# Computational Creativity by Structural Analogy between Acoustic Bayesian Models

Michael A. Casey
*Centre for Computational Creativity
Department of Computing
City University, London
casey@soi.city.ac.uk

**Abstract**

We present a method for generative modeling of audio content that performs mappings between minimum entropy hidden Markov models learnt from audio data. By training with a minimum entropy prior, compact, low-complexity models of the latent structure in audio source samples are obtained. Synthesis of new audio content is achieved by mapping the state sequence of a nominated *structure* model onto a nominated *content* model. The mapping is chosen such that the cross-entropies between the state variables of the nominated models are minimised. This creates an analogy between the models' structures, even when the specific content of the models varies significantly. The re-mapped content state sequences are inverted to yield spectral vectors that consist of the higher-order pattern information of the structure model and the low-order spectral features of the content model. To illustrate the methods, we present examples of mapping audio structure and content between drum beat samples in different styles.

## 1 Introduction

Audio content repurposing has a significant presence in musical culture. Much of the electronic music currently being produced is based on transforming sampled materials in some way, and there exist a number of commonly-used signal processing approaches for such transformations. Notable amongst them is the *phase vocoder*, (Portnoff, 1976), which uses the short-time Fourier transform for analysis and modified resynthesis of spectral data and has been used by composers for creating new acoustic content by spectral blending of two or more source sounds. Phase vocoding is a remarkably effective tool that enables a degree of independence between temporal and spectral transformations not possible with looping and resampling techniques. However, as with most signal processing transformations, it is agnostic to any patterns that exist in the audio data, thus the transformations are structurally naive. In this paper, we describe a structured cross-synthesis technique that maps higher-order structural features between audio samples; with structure determined automatically using minimum entropy learning methods for Bayesian models.

Hidden Markov models are widely used for acoustic modeling in speech and for classification of audio, (Rabiner, 1989) (Casey, 2002). This is due, in large part, to the Baum-Welch algorithm for maximum likelihood inference of model parameters. However, little work has been undertaken on using HMMs to synthesise audio content outside of the speech synthesis literature, (Yoshimuray and Tokuday, 1998). We use HMMs to extract latent structure from audio signals and use this structure for generating novel audio content that exhibits a blend of salient features from two source samples.

## 2 Structure Discovery HMMs

A large number of algorithms, and heuristics, exist for fitting models to data. But one must normally specify the structure of the model; for example, the number of states to use and the linkage of the transition matrix for hidden Markov models. To automatically learn such structure from features, we use a minimum entropy prior on the form of the internal variables for an HMM. This strategy combines the problem of model structure estimation with the problem of optimal parameter estimation from data and is solved using maximum *a-posteriori* (MAP) estimation. A hidden Markov model consists of multinomital parameters representing an initial state distribution, $\pi_i = P(s_1 = i)$ and between-state transition probabilities, $T_{ij} = P(s_{t+1} = i|s_t = j)$, for states $i, j \in \{1, \ldots, K\}$ and time $t \in \{1, \ldots, T\}$. States are parameterised by multidimensional Gaussian probability distributions over the space of observations. To compute the entropy of an HMM, the model parameters are concatenated into a stochastic vector $\theta = [\theta_1, \ldots, \theta_N]$, thus creating a multinomial over the model's parameter space. The conditional *posterior* distribution , with respect to the model parameters $\theta$ and observable evidence $\mathbf{x}$, can be factored according to Bayes' rule:

$$P(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta)P_e(\theta) \quad (1)$$

with entropic prior $P_e(\theta)$ of the form:

$$P_e(\theta) \propto \exp^{-H(\theta)} = \exp\big[\sum_{i=1}^{N} \theta_i \log \theta_i\big] \qquad (2)$$

Brand (1999) provides a method for maximum *a-posteriori* (MAP) estimation using the entropic prior, in Equation 2, that yields models that are biased towards representing the structure, sparsity and determinism inherent in the training data. Figure 1 shows a minimum entropy HMM of a 16-bar Latin percussion sample. Only the first two dimensions of the probability space are shown.

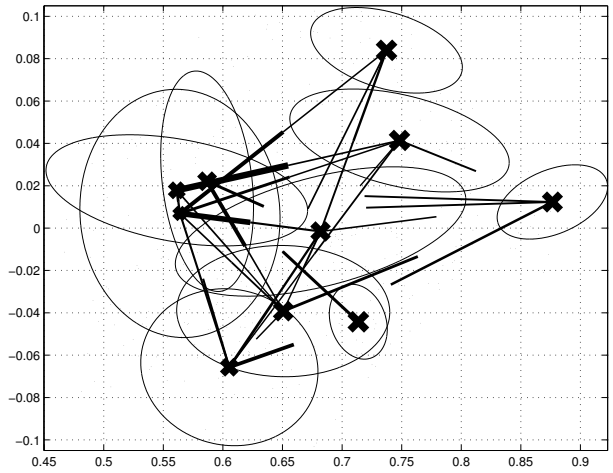

Figure 1: A 10-state hidden Markov model with sparse transition structure. Elipses represent Gaussian probability isocontours and lines represent transition probabilities between states. The size of 'X'in each state denotes the probability of self-transition.

Figure 2 illustrates the structure that is represented by the HMM state variables. Noisy low-level spectral audio features are converted into a representation that encodes the higher-order event structure in the data. It is easy to 'pick out' the states that correspond to onset and sustain phases of the percussive timbres in the figure.

## 2.1 Audio Feature Extraction

Log spectral features were extracted according to the specifications of the MPEG-7 low-level audio descriptors standard, see ISO (2001). Below, we also extract linear basis functions and low-dimensional projection coefficients using the MPEG-7 specification. Thus the feature-extraction and acoustic Markov modeling methods presented herein constitute an application of the MPEG-7 standard to structured content repurposing. We briefly describe these feature extraction methods in this section.

We recorded a set of 20 audio samples from a high-quality drum machine. Each sample consisted of a unique drum pattern in one of two styles (Latin or Techno) with a tempo of 75bpm and 16 bars duration in $\frac{4}{4}$ metre. Each sample was therefore 1280 spectral frames in duration and

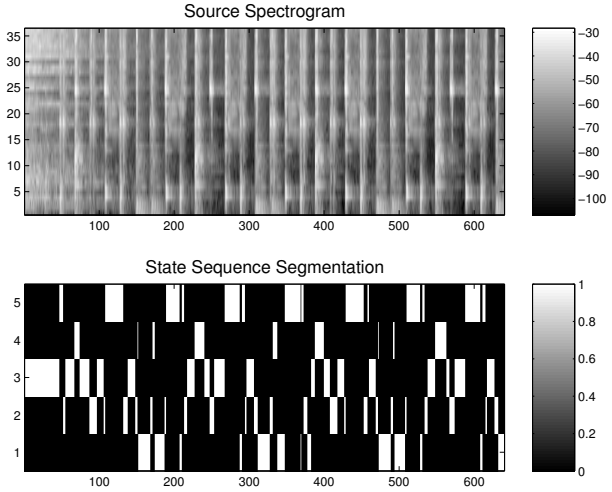


Figure 2: (Top) A constant-Q spectrogram of a Latin percussive sample lasting 6.4 seconds (10ms frames). (Bottom) Segmentation of the sample into 5 discrete states by a content-specific hidden Markov model.

the ensemble of training samples consisted of a total of 256 seconds of audio.

The signals were reduced to 16kHz mono 16-bit linear PCM samples and segmented by 30ms hamming windows, with a 10ms advance per frame, and transformed to the spectral domain using a 512-point FFT. The FFT elements were then converted to power spectrum coefficients and re-binned into $\frac{1}{8}$th-octave bands, taking care to preserve the total power. The resulting log-spectral frames were 39-dimensional and preserved the power of the signal over the analysis window. These frames were stacked row-wise into observation matrices, $\mathbf{X}$, with the ensemble of stacked matrices representing all source samples, $\{\mathbf{X}\}$. Vectors were re-scaled to decibels and normalised by their L2-norm coefficients to yield unit-length spectral-shape vectors. The L2-norm coefficients were pre-pended to the ensemble vectors to create 40-dimensional vectors consisting of the RMS of spectral power coefficients in the first element and normalised spectral shape vectors in the remaining elements.

## 2.2 Audio Probability Space

To cross-synthesise audio data from multiple HMMs, one must ensure that the parameters of each model are defined over a common probability space. To do this, we first extracted short-time spectral features from each source audio sample, then we computed a linear basis over the spectral space.

The spectral linear basis was extracted by computing the singular value decomposition (SVD) over the ensemble of observation matrices, $\{\mathbf{X}\}$. The first five basis functions, $\mathbf{V}$, were retained and used to project observations into a low-dimensionality probability space:

$$\mathbf{Y} = \mathbf{X}\mathbf{V} \tag{3}$$

This approach to audio feature extraction is described in detail, along with quantitative test experiments, in Casey (2002) and ISO (2001).

# 3 Content Repurposing

## 3.1 Structure, Content and Style

Methods for mapping between Bayesian models for style and content repurposing have previously been proposed in the fields of machine vision and computer graphics, (Brand, 2000) (Freeman and Tenenbaum, 1997). Brand's method uses learning to capture structure and style in bipedal motion performed by human subjects. Style hidden Markov models are used to synthesize new motion sequences by interpolating parameters between sample-specific models thus creating novel animations from existing content. Freeman and Tenenbaum (1997) use bi-linear models to separate style and content in images thereby enabling independent control over stylistic aspects of the data, such as the effects of lighting and pose in images of faces. There has also been some work in the area of speech synthesis on controlling HMM state variables for expressive articulation, (Yoshimuray and Tokuday, 1998).

We adopted a similar framework for music and general audio repurposing. As such, this represents a departure from usual audio signal processing practices in that higher order structure is considered along in addition to the low-order spectral structure.

## 3.2 Structure Mapping

Since the models were constrained to exist in a compatible probability space, by projection onto the linear basis $V$, the state parameters in different models could, meaningfully, be computed. Starting with two content-secific models trained on audio samples, we nominate one as the structure model, $S$, and the other as the content model, $C$. The best structural analogy was selected to be the map that minimised the sum of cross-entropies between all pairs of between-model states:

$$M(i) = \arg\min_j H_X(p_i; q_j) \tag{4}$$

The cross entropy, $H_X(p;q)$, between probability distributions $p(x)$ and $q(x)$ over $\mathbf{X}$ is defined as:

$$\begin{aligned} H_X(p;q) &= H(X) + D(p||q) \\ &= -\sum_{x \in X} p(x) \log(q(x)) \end{aligned}$$

with entropy $H(\mathbf{X})$ and relative entropy $D(p||q)$. Given two Gaussian states, $p_i$ in model $S$ and $q_j$ in model $C$, with mean and covariance parameters $(\mathbf{m}_1, \mathbf{K}_1)$ and $(\mathbf{m}_2, \mathbf{K}_2)$ respectively, and dimensionality $d$, the cross entropy calculated by:

$$\begin{aligned} H_X(p_i; q_j) &= \frac{1}{2}[d \log 2\pi e + \log |\mathbf{K}_i|] \\ &+ \frac{1}{2}[\log |\mathbf{K}_i| - \log |\mathbf{K}_j| \\ &+ \sum_{mn} ({\mathbf{K}_i}^{-1})_{mn} ({\mathbf{K}_j}^{-1})_{mn} \\ &+ (\mathbf{m}_i - \mathbf{m}_j)^T {\mathbf{K}_i}^{-1} (\mathbf{m}_i - \mathbf{m}_j) - d] \end{aligned}$$

We define a map, $M : S \mapsto C$, using Equation 4, and denote it using the functional $c_t = M(s_t)$ for states $c$ in $C$ and $s$ in $S$ with time index $t$. Applying the map to the structure state sequence, $(s_1, s_2, \ldots, s_T)$, returns the new content sequence, $(c_1^\star, c_2^\star, \ldots, c_T^\star)$, which has the higher-order patterning of the structure model but the low-level spectral features of the content model.

To maximize the structural analogy between models we initialized the training process for models $C$ and $S$ with a generic model, $G$, learnt from the ensemble observation probability space, $P(\{\mathbf{X}\})$, and minimized the cross-entropies between corresponding states in each model during MAP estimation. Therefore the models were constrained to exhibit as much structural similarity as possible whilst capturing the specific content of the individual samples from which they were trained.

Figure 3 illustrates structural analogy using the minimum cross-entropy map between two models of drum beat patterns. The top model was trained on a *Latin* style sample and the bottom was trained on *Techno* style sample. Using such maps we computed new audio sequences, by mapping the states of one model onto the states of the other.

## 3.3 Model Inversion

To invert the new state sequence, $(c_1^\star, c_2^\star, \ldots, c_T^\star)$, we used the means of the Gaussian states in $C$. Whilst this is clearly a simplication, constraining the emitted observations to visit only the centre of each state, it is sufficient for generating an approximation to the inverse of an HMM given a state sequence. New sequences of spectral acoustic vectors were generated by first projecting the mean vector of each state onto the inverse linear basis, $V^T$:

$$\mathbf{x}^\star = \mathbf{m}_{c^\star} V^T \tag{5}$$

then taking the first RMS power element from $\mathbf{x}^\star$, multiplying it with the remaining elements, and inverse decibel scaling. The resulting log-frequency power spectrum coefficients were used to drive a filterbank for synthesis of an acoustic waveform.

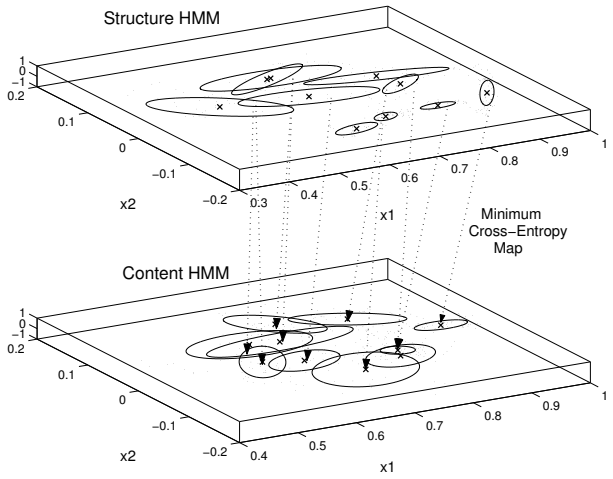


Figure 3: Structural analogy between two sample-specific models of a *Latin* drum sample (top) and a *Techno* drum sample (bottom). The arrows between-models depict the minimum cross-entropy mapping between the Gaussian states of each model.

Figures 4, 5 and 6 show the results of structure mapping and model inversion for combinations of drum patterns in different styles. In each case, two source samples were used to generate new hybrid state sequences. The entire process was automatic so the results reflect the ability of structure discovery methods to find salient higher-order structure in acoustic data without supervision.

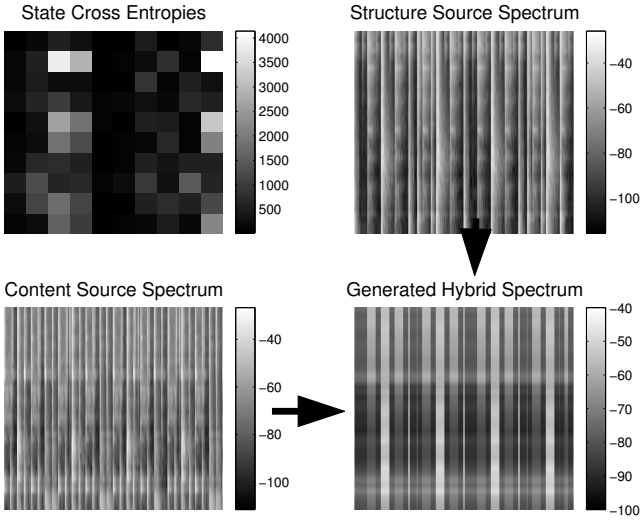


Figure 4: Generation of new structured audio content by remapping structure between source models (Top L) Cross entropy map of between-model states (Top R) spectrum of structure source [Latin:1] (Bottom L) spectrum of content source [Techno:1] (Bottom R) new spectrum.

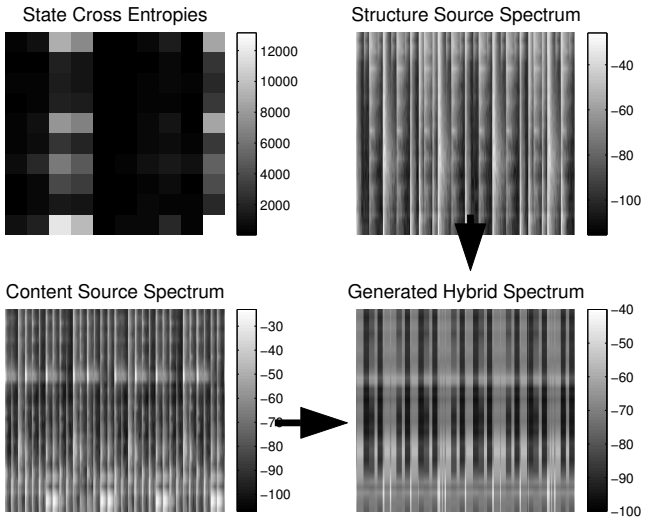


Figure 5: (Top L) Cross entropy map of between-model states (Top R) spectrum of structure source [Latin:2] (Bottom L) spectrum of content source [Techno:1] (Bottom R) new spectrum.

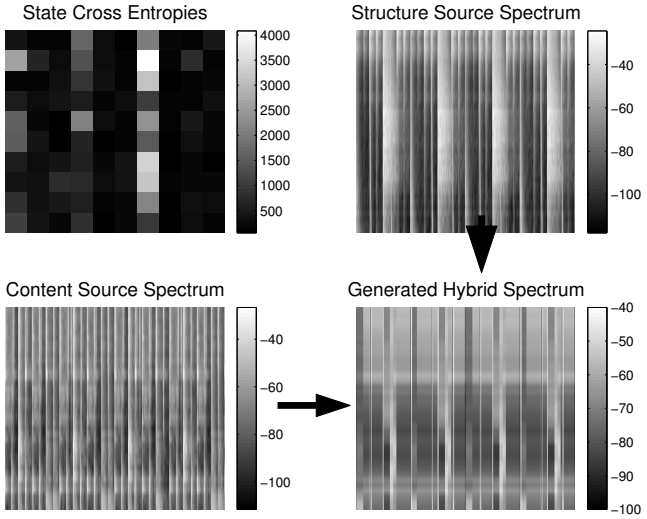


Figure 6: (Top L) Cross entropy map of between-model states (Top R) spectrum of structure source [Latin:1] (Bottom L) spectrum of content source [Techno:2] (Bottom R) new spectrum.

# 4 Summary

To summarise the method of content repurposing by structural analogy we enumerate the steps for training, mapping and inverting sample models.

1. Extract audio features from an ensemble of source samples, generating observation matrices $\{\mathbf{X}\}$.

2. Extract an ensemble linear basis,$V$, by singular value decomposition of $\{\mathbf{X}\}$.

3. Train a hidden Markov model on the ensemble observation matrix to produce a generic model $G$.

4. Train an ensemble of sample-specific hidden

Markov models, initialised to $G$, on individual observation matrices $\mathbf{X_j}$. Optionally constrain the models to have minimum cross-entropies between model states after each training epoch. This step produces a set of source models $\{\mathbf{S}\}$.

5. Nominate a structure source model and a content source model, $S, C \in \{\mathbf{S}\}$. Then, map the state trajectories for model $S$ onto the model $C$ by finding the minimum cross-entropies between model states thereby creating a structural analogy. This step creates a new state sequence in $C$, see Equation 4.

6. Invert the new state sequence with respect to $C$ to obtain a series of observation vectors and invert the observation vectors to generate a series of spectral frames. Invert the spectral frames via filterbank resynthesis to produce an audio signal.

# 5 Conclusion

In this paper we presented new techniques for automatically discovering structure in audio sample data using learning in Bayesian models with minimum entropy priors, and for repurposing such structure to create novel audio content. The methods were shown to capture salient event and pattern information in complex audio data. Structural analogy mapping between content-specific models, using the minimum cross-entropy between states, was introduced as means for generating new state sequences from trained models. Methods for inverting such sequences, and examples of novel content generation, were also presented.

We conclude that the methods presented herein are a good starting point for exploration of creativity by Bayesian inference. Whilst the methods produced good results for representing macrostructure, the current model inversion methods could be improved for generating spectral vector sequences with morphologies that better represent the micro-temporal structures in the training data. One such extension is to use a Bayesian method that maps smooth state trajectories in a structure model onto smooth state trajectories in a content model by learning the dynamic between-model relationships from pairs of training examples. Due to the additional complexity this has not yet been implemented, but will be the subject of future extensions to our methods.

# References


M. Brand. Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, 11(5):1155–1182, 1999.

M. Brand. Style machines. In *Proceedings of SIGGRAPH*, New Orleans, Louisiana, USA, 2000.

M. Casey. *Introduction to MPEG-7: Multimedia Content Description Language*, chapter Sound Classification and Similarity Tools. J. Wiley, NY, 2002.

W. Freeman and J. Tenenbaum. Learning bilinear models for two-factor problems in vision. In *Proceedings, Conf. on Computer Vision and Pattern Recognition*, San Juan, 1997.

ISO. *ISO 15938-4:2001 (MPEG-7: Multimedia Content Description Interface, Part 4:Audio)*. ISO, 2001.

M.R. Portnoff. Implementation of the digital phase vocoder using the fast fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24:243–248, 1976.

L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

T. Yoshimuray and K. Tokuday. Duration modeling for hmm-based speech synthesis. In *Proceedings of the International Conference on Speech and Language Processing*, Australia, 1998.