AUTOMATIC MUSIC TAGGING WITH TIME SERIES MODELS

Emanuele Coviello	Luke Barrington	Antoni B. Chan	Gert. R. G. Lanckriet
University of California, San Diego	University of California, San Diego	City University of Hong Kong	University of California, San Diego
Dept. of Electrical and Computer Engineering	Dept. of Electrical and Computer Engineering	Dept. of Computer Science	Dept. of Electrical and Computer Engineering
ecoviell@ucsd.edu	lukeinusa@gmail.com	abchan@cityu.edu.hk	gert@ece.ucsd.edu

ABSTRACT

We present a system for automatic music annotation that leverages temporal (e.g., rhythmical) aspects as well as timbral content. Our system estimates a dynamic texture mixture (DTM) density over times series of acoustic features (instead of on individual features) for each tag in a semantic vocabulary. When analyzing a new song, our system processes the time series of acoustic features of the song and outputs a semantic multinomial, i.e., a vector of tag-affinities. A song is then annotated by selecting the top-ranking tags in its semantic multinomial. Tag-DTM models are estimated efficiently with the hierarchical EM algorithm for DTM (HEM-DTM) from all the DTMs modeling individual songs associated with a tag.

E. Coviello, L. Barrington, A. Chan, and G. Lanckriet, **"Automatic Music Tagging with Time Series Model"**, *ISMIR 2010*, Utrecht (Netherlands). 9 - 13 Aug. 2010

E. Coviello, A. Chan, and G. Lanckriet, **"Time Series Models for Semantic Music Annotation"**, *Transactions on Audio, Speech and Language Processing*, 19-5, pp 1343 - 1359 .

1. MODELING AUDIO AND TAGS

Our auto-tagging music information retrieval (MIR) system takes as input an audio track and computes the relevance of all the tags in a vocabulary to the audio track. The systems is based on the models in [3, 4] previously applied to video and audio retrieval.

2. MODELING SONGS

The acoustic content of each song in the collection is represented by computing a time series $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$ of 34-bin Mel-frequency spectral features extracted over half-overlapping windows of 92 ms of audio signal, where T depends on the length of the song. The Mel-frequency spectral features are grouped into fragments of $\tau = 125$ consecutive feature vectors (corresponding to approximately $6s$) with 80% overlap. The acoustic content of a song is hence represented by an unordered bag of audio fragments $\mathcal{Y} = \{\boldsymbol{y}^1_{1:\tau}, \ldots, \boldsymbol{y}^F_{1:\tau}\}$, where F depends on the length of the song.

The timbral content and temporal aspect of a single audio fragment $\boldsymbol{y}_{1:\tau}$ are captured by a dynamic texture (DT) component $\Theta = \{A, Q, C, R, \mu, S, \bar{y}\}$. The DT model consists of two random variables, y_t, which encodes the acoustic component (audio feature vector) at time t, and x_t, which encodes the dynamics (evolution) of the acoustic component over time. The two variables are modeled as a *linear dynamical system*,

$$x_t = Ax_{t-1} + v_t, \tag{1}$$
$$y_t = Cx_t + w_t + \bar{y}, \tag{2}$$

where $x_t \in \mathbb{R}^n$ and $y_t \in \mathbb{R}^m$ are real vectors (typically $n \ll m$). The matrix $A \in \mathbb{R}^{n\times n}$ is a *state transition matrix*, which encodes the dynamics or evolution of the hidden state variable (e.g., the evolution of the audio track), and the matrix $C \in \mathbb{R}^{m\times n}$ is an *observation matrix*, which encodes the basis functions for representing the audio sequence. The vector $\bar{y} \in \mathbb{R}^n$ is the mean of the dynamic texture (i.e. the mean audio feature vector). v_t is a *driving noise process*, and is zero-mean Gaussian distributed , e.g., $v_t \sim \mathcal{N}(0, Q)$, where $Q \in \mathbb{R}^{n\times n}$ is a covariance matrix. w_t is the *observation noise* and is also zero-mean Gaussian, e.g., $w_t \sim \mathcal{N}(0, R)$, where $R \in \mathbb{R}^{m\times m}$ is a covariance matrix. Finally, the *initial condition* is specified as $x_1 \sim \mathcal{N}(\mu, S)$, where $\mu \in \mathbb{R}^n$ is the mean of the initial state, and $S \in \mathbb{R}^{n\times n}$ is the covariance.

The timbral content and the temporal aspects of a song $\mathcal{Y}$ are modeled with a dynamic texture mixture (DTM) [2] probability density over audio fragments:

$$p(\boldsymbol{y}_{1:\tau}|\mathcal{Y}) = \sum_{s=1}^{K_s} a_s^{(\mathcal{Y})} p(\boldsymbol{y}_{1:\tau}|\Theta_s^{(\mathcal{Y})}) \,, \tag{3}$$

where K_s is the number of mixture components and $\Theta_s^{(\mathcal{Y})}$ is the s^{th} DT component. The parameters $\{a_s^{(\mathcal{Y})}, \Theta_s^{(\mathcal{Y})}\}_{r=1}^{K_s}$ are estimated based on the audio fragments extracted from the song, i.e., $\mathcal{Y} = \{\boldsymbol{y}^1_{1:\tau}, \ldots, \boldsymbol{y}^F_{1:\tau}\}$, using the EM algorithm for DTM [2]. Each dynamic texture (DT) component $\Theta_s = \{A_s, Q_s, C_s, R_s, \mu_s, S_s, \bar{y}_s\}$ model homogeneous, perceptually similar segments of the song (corresponding to what a human listener would label as chorus, verse, bridge, etc.) by capturing temporal as well as textural aspects of the musical signal [1].

3. MODELING TAGS

Tag models are learned from a database $\mathcal{D} = \{(\mathcal{Y}_d, \boldsymbol{c}_d)\}_{d=1}^{|\mathcal{D}|}$ of songs annotated with respect to a semantic vocabulary $\mathcal{V}$. The binary vector $\boldsymbol{c}_d = (c_{d,1}, \ldots, c_{d,|\mathcal{V}|})$ encodes the semantic content of the d^{th} song, with $c_{d,i} = 1$ only if there is a positive association between the song and the tag $w_i \in \mathcal{V}$.

The distribution for tag $w_i \in \mathcal{V}$ is modeled with a dynamic texture mixture (DTM) [2] probability density over sequences of audio feature vectors:

$$p(\boldsymbol{y}_{1:\tau}|w_i) = \sum_{r=1}^{K_t} a_r^{(w_i)} p(\boldsymbol{y}_{1:\tau}|\Theta_r^{(w_i)}) \,, \quad (4)$$

where K_t is the number of mixture components and $\Theta_r^{(w_i)}$ is the r^{th} DT component. The parameters $\{a_r^{(w_i)}, \Theta_r^{(w_i)}\}_{r=1}^{K_t}$ for the tag model for tag w_i are estimated from all the audio fragments extracted from the songs in $\mathcal{D}$ that are positively associated with tag w_i, i.e., $\{\mathcal{Y}_d|\, c_{d,i} = 1\}$.

Learning the tag-distribution directly on this data could be computationally inefficient. To allow efficient training in both computation time and memory requirements, we use the hierarchical EM algorithm for DTM (HEM-DTM) [3] to learn the tag-distribution directly from the song-level distributions $p(\boldsymbol{y}_{1:\tau}|\mathcal{Y}_d)$ associated to w_i

4. AUTOMATIC ANNOTATION

Given the audio fragments of a novel song $\mathcal{Y}$, the relevance of tag w_i is computed using Bayes' rule:

$$p(w_i|\mathcal{Y}) = \frac{p(\mathcal{Y}|w_i)\, p(w_i)}{p(\mathcal{Y})}, \quad (5)$$

where $p(w_i)$ is the prior of the i^{th} tag, and $p(\mathcal{Y}) = \sum_{i=1}^{|\mathcal{V}|} p(\mathcal{Y}|w_i)p(w_i)$ is the song prior. To promote annotation using a diverse set of tags, we assume an uniform prior, $p(w_i) = 1/|\mathcal{V}|$. The likelihood term is computes with the geometric average of the individual sequence likelihoods smoothed by the sequence length τ, i.e., $p(\mathcal{Y}|w_i) = \prod_{f=1}^{F} \left(p(\boldsymbol{y}_{1:\tau}^{f}|w_i)\right)^{\frac{1}{T\tau}}$.

Finally, the song can be represented as a semantic multinomial, $\mathbf{p} = [p_1, \ldots, p_{|\mathcal{V}|}]$, where each $p_i = p(w_i|\mathcal{Y})$ represents the relevance of the i^{th} tag for the song, and $\sum_{i=1}^{|\mathcal{V}|} p_i = 1$.

We annotate a song with the most likely tags according to $\mathbf{p}$, i.e., we select the tags with the largest probability. For retrieval, we define a song's relevance to the query tag w_i as the posterior probability of the tag, $p(w_i|\mathcal{Y})$. Hence, retrieval involves rank-ordering the songs in the database based on the i^{th} entry (p_i) of the semantic multinomials $\mathbf{p}$.

Acknowledgments E.C., L.B. and G.R.G.L. wish to acknowledge support from NSF grants DMS-MSPA 0625409 and CCF-0830535.

5. REFERENCES

[1] L. Barrington, A.B. Chan, and G. Lanckriet. Modeling music as a dynamic texture. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):602–612, 2010.

[2] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):909–926, 2008.

[3] A.B. Chan, E. Coviello, and G. Lanckriet. Clustering dynamic textures with the hierarchical EM algorithm. In *Proc. IEEE CVPR*, 2010.

[4] E. Coviello, A. Chan, and G. Lanckriet. Time Series Models for Semantic Music Annotation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(5):1343–1359, July 2011.