MULTI-INSTRUMENT DETECTION IN POLYPHONIC MUSIC USING GAUSSIAN MIXTURE BASED FACTORIAL HMM

Ranjani H. G., T. V. Sreenivas

Dept. of ECE., Indian Institute of Science, Bangalore-560 012, India
{ranjanihg , tvsree}@ece.iisc.ernet.in

ABSTRACT

We formulate the problem of detecting the constituent instruments in a polyphonic music piece as a joint decoding problem. From monophonic data, parametric Gaussian Mixture Hidden Markov Models (GM-HMM) are obtained for each instrument. We propose a method to use the above models in a factorial framework, termed as Factorial GM-HMM (F-GM-HMM). The states are jointly inferred to explain the evolution of each instrument in the mixture observation sequence. The dependencies are decoupled using variational inference technique. We show that the joint time evolution of all instruments' states can be captured using F-GM-HMM. We compare performance of proposed method with that of Student's-t mixture model (tMM) and GM-HMM in an existing latent variable framework. Experiments on two to five polyphony with 8 instrument models trained on the RWC dataset, tested on RWC and TRIOS datasets show that F-GM-HMM gives an advantage over the other considered models in segments containing co-occurring instruments.

***Index Terms*—** Factorial HMM, Latent Variable, Polyphony, F-GM-HMM

1. INTRODUCTION

Identifying multiple instruments present in polyphonic music is indispensable in content based Music Information Retrieval, with applications in separation [1–5] and detection [6–13]. Latent Variable (LV) formulations are used to achieve the task with non-parametric [2, 3, 8, 11, 14] or parametric models [15]. The LVs indicate presence/contribution of each of the instruments. A commonality in such probabilistic approaches is that likelihood of polyphonic signal observation vector is modeled in terms of individual instrument models. A generic approach for the same is proposed in [15] with focus on detecting underlying instruments.

Factorial models are a natural choice to analyze observations that are coupled due to simultaneous emissions from multiple, otherwise independent, generative processes. For this application, each generative process is modeled using Gaussian Mixture Hidden Markov Models (GM-HMM) to capture attack, sustain and release (A-S-R) sequence of a given instrument note. In this paper, we propose a factorial framework to jointly analyze these temporal evolutions of all instruments in a polyphonic signal. We refer to this as Factorial GM-HMM (F-GM-HMM). It is a generalization of a specific type of factorial HMM with a single Gaussian emission density per state (F-G-HMM), proposed in [16]. Variational inference technique is used to decouple dependencies of chains on observations to infer the joint time evolution of all instruments' states in addition to (an introduced) silence state from the polyphony given monophonic models. Further, we use F-GM-HMM in LV framework to detect instruments and evaluate its performance. The contributions of this paper are: 1) A generative F-GM-HMM model to explain A-S-R patterns of instruments in a polyphony, 2) Formulation for multiple instrument detection using F-GM-HMM model in a LV framework, 3) Generalization of F-G-HMM to F-GM-HMM, and insight into decoupling in LV-framework against that in factorial framework, and 4) Comparison of performance of F-GM-HMM with that of tMM (Student's-t mixture model) [15], & GM-HMM in a LV-framework on 8 instruments (clarinet, flute, guitar, harp, mandolin, piano, trombone & violin) from RWC database [22]. Relation to prior work is discussed in Section (5).

2. F-GM-HMM MODEL FOR POLYPHONY

It is well accepted that A-S-R portions of a music note strongly contributes to characterize an instrument [17]. For a given instrument, variability in attributes such as instrument make, style of playing, note being played etc., reflects as variability in A-S-R features of the instrument. Stochastic models can be used to capture these A-S-R variabilities [11, 18–20]. We model the evolution of A-S-R patterns for different notes of a given instrument using a 3-state left to right (LR) HMM with each state emission distribution being modeled using GMMs to represent A-S-R pattern variability across notes of a given instrument.

In a polyphonic signal, $y(t)$, let the features of a T length segment be denoted by $\{Y_t\}_{t=1}^{T} = [Y_1, Y_2, \dots Y_T]$. Each vector, $Y_t \in \mathbb{R}^{D\times 1}$, is comprised of contributions from one or more of M instruments. It is likely that in a given polyphonic piece, attack of an instrument overlaps with release of another, while a few other instruments are in a sustain state. i.e., the instrument streams independently evolve over time, but jointly contribute to polyphonic signal Y_t at frame t. So, Y_t can be modeled as a coupled emission from any state of all instruments. We propose to model this independent behaviour of instruments, and address the coupling in $\{Y_t\}$ due to various instrument states, from a generative model perspective. The proposed F-GM-HMM model uses factorial framework with each instrument being modeled by a GM-HMM.

A graphical model of F-GM-HMM model is shown in Fig. 1(a). $\{Y_t\}$ are shaded to indicate that they are observed polyphonic feature vectors. Let each horizontally connected variables of the graph (referred to as chain) denote K-states (3 for A,S,R) of first order LR-HMM model of one instrument. M such chains indicate possible contributions from M instruments. A $K \times M$ matrix, $\mathbf{S}_t \triangleq [S_t^1, S_t^2, \dots S_t^M]$, denotes the latent states such that $S_t^m(k_m) = 1$, if the m^{th} instrument is in k_m^{th} state at t^{th} time frame and 0 otherwise, $\forall k_m \in [1, K]$, $m \in [1, M]$, $t \in [1, T]$. Let each latent state $S_t^m(k_m)$ generate another latent variable Z_t^{m,k_m} to indicate the state's mixture density component chosen. Then, $\mathbf{Z}_t^m$, a $\check{P} \times K$ LV matrix, is such that $Z_t^{m,k_m}(p_k) = 1$, if at t^{th} instant, p_k^{th} mixture component ($p_k \in [1, \check{P}]$) from k_m^{th} state of the m^{th} instrument is emitting observations. The complete likelihood of latent sequences and the observed vectors can be written as:

$$\begin{aligned} P(\{\mathbf{S}_t, \mathbf{Z}_t, Y_t\}|\boldsymbol{\theta_a}) &= P(\{\mathbf{S}_t\}|\boldsymbol{\theta_a})P(\{\mathbf{Z}_t\}|\{\mathbf{S}_t\}, \boldsymbol{\theta_a}) \\ &\quad P(\{Y_t\}|\{\mathbf{Z}_t\}, \boldsymbol{\theta_a}) \end{aligned}$$

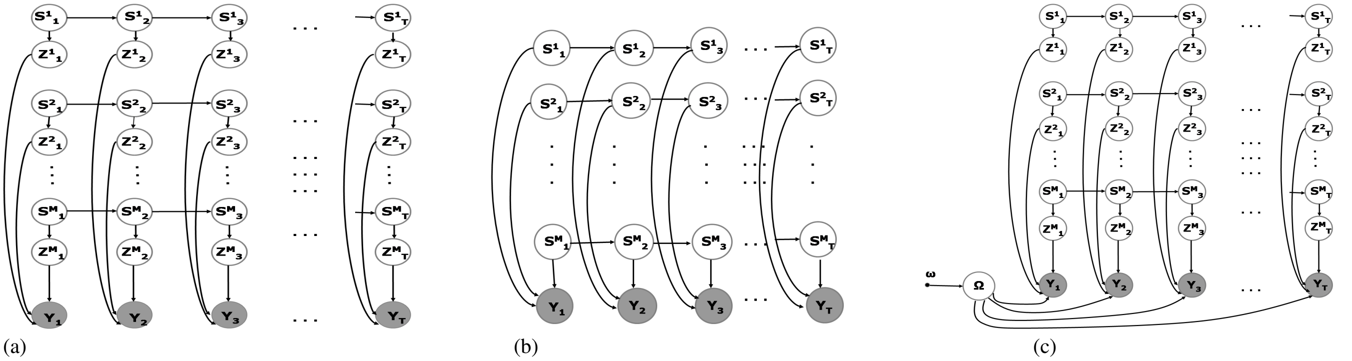
S^1_1 S^1_2 S^1_3 . . . S^1_T
Z^1_1 Z^1_2 Z^1_3 Z^1_T
S^2_1 S^2_2 S^2_3 . . . S^2_T
Z^2_1 Z^2_2 Z^2_3 Z^2_T
S^M_1 S^M_2 S^M_3 . . . S^M_T
Z^M_1 Z^M_2 Z^M_3 Z^M_T
Y_1 Y_2 Y_3 . . . Y_T
ω
Ω
(a)
(b)
(c)

Fig. 1. Graphical models depicting (a) F-GM-HMM, (b) F-G-HMM (c) LV-F-GM-HMM.

$$= \prod_{t=1}^{T} P(\mathbf{S}_t, \boldsymbol{\theta_a}) P(\mathbf{Z}_t|\mathbf{S}_t, \boldsymbol{\theta_a}) P(Y_t|\mathbf{Z}_t, \boldsymbol{\theta_a}) \tag{1}$$

$$\text{where, } P(\mathbf{S}_1, \boldsymbol{\theta_a}) = \prod_{m=1}^{M} P(S_1^m|\theta_a^m), \tag{2}$$

$$P(\mathbf{S}_t, \boldsymbol{\theta_a}) \quad = \quad \prod_{m=1}^{M} P(S_t^m|S_{t-1}^m, \theta_a^m) \; \forall t > 1, \tag{3}$$

$$P(\mathbf{Z}_t|\mathbf{S}_t, \boldsymbol{\theta_a}) = \prod_{m=1}^{M} \prod_{k_m=1}^{K} P(Z_t^{m,k_m}|S_t^m(k_m), \theta_a^m) \tag{4}$$

$$P(Y_t|\mathbf{Z}_t, \boldsymbol{\theta_a}) = P(Y_t|\{\mathbf{Z}_t^m\}_{m=1}^M, \boldsymbol{\theta_a}) \tag{5}$$

With each state emission probabilities modeled as mixture Gaussian distributions, we impose the following on polyphonic observation Y_t conditioned on the state and emission variables :

$$P(Y_t|Z_t^{1,k_1}, Z_t^{2,k_2}, \dots Z_t^{M,k_M}, \boldsymbol{\theta_a}) \triangleq$$

$$\mathcal{N}\big(Y_t; \sum_{m=1}^{M} \boldsymbol{\mu}^m \mathbf{Z}_t^m, \sum_{m=1}^{M} \boldsymbol{C}^m \mathbf{Z}_t^m\big) \tag{6}$$

where $\boldsymbol{\theta_a} = [\theta_a^1, \theta_a^2, \dots \theta_a^M]$ and $\theta_a^m = [\pi^m, A^m, \boldsymbol{\alpha}^m, \boldsymbol{\mu}^m, \boldsymbol{C}^m]$ are the parameters of the monophonic instrument GM-HMM model , $\forall m \in [1, M]$. $\boldsymbol{\mu}^m$ is a $D \times \check{P} \times K$ matrix containing $\check{P}$ means (μ) for each of the K states of m^{th} instrument. The latent variable Z_t^{m,k_m} chooses the relevant mean from $\boldsymbol{\mu^m}$, while $\mathbf{C}^m$ is its corresponding covariance matrix.

In effect, Eqn. (6) assumes that the polyphonic Y_t is sum of Gaussian random variables from M instruments, whose evolution is indicated by latent variables $\mathbf{S}_t$ and $\mathbf{Z}_t$ resulting in joint decoding of instruments. We can also interpret it as observations, Y_t compared with not just each θ_a^m, but a combination of them. In other words, Eqn. (6) attempts to find the combination of distributions that can best explain the observations.

However, it is difficult to separate $\mathbf{Z}_t^m$ term entangled in the inverse of sum of covariance matrices in Eqn.(6), for inference purposes. Moreover, $\mathbb{C} \triangleq (\mathbb{E}(\{Y_t^2\}_{t=1}^T) - \mathbb{E}(\{Y_t\}_{t=1}^T)^2)$ and, $\mathbb{C} = \sum_{m=1}^{M} \boldsymbol{C}^m \mathbf{Z}_t^m$. Hence, we equivalently use $\mathbb{C}$, evaluated for every observation segment, in Eqn.(6).

Now, we can observe from Fig. 1(a), Eqn. (1-6), that the F-G-HMM proposed in [16] using a single Gaussian emission for each state of a chain HMM can be seen as a special case of F-GM-HMM i.e., for $\check{P} = 1$, one can arrive at F-G-HMM formulations in [16]. Equivalent to Eqn. (5) and Eqn. (6), we get:

$$P(Y_t|\mathbf{S}_t, \boldsymbol{\theta_b}) = P(Y_t|\{\mathbf{S}_t^m\}_{m=1}^M, \boldsymbol{\theta_b}) \tag{7}$$

$$P(Y_t|S_t^1, S_t^2, \dots S_t^M, \boldsymbol{\theta_b}) \triangleq \mathcal{N}\big(Y_t; \sum_{m=1}^{M} \boldsymbol{\mu}^m \mathbf{S}_t^m, \mathbb{C}\big) \tag{8}$$

where $\boldsymbol{\theta_b} = [\theta_b^1, \theta_b^2, \dots \theta_b^M]$ and $\theta_b^m = [\pi^m, A^m, \boldsymbol{\mu}^m, \boldsymbol{C}^m]$ are the parameters of the G-HMM monophonic instrument model, $\forall m \in [1, M]$, with $\mathbb{C}$ being calculated from data of every observation segment. A graphical model of F-G-HMM is shown in Fig. 1(b). Originally, F-G-HMM in [16] is developed to model real time series data that may have complex internal structure. One can also use F-G-HMM with A-S-R still modeled by 3-state single Gaussian LR HMM. However, we would not be capturing the variabilities in attack, sustain or release state emission accurately. Also, experimentally, we observe GM-HMMs performing significantly better than G-HMMs on monophonic instrument identification as shown in Table. 1. We expect similar performance in the factorial approach.

In order to capture hidden dynamics of all instruments, best explaining a given polyphonic feature vector Y_t, we must infer the probability of states, $\mathbf{S}_t$ and its emission component $\mathbf{Z}_t$ of the F-GM-HMM model. An exact inference for $\mathbf{Z}_t$ and $\mathbf{S}_t$ will require a large number of forward backward recursions over latent variable $\mathbf{Z}_t^m$ which in turn has $\check{P} \times K$ values in each m, resulting in time complexity of order of $\mathcal{O}(TM(\check{P}K)^{M+1})$. Hence, we use variational inference algorithm to accommodate the mixture emission model.

We derive structured variational inference for F-GM-HMM analogous to F-G-HMM in [16]. The complete likelihood of the variables of graphical model in Fig. 1(a) is given by Eqn. (1). The first 2 terms of RHS in Eqn. (1) is determined by θ_a^m alone [from Eqn. (2-4)]. It is required to evaluate the 3rd term for which we introduce an approximate distribution Q_v to minimize the KL divergence between complete likelihood given by Eqn. (1), and an approximate distribution $Q(\{\boldsymbol{S_t}\}, \{\boldsymbol{Z_t}\}|\boldsymbol{\theta_a}) \triangleq Q_v$ where,

$$Q_v \quad = \quad \frac{1}{Z_Q} \prod_{m=1}^{M} P(S_1^m|\theta_a^m) \prod_{t=2}^{T} \prod_{m=1}^{M} P(S_t^m|S_{t-1}^m, \theta_a^m)$$

$$\prod_{m=1}^{M} \prod_{t=1}^{T} Q(Z_t^{m,k_m}|S_t^m, \boldsymbol{\theta_a}) \tag{9}$$

$$\text{with,} \quad Q(Z_t^{m,k_m}|S_t^m, \boldsymbol{\theta_a}) = \prod_{p_k=1}^{\check{P}} [h_t^{m,k_m}(p_k)]^{Z_{t,p_k}^{m,k_m}} \tag{10}$$

and Z_Q is for normalization. By minimizing, $\mathcal{KL}(Q_v||P)$ w.r.t $\log h_t^{n,k_n}$, we get a closed form expression for h_t^{m,k_m} as:

$$h_t^{m,k_m} \quad = \quad \alpha^{m,k_m} \exp\Big\{\mu^{m,k_m\,T} \mathbb{C}^{-1} \big[Y_t -$$

$$\sum_{n \neq m} \mu^{n,k_n} \mathbb{E}(Z_t^{n,k_n})\big] - \frac{1}{2}\delta^{m,k_m}\Big\} \tag{11}$$

where, $\delta^{m,k_m} = \text{diag}\{\mu^{m,k_m\,T} \mathbb{C}^{-1} \mu^{m,k_m}\}$. Eqn. (11) is very similar to the variational inference based solution for Eqn. (8) of F-G-HMM [16] i.e.,

$$h_t^m = \exp\{\mu^{m\,T} \mathbb{C}^{-1}(Y_t - \sum_{n \neq m} \mu^n \mathbb{E}(S_t^n)) - \frac{1}{2}\delta^m\} \tag{12}$$

Eqn. (11) is inclusive of the weights $\boldsymbol{\alpha^m}$ of the emission Gaussian mixtures of $\boldsymbol{\theta_a}$. For the considered application, we do not update $\boldsymbol{\theta_a}$. The algorithm to calculate the likelihood of $\{Y_t\}$ for F-GM-HMM is drafted below.

```
*Algo 1: F-GM-HMM*
  *Training*: Find parameters $\theta_a^m$ of each of $M$ instruments.
  *Testing*: Initialize all $\mathbb{E}(Z_t^{m,k_m})$ to be equal.
    (i) Calculate $\check{P} \times 1$ vector, $h_t^{m,k_m}$ vector, using Eqn. (11).
    (ii) Use $\mathbf{h}_t^m = P(Y_t|Z_t^{m,k_m})$ in the forward backward algo in $M$
    HMMs to evaluate $\mathbb{E}(Z_t^{m,k_m})$, total likelihood. $\theta_a^m$ are unaltered.
*Stopping criteria*: Repeat (i) & (ii) till likelihood change $\leq \epsilon$ (0.001)
```

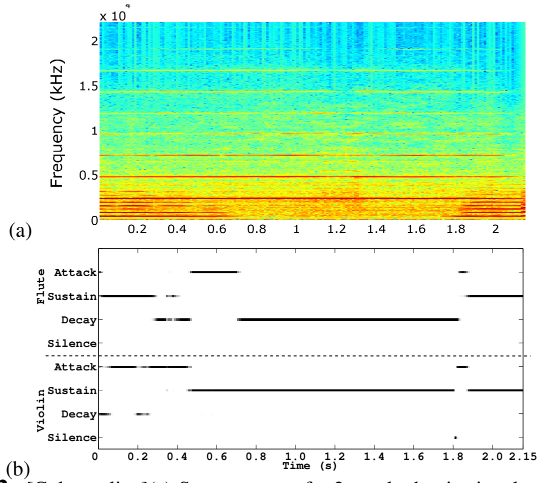

Fig. 2. [Color online](a) Spectrogram of a 2 s polyphonic signal comprising of flute & violin instruments (b) Instrument evolution in the polyphony depicted by forward likelihood, $P(\{Y_t\}, S_t^m(k)|\boldsymbol{\theta_a})$ for all states and instruments calculated using F-GM-HMM.

In order to achieve decoupling in F-GM-HMM, the variable h_t^{m,k_m} approximates the likelihood of data w.r.t. latent variables $\{S_t^m\}$, $\{Z_t^{m,k_m}\}$ for all possible combinations of the variables and in turn calculates likelihood of all M models.

The advantage of factorial models is that it is possible to simultaneously calculate the probability of all states, of all instruments in a polyphonic signal. For the testing phase, we have introduced an additional silence state for each instrument so that no instrument is forced into any of its K trained (attack-sustain-release) states. This is important when $< M$ instruments contribute to the polyphony. The $(K+1) \times (K+1)$transition matrix for achieving the same is engineered by subtracting small amounts (∼0.05) from the diagonal terms of A^m. The μ^m matrices are appended with feature vectors containing zero entries to simulate silence state. We illustrate the same in Fig. 2(b) for a 2 s mixture signal comprising of 2 instruments (flute and violin) along with its spectrogram to depict evolution. The mixture signal contains a decaying flute note for 0.5 s, followed by another onset at 1.8 s; and, a violin note is present in the entire 2 s duration, with its attack lasting for 0.5 s (note the gradual increase in strength of higher harmonics). We calculate $P(\{Y_t\}, S_t^m|\boldsymbol{\theta_a})$ using forward-backward algorithm for each instrument, after decoupling $P(Y_t|Z_t^{m,k_m})$ for both flute & violin. An instrument is said to be active if it contributes to the polyphony else passive.

3. LV-GM-HMM AND LV-F-GM-HMM MODELS

Although likelihoods from F-GM-HMM generative model for a polyphony indicates presence of several instruments in different states, it is difficult to estimate active instruments directly as discriminability between active and passive instrument is lesser. We therefore propose to use these likelihoods in generic LV framework of [15] (and evaluated for tMM) for detecting instruments in a polyphony.

Let $\Omega \in \mathcal{B}^M$ & $\mathcal{B} = \{0, 1\}$ be a LV to flag an active instrument. Let $\boldsymbol{\theta}$ indicate the parameter set of all M individual instrument models. The likelihood of m^{th} instrument using F-GM-HMM (*Algo 1*), for smaller segments of $\{Y_t\}$ of duration $\tau < T$, $\forall t \in [1, T/t']$ is given by:

$$P_m(\{Y_t\}_{t=t'-\tau}^{t'}|\boldsymbol{\theta_a}) = \sum_{k_m=1}^{K} P(\{Y_t\}, S_t^m(k_m) = 1|\boldsymbol{\theta_a}) \qquad (13)$$

We use this likelihood in the generic LV framework given by:

$$P(Y_t|\boldsymbol{\theta_a}) = \sum_{m=1}^{M} \omega_m P_m(\{Y_t\}_{t=t'-\tau}^{t'}|\boldsymbol{\theta_a}) \; s.t., \sum_{m=1}^{M} \omega_m = 1 \qquad (14)$$

where, $\omega_m = P(\Omega(m) = 1)$ indicates probability of m^{th} instrument being active. We refer to this approach as LV-F-GM-HMM (graphical model in Fig. 1(c)). Similarly, using m^{th} GM-HMM model, θ_a^m to calculate likelihood, $P(\{Y_t\}_{t=t'-\tau}^{t'}|\theta_a^m)$ of m^{th} instrument, we get LV-GM-HMM model as:

$$P(Y_t|\boldsymbol{\theta_a}) = \sum_{m=1}^{M} \omega_m P(\{Y_t\}_{t=t'-\tau}^{t'}|\theta_a^m) \; s.t., \sum_{m=1}^{M} \omega_m = 1 \qquad (15)$$

The EM algorithm to evaluate ω_m is as follows:

```
*Algo 2: LV-GM-HMM or LV-F-GM-HMM*
  *Training*: Find parameters $\theta_a^m$ of each of $M$ instruments.
  *Testing*: Initialize all $\omega_m$ to be equal.
    (i) Find posterior $\gamma_{mt'} = \frac{\omega_m P(\{Y_t\}_{t=t'-\tau}^{t'}|\theta_a^m)}{\sum_m \omega_m P(\{Y_t\}_{t=t'-\tau}^{t'}|\theta_a^m)}$
    (ii) Evaluate $\omega_m = \frac{t'}{T} \sum_{t=1}^{T/t'} \gamma_{mt'}$
*Stopping criteria*:Repeat (i) & (ii) till posterior change $\leq \epsilon'$(0.001)
```

For insight into differences between LV-GM-HMM and LV-F-GM-HMM methods, we note that in the former approach, likelihood of polyphonic signal is evaluated using each monophonic signal model. The major assumption in this model defined by Eqn. (15) is that $P(Y_t|\Omega(m) = 1, \boldsymbol{\theta_a}) \triangleq P(\{Y_t\}|\theta_a^m)$. This assumption decouples the set of models $\boldsymbol{\theta_a}$ in LV-GM-HMM. The same assumption is true for LV-F-GM-HMM too, but $P_m(\{Y_t\}|\boldsymbol{\theta_a})$ is calculated jointly considering all M models .i.e., decoupling of likelihood by F-GM-HMM takes into account all possible instruments' states and components, through Eqn. (6,13) (*Algo 1*). Individual likelihood is then used in LV framework to equivalently arrive at active set of instruments, through Eqn. (14)(*Algo 2*). Since, we take into account all the instruments' states before calculating individual likelihood, a better performance in segments containing overlapping instruments can be expected.

4. EXPERIMENTS AND RESULTS

The performance of the proposed approach is evaluated on RWC database [22] on three models: LV-GM-HMM, LV-tMM [15] & LV-F-GM-HMM models. The 8 chosen instruments are: clarinet, flute, guitar, harp, mandolin, piano, trombone and violin. 12 dimensional Mel-frequency cepstral co-efficients (MFCC) with Δ & Δ^2 co-efficients are used as feature vectors for training after silence removal. We have used a frame length of 25 ms, a frame shift of 10 ms for obtaining the MFCC using HTK [23]. An analysis segment, $\{Y_t\}_{t=1}^{T}$, constitutes T such consecutive frames.

4.1. Instrument models

Monophonic training data comprises of individual notes from atleast 5 min of data from RWC dataset for each instrument with silence removal. We train using tMM [15], GMM and HMM models.

We use diagonal covariance matrices in the mixture models. For HMMs, we have used 3-state Left to Right (LR) latent variable such that each state captures attack, sustain and release portions of an instrument note. τ for HMM is set to 50 ms. A set of ∼8000 randomly selected notes, excluding training set, is chosen for all instruments in the testing phase. The instrument model yielding highest likelihood for given data is taken as the detected instrument. Each train or test data length is at-most 5 s. The test data and training data of all instruments are such that they differ in either the artist or instrument manufacturer [22]. Table. 1 shows the performance accuracy evaluated on test notes, for solo instrument recognition task, with varying mixture components in each model. As expected, increasing mixture components results in a better model. The improvement is lesser as mixture components increase. tMM and GMM show similar performance, while HMMs consistently show better performance justifying A-S-R contribution in instrument recognition.

Table 1. Average note-wise recognition accuracy for 8 monophonic instruments using different models.

Mixtures	1	4	8	16	32
GMM	58.61	70.49	73.64	78.09	80.60
tMM	59.57	66.54	74.82	77.82	80.83
3-state LR-HMM	65.54	75.26	77.30	79.48	81.27

4.2. Multi-Instrument Recognition in Polyphonic Music

Multi-instrument polyphonic test signals are created by linear addition of amplitude normalized monophonic test data of the $M = 8$ instruments. A L-polyphony test set comprises of all $\binom{M}{L}$ combination of instruments, with $L \in [2, 5]$. Performance of detecting all instruments in each 5 s segment is measured using F-measure. An instrument in T length segment is considered detected if any $\omega_m > \epsilon_{th}$, with ϵ_{th} as threshold. For polyphonic detection, we have used only 12 MFCCs of test and training models, as the mixture signal will not have Δ and Δ^2 co-efficients linearly related to individual instruments. MFCCs themselves are also not additive, but can be approached through Parallel Model Combination [24]. However, combining multiple such mixture distributions is not straightforward and is beyond the scope of this paper. The detection performance per frame using LV-tMM (32-component tMM), and LV-GM-HMM and LV-F-GM-HMM (both with 8 component Gaussian for each of 3 states LR model) is shown in Fig. (3). We report results for T of 500 frames (5 s), with a segment shift of 100 frames (1 s).

The LV-F-GM-HMM exhibits better detection accuracy as number of polyphony increases. This can be explained because LV-tMM or LV-GM-HMM rely on the regions where there is little or no overlap for its performance. When the signals overlap completely, the accuracy drops drastically, more so in the higher polyphony case. In such signals, LV-F-GM-HMM are seen to be showing greater accuracy. LV-F-GM-HMM is observed to exhibit lower performance scores in lower polyphony. This behaviour is observed to be linked to Eqn. (11), which can be seen as a search for best possible combination from all the M instrument models to explain the data and hence unlikely to yield a sparse solution for lesser polyphony, even if the signals are additive in nature, in absence of sparsity constraints to control the number of false detections.

We have tested on TRIOS dataset [25]. We have chosen the common instruments i.e., clarinet, piano & violin signals to create 2 and 3 polyphony signals. The total duration of constructed 2 and 3 polyphony is $\sim$ 500 min. We use the RWC trained models. The mean of F-measures are plotted in Fig. 3(b). We find that the accuracy of LV-F-GM-HMM is better than that of LV-tMM or LV-GM-HMM for 3- polyphony. The overall detection accuracy itself is quite low. This is attributed to training-testing cross database environment mismatch & presence of chords in the piano portions, which are not in the training set.

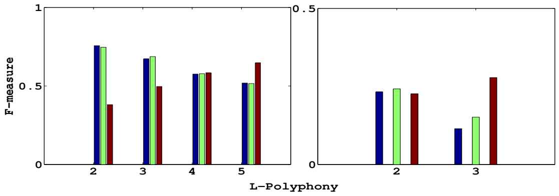

Fig. 3. [Color online] Average of segment-wise F-measure of LV-tMM (blue), LV-GM-HMM (green) and LV-F-GM-HMM (red) for L-polyphony generated from (a) RWC dataset (b) TRIOS dataset. $T = 5\ s$, $t' = 100\ ms$, $\tau = 500\ ms$ & $\epsilon_{th} = 0.1$.

5. DISCUSSION

We have proposed a generative F-GM-HMM model for detecting instruments and their evolution in polyphonic signals. We have shown F-GM-HMM is shown to be a generalized version of F-G-HMM. Simultaneous capture of A-S-R patterns in polyphony using F-GM-HMM is demonstrated. For analysis of structures in mixture signals, F-GM-HMM being a more general version, can be expected to give more flexibility where detection/classification (rather than parameter estimation) is vital.

Two closely related approaches using factorial HMM are, Factorial Scaled HMM (FSHMM) in [26] and Non-negative factorial HMM (NFHMM) in [27], for source separation. In FSHMM, the mixture signal is modeled as sum of the components from latent states of the HMMs. The components of each state of HMM are assumed to exhibit zero mean Gaussian mixture distributions. The spectral characteristics and amplitude factors are accounted in the variance of the Gaussians. Succeeding parameter estimation step, using multiplicative updates, is the inference step to obtain the amplitude factors to separate the sources. The amplitude factors play a similar role as Z_t^{m,k_m} in our formulation, even though approaches for modeling the spectral envelope are different. Non-negative Matrix Factorization (NMF) with Markov chained bases (called MNMF) have also been proposed in [28] to learn A-S-R patterns from spectrogram and it is shown that FSHMM is a particular case of MNMF. NFHMM, a non-parametric model, uses a latent variable to identify elements as well as to assign proportion from a combined dictionary over both sources (as speech sources may have spectrally similar aspects). We note that using discrete density in Eqn. (5) leads to formulation similar to N-FHMM in [27]. For our application, spectral sources do not share a dictionary/model (owing to different distributions) resulting in a different graphical model as against in [27, 29]. Further, we decouple only the probability of the polyphonic signal and not the signal itself as in signal separation application. For detecting multiple instruments, the instrogram technique in [10] uses HMM, similar to that of $P(Y_t|\theta_m^a)$ likelihood calculation (for a set of 28 features). However, no latent variable is used, and the likelihood is multiplied with a non-specific instrument existence probability to accentuate the position of notes. Whereas in [11, 12], temporal and durational constraints are added to discrete HMMs to capture the dynamics of sound (in terms of attack, sustain and release) along with notes. The approach of identifying the instruments can be seen as similar to LV-GM-HMM formulation with discrete distributions.

We have decoupled each instrument contribution in total likelihood, and used them in LV framework for identifying multiple instruments in polyphony. We have evaluated on cross database using LV based F-GM-HMMs and found them to be more advantageous for higher polyphony, than LV based mixture models. This is significant when many harmonics of the component instruments overlap.

6. REFERENCES

[1] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. Int. Soc. on Music Info. Retr. Conf (ISMIR)*, 2009, pp. 327–332.

[2] B. Raj and P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *Proc. Workshop on Applic. of Sig. Process. to Audio and Acoust.*, 2005, pp. 17–20.

[3] B. Raj, M. V. S. Shashanka, and P. Smaragdis, "Latent dirichlet decomposition for single channel speaker separation," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, May 2006, vol. 5.

[4] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[5] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 3, pp. 529–540, march 2008.

[6] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 1, pp. 68 – 80, Jan. 2006.

[7] J. Wu, E. Vincent, S. A. Raczynski, T. Nishimoto, N. Ono, and S. Sagayama, "Polyphonic pitch estimation and instrument identification by joint modeling of sustained and attack sounds," in *IEEE J. Sel. Topics Signal Process.*, vol. 5, pp. 1124–1132.

[8] G. Grindlay and D. P. W. Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1159–1169, Oct. 2011.

[9] D. Giannoulis and A. Klapuri, "Musical instrument recognition in polyphonic audio using missing feature approach," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 9, pp. 1805–1817, 2013.

[10] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. Okuno, "Musical instrument recognizer instrogram and its application to music retrieval based on instrumentation similarity," in *Proc. of Int'l Symp. on Mult.*, 2006, pp. 265–274.

[11] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model," *J. Acoust. Soc. Amer.*, vol. 133, no. 3, pp. 1727–1741, 2013.

[12] E. Benetos and T. Weyde, "Explicit duration hidden Markov models for multiple-instrument polyphonic music transcription.," in *Proc. Int. Soc. on Music Info. Retr. Conf (ISMIR)*, 2013, pp. 269–274.

[13] K. Yoshii and M. Goto, "Infinite composite autoregressive models for music signal analysis," in *Proc. Int. Soc. on Music Info. Retr. Conf (ISMIR)*, Porto, Portugal, October 8-12 2012.

[14] K. Yoshii and M. Goto, "Infinite latent harmonic allocation: A nonparametric bayesian approach to multipitch analysis," in *Proc. Int. Soc. on Music Info. Retr. Conf (ISMIR)*, Utrecht, The Netherlands, August 9-13 2010, pp. 309–314.

[15] H. Sundar, H. G. Ranjani, and T. V. Sreenivas, "Student's-t mixture model based multi-instrument recognition in polyphonic music," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2013, pp. 216–220.

[16] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, no. 2-3, pp. 245–273, 1997.

[17] K. W. Berger, "Some Factors in the Recognition of Timbre," *J. Acoust. Soc. Amer.*, vol. 36, pp. 1888, 1964.

[18] T. Virtanen and T. Heittola, "Interpolating hidden Markov model and its application to automatic instrument recognition," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.* IEEE, 2009, pp. 49–52.

[19] A. Klapuri and T. Virtanen, "Representing musical sounds with an interpolating state model," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 3, pp. 613–624, 2010.

[20] M. Eichner, M. Wolff, and R. Hoffmann, "Instrument classification using hidden Markov models," *Proc. Int. Soc. on Music Info. Retr. Conf (ISMIR)*.

[21] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998.

[22] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database : Music genre database and musical instrument sound database," in *Proc. Int. Soc. on Music Info. Retr. Conf (ISMIR)*, pp. 229–230.

[23] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[24] Mark Gales, Steve Young, and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 352–359, 1996.

[25] J. Fritsch, *High quality musical audio source separation*, Ph.D. thesis, Centre for Digital Music, 2012.

[26] A Ozerov, C. Fevotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. Workshop on Applic. of Sig. Process. to Audio and Acoust.*, Oct 2009, pp. 121–124.

[27] G. J. Mysore and M. Sahani, "Variational inference in non-negative factorial hidden Markov models for efficient audio source separation," in *IEEE Int. Conf. Machine Learning*, 2012, pp. 1887–1894.

[28] Masahiro Nakano, Jonathan Le Roux, Hirokazu Kameoka, Yu Kitano, Nobutaka Ono, and Shigeki Sagayama, "Non-negative matrix factorization with markov-chained bases for modeling time-varying patterns in music spectrograms," in *Latent Variable Analysis and Signal Separation*, pp. 149–156. Springer, 2010.

[29] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *Latent Variable Analysis and Signal Separation*, pp. 140–148. Springer, 2010.