(\hat{D}_c \cdot \mathbf{V} - \sum_{i=1}^n \hat{D}_c \cdot \hat{N}_i N_{\max}^i / 2 - \hat{D}_c \cdot \mathbf{S}_g)}{\sum_{i=1}^n (\hat{D}_c \cdot \hat{N}_i N_{\max}^i)^2 / 12}$$

to bring about an equivalence. Hence, our marginal variate $\hat{D}_c \cdot \mathbf{V}$ becomes asymptotically normally-distributed with mean $\hat{D}_c \cdot \mathbf{S}_g + \sum_{i=1}^n \hat{D}_c \cdot \hat{N}_i N_{\max}^i / 2$ and variance $\sum_{i=1}^n (\hat{D}_c \cdot \hat{N}_i N_{\max}^i)^2 / 12$, obeying the overall formula:

$$P(\hat{D}_c \cdot \mathbf{V} | \mathbf{S}_g) = \frac{e^{\left[\frac{(\hat{D}_c \cdot \mathbf{V} - (\hat{D}_c \cdot \mathbf{S}_g + \sum_{i=1}^n \hat{D}_c \cdot \hat{N}_i N_{\max}^i / 2))^2}{2(\sum_{i=1}^n (\hat{D}_c \cdot \hat{N}_i N_{\max}^i)^2 / 12)} \right]}}{\left[2 \sum_{i=1}^n (\hat{D}_c \cdot \hat{N}_i N_{\max}^i)^2 / 12 \pi \right]^{\frac{1}{2}}} \quad (11)$$

Thus our mapping of the marginal observational states onto the ‘hidden’ states of the ICA+FS transformed *total* observational configuration space implies the existence of a hidden Markovian model of Gaussian observed state emission

characteristics (of the kind familiar from, in particular, automated speech recognition). It should be emphasised that this form is in no way imposed *a priori*; the relationship between, and distribution of, hidden and observed states is determined purely by the output of the ICA+FS procedure.

The f_n -dimensional generalisation of the above argument for single-dimensional marginal projections is complicated by the issue of covariance. However, it is possible to see that there must always exist a coordinate transformation of the f_n -dimensional space such that each of the noise source vectors $N^i \hat{N}_i$ supplies an equivalent aggregate contribution to the signal vector: that is, a transformation such that unit vectors in the transformed space, $\hat{u}_i$, satisfy the constraint: $\langle \hat{u}_i, \hat{N}_j \rangle = \text{constant} \quad \forall i, j : 1 \leq i \leq f_n, 1 \leq j \leq n$. Full details of the derivation of this transformation, T , are set out in [7]: critically, the coordinate transformation is both linear and invertible.

That the signal-vector distribution in the transformed space is now composed of normalised uniform distributions again implies, via the central limit theorem, that they collectively approximate symmetric Gaussian distributions of variance: $1/(\sqrt{2^{f_n-1}}12)$ for each marginal distribution of the transformed space. Since these marginal distributions are independent of each other, we can take their tensor product in order to recover the distribution over the full transformed space. Consequently, since the tensor product of Gaussian distributions is itself Gaussian, we find that the overall distribution is a covariance-less $(F - n)$ -dimensional Gaussian distribution over the vectorial range of the transformed space. The bijective linear transformation back into the original space, T^{-1} , maintains this Gaussian form, but generates a unique covariance matrix Σ defined by the various shear and rotation deformations implicit in T^{-1} ¹.

It is hence this potential to constrain unique Σ and μ values that constitutes the effective Gaussian parameterisation of the ICA+FS method in restricted spaces, and brings it in to an operational equivalence with the Gaussian mixture-model estimators such as Baum-Welch re-estimation, which typically [cf eg. 2] assume models of the form:

$$b_j(\mathbf{o}_t) = \prod_{w=1}^W \left[\sum_{m=1}^{M_w} c_{jwm} \mathcal{N}(\mathbf{o}_{wt}; \mu_{jwm}, \Sigma_{jwm}) \right] \quad (12)$$

with;

$$\mathcal{N}(\mathbf{o}; \mu, \Sigma) = \frac{1}{((2\pi)^n |\Sigma|)^{\frac{1}{2}}} e^{\frac{1}{2}(\mathbf{o}-\mu)' \Sigma^{-1}(\mathbf{o}-\mu)} \quad (13)$$

(M is the number of Gaussians in the mixture, W the number of distinct observational feature-spaces, and $\mathbf{o}$ the within-stream observational vector).

¹ It is additionally the case that multiple Gaussians may be formed by this type of marginal projection where there exist distinct signal regions within the observational space, as might be obtained by regional application of the ICA+FS procedure. The key requirement for Gaussianisation of the emission states in either scenario is that there exists a *projective* mapping from a higher to a lower dimensional space.

3 Conclusions

We have set out a technique for automatically obtaining parameterised hidden Markov models from observed feature transition sequences through a process of feature-selection in an ICA-transformed space, on the assumption of the presence of independent noise sources within the data. This involves the mapping of observed states onto a space of hidden states induced by the orthogonalisation and removal of the noise sources within the observed features. The stochasticity of this mapping accounts for the generated entity being a *hidden*, rather than a simply Markovian model, and thus allows a much greater generalising capacity than would a simple Markovian description of the observed transitions.

When an artificial distinction between data streams (such as position and velocity) is assumed, or when the observed feature-space represents an incomplete description of the free experimental parameters, the observed state distribution in the HMM models generated by the ICA+FS technique was shown to default to a Gaussian emission model. The proposed methodology under these circumstances thus reproduces the familiar Gaussian mixture model parameterises, without, however, having to assuming this form on an *a priori* basis. The technique is thus very much more general than existing HMM parameterisation techniques, having fewer underlying assumptions, and can be treated as first recourse when the nature of the emission state/hidden state transition relationships are unknown. Empirical studies suggest a consistent 50% performance improvement over equivalent ‘naive’ Markovian methods.

Acknowledgements

This paper gratefully acknowledges the financial support of INRIA within the context of the Cognitive Vision Systems framework collaboration.

References

1. L. Baum and J. Eagon, American Mathematical Society Bulletin, 73:360-363, 1967.
2. Steve Young, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland. The HTK Book for HTK Version 2.1. Cambridge University, 1997.
3. A. Hyvärinen, E. Oja, “Independent Component Analysis: Algorithms and Applications”, Neural Networks, 13(4-5):411-430, 2000
4. P. Comon, “Independent component analysis, A new concept?,” IEEE Signal Processing Mag., vol. 36, no. 3, 1994.
5. P.A. Devijver and J. Kittler, “Pattern Recognition: A Statistical Approach”, Prentice-Hall, 1982
6. Asymptotics of graphical projection pursuit, P. Diaconis & D. Freedman, Ann. Stat., vol. 12, 793-815
7. D. Windridge, R. Bowden, J. Kittler, (Univ. of Surrey Technical Report: VSSP-TR-3/2004), UNIS, UK