STUDENT'S-t MIXTURE MODEL BASED MULTI-INSTRUMENT RECOGNITION IN POLYPHONIC MUSIC

Harshavardhan Sundar, Ranjani H. G., and T. V. Sreenivas

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore-560 012, India
{harsha,ranjanihg,tvsree}@ece.iisc.ernet.in

ABSTRACT

We address the problem of multi-instrument recognition in polyphonic music signals. Individual instruments are modeled within a stochastic framework using Student's-t Mixture Models (tMMs). We impose a mixture of these instrument models on the polyphonic signal model. No a priori knowledge is assumed about the number of instruments in the polyphony. The mixture weights are estimated in a latent variable framework from the polyphonic data using an Expectation Maximization (EM) algorithm, derived for the proposed approach. The weights are shown to indicate instrument activity. The output of the algorithm is an Instrument Activity Graph (IAG), using which, it is possible to find out the instruments that are active at a given time. An average F-ratio of 0.75 is obtained for polyphonies containing 2-5 instruments, on a experimental test set of 8 instruments: clarinet, flute, guitar, harp, mandolin, piano, trombone and violin.

Index Terms— Student's-t Mixture Models, Latent Variable, Polyphony, Instrument Recognition, Instrument Activity Graph.

1. INTRODUCTION

The primary challenge in Music Information Retrieval (MIR) is to unravel the underlying polyphonic texture and multi-instrument structure of the music signal. Apart from obtaining melodies, multi-instrument recognition in polyphonic music signals plays a major role in the science of MIR.

Automatic instrument recognition approaches can be broadly classified into two categories [1]. In the first category, instrument recognition is performed after separating the individual instrument signals from the polyphonic music signal [2–5], using techniques such as Probabilistic Latent Component Analysis (PLCA), Non-negative Matrix Factorization (NMF), instrument specific harmonic models. In the second category, detection of the constituent instruments in the polyphony is addressed without separation [6–8]. In the recent years, the problem of jointly addressing transcription and instrument recognition is seen in [9–12].

In this paper, the focus is on multi-instrument recognition. We propose a latent variable (LV) framework for identifying the constituent instruments of a polyphonic music signal (Section (2)). Contrary to other LV approaches, we perform detection without signal separation. Stochastic models are used to model individual instruments (Section (3.1)). The probability density function of the polyphonic signal is then modeled as a convex combination of the individual instrument models. The weights of the convex combination are estimated from the polyphonic music signal using an Expectation Maximization (EM) algorithm. It is shown that these weights indicate the presence or absence of an instrument in the polyphony. We construct an "Instrument Activity Graph" (IAG) (Section (2.1)) using the model-weights, to indicate the activity of each modeled instrument across time. Performance of the proposed approach on polyphonic signals from RWC database [13] is shown in Section (3.2). The contributions of this paper are: i)A generic LV approach for music instrument recognition, which admits different kinds of instrument models. ii) Experimental evaluation of Student's-t Mixture Models (tMMs) as efficient instrument models for recognition. iii) A graphical display of instrument activity over time (IAG). iv) Performance analysis on varying test data length for online applications. Relation to prior work is detailed in Section (4).

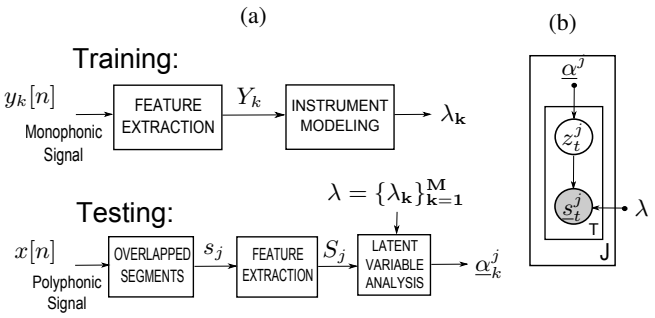


Fig. 1: a) Block diagram of the proposed approach. b) Probabilistic graphical model of the proposed approach.

2. THE PROPOSED APPROACH

Figure 1a shows a block diagram of the proposed approach. The training phase consists of building stochastic models for each instrument. Consider a set of M instruments $I = \{I_k\}_{k=1}^{M}$. Each instrument I_k is trained on feature

vectors $Y_k = [\underline{f}_{k_1}, \underline{f}_{k_2}, \cdots, \underline{f}_{k_{T_k}}]$, obtained from its monophonic signal $y_k[n]$. $\underline{f}_{k_t} \in \mathbb{R}^d$ is a feature vector at t^{th} frame. Let $\lambda_{\mathbf{k}}$ denote the parameters of the probability density function (p.d.f.) used to model the instrument I_k.

Consider a polyphonic signal $x[n]$. We refer to the instruments contributing to this signal as active instruments. The objective is to identify the active instruments in any given interval of a polyphonic signal. In the testing phase, we subject the features estimated from $x[n]$ to the proposed LV analysis to obtain the activity of each of the M instruments as described below.

Consider the j^{th} segment of $x[n]$: $s_j \triangleq \{x[n]\}_{n=j_0+1}^{j_0+N_w}$, where $j_0 \triangleq (j-1)N_{sh}$, N_w is the segment length and N_{sh} is the segment shift. Let $S_j = \left[\underline{s}_1^j, \underline{s}_2^j, \ldots, \underline{s}_T^j\right]$ denote T feature vectors estimated from the segment s_j, where $\underline{s}_t^j \in \mathbb{R}^d$. The feature extraction scheme used to obtain S_j from s_j is identical to that used to obtain Y_k from $y_k[n]$. We introduce a latent variable $\underline{Z}_t^j \in \mathcal{B}^M$; where the set $\mathcal{B} = \{0, 1\}$, to discover the active instruments in the j^{th} segment of $x[n]$. Let the k^{th} element of the vector $\underline{Z}_t^j$ be denoted as $z_t^j(k)$. $z_t^j(k) = 1$ if $\underline{s}_t^j$ has contribution from the k^{th} instrument I_k. Hence, one or more elements of the vector $\underline{Z}_t^j$ can be unity, indicating the presence of corresponding instruments. Let $\lambda = \{\lambda_{\mathbf{k}}\}_{\mathbf{k=1}}^{\mathbf{M}}$ denote conditional p.d.f. parameter set. We model each vector of j^{th} segment in terms of the known instruments models as:

$$p\left(\underline{s}_t^j; \lambda\right) \triangleq \sum_{k=1}^{M} p\left(z_t^j(k) = 1\right)\ p\left(\underline{s}_t^j \mid z_t^j(k) = 1; \lambda\right); \quad (1)$$

$s.t.\ \sum_{k=1}^{M} p\left(z_t^j(k) = 1\right) = 1$. Equation (1) follows from the assumption that the music signal has contributions at least from one of the M modeled instruments. Therefore, (1) holds even when Y_k and S_j are not linearly related. The objective of using such a mixture of instrument models is to express the unknown (polyphonic signal) in terms of the known (individual instrument models). The validity of such a formulation for convolutive mixtures is explored in the context of speech in [14].

The significance of the above formulation is that, in general, a non-linear model in the feature vector space is viewed as a linear model in the probability space. The p.d.f. of the polyphonic signal is expressed in terms of the known instrument-models. From the definition of the latent variable, it follows that: $p\left(\underline{s}_t^j \mid z_t^j(k) = 1; \lambda\right) = p\left(\underline{s}_t^j; \lambda_{\mathbf{k}}\right)$.

Assuming that the distribution of latent variable does not vary within the j^{th} analysis segment, we denote $p\left(z_t^j(k) = 1\right)$ as α_k^j. Thus, (1) can now be written as,

$$p(\underline{s}_t^j; \lambda) \triangleq p\left(\underline{s}_t^j; \lambda, \underline{\alpha}\right) = \sum_{k=1}^{M} \alpha_k^j\ p\left(\underline{s}_t^j; \lambda_{\mathbf{k}}\right); \quad (2)$$

$s.t \sum_{k=1}^{M} \alpha_k^j = 1$, where $\underline{\alpha}^j \triangleq \left[\alpha_1^j, \alpha_2^j, \cdots, \alpha_M^j\right]$ where, $\underline{\alpha}^j \in \mathbb{R}^M$.

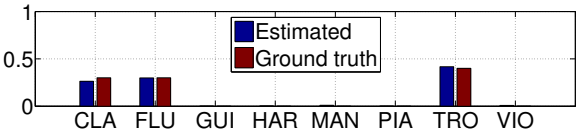


Fig. 2: *(Color Online):* An example indicating the estimated $\underline{\alpha}$ and the ground truth for a 3-polyphony.

Figure 1b shows the graphical model summarizing the proposed latent variable approach. The model-weights, $\underline{\alpha}^j$ estimate the contribution of different instrument models to the mixture model, $p(\underline{s}_t^j; \lambda)$. Assuming that the feature vectors in S_j are independent and identically distributed across t, we have:

$$p\left(S_j; \lambda, \underline{\alpha}\right) = \prod_{t=1}^{T} p\left(\underline{s}_t^j; \lambda, \underline{\alpha}^j\right). \quad (3)$$

To estimate the contribution of individual instrument models, $\underline{\alpha}^j$ are learnt from $\{\underline{s}_t^j\}$ and the instrument models parameterized by λ. We use the maximum likelihood estimation (MLE) approach to solve for α_k^j, using the EM algorithm [15]. Let the posterior probability at the m^{th} iteration be denoted as $\gamma_{kt}^j(m)$ and, $\gamma_{kt}^j(m+1)$ is the probability of $z_t^j(k) = 1$ at the $(m+1)^{th}$ iteration given $\underline{s}_t^j$, i.e.,

$$\gamma_{kt}^j(m+1) = \frac{\alpha_k^j(m)\ p\left(\underline{s}_t^j; \lambda_{\mathbf{k}}\right)}{\sum_{k=1}^{M} \alpha_k^j(m)\ p\left(\underline{s}_t^j; \lambda_{\mathbf{k}}\right)}, \quad (4)$$

where $\alpha_k^j(m)$ denotes the value of α_k in the j^{th} segment at the m^{th} iteration. We formulate the Q function as:

$$Q\left(\Psi, \Psi^{(m)}\right) = \sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_{kt}^j(m) \left[\log \alpha_k^j + \log p\left(\underline{s}_t^j; \lambda_k\right)\right],$$

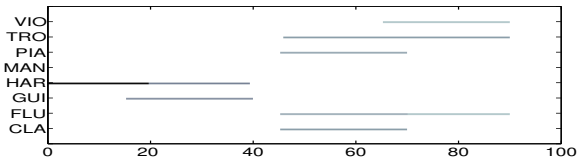
where $\Psi = \{\underline{\alpha}, \lambda\}$ and $\Psi^{(m)} = \{\underline{\alpha}(m), \lambda\}$. The parameters λ are fixed, and estimated a priori for individual instruments. The update for the parameters $\underline{\alpha}$ is given by:

$$\alpha_k^j(m+1) = \frac{1}{T} \sum_{t=1}^{T} \gamma_{kt}^j(m+1). \quad (5)$$

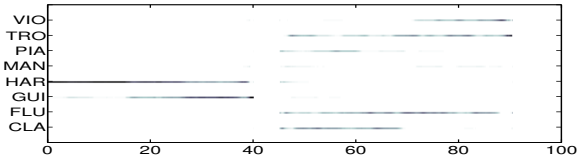
Figure 2 depicts the estimated $\underline{\alpha}$ for a 3-polyphony. The ground truth is established using the relative mixing ratios of individual instrument signals in the polyphonic signal. It is observed that the active instruments are accurately estimated.

2.1. Instrument Activity Graph

Let the model-weights obtained on convergence of EM algorithm be denoted as $\underline{\alpha}^{j*}$. Let α be the collection of these model-weights obtained for all segments, $1 \leq j \leq J$ i.e., $\alpha \triangleq \left[\underline{\alpha}^{1*}, \underline{\alpha}^{2*}, \ldots, \underline{\alpha}^{J*}\right]$ and $\alpha \in \mathbb{R}^{M \times J}$. We refer to a graphical display of α as "Instrument Activity Graph (IAG)". The y-axis of the plot indicate the M instrument indices and x-axis denotes time. An example IAG is shown in Figure 3. One can get an idea of the instruments active at any given time using an IAG. The specifics related to obtaining α (and hence IAG), are detailed in Section (3).



(a) Ground truth indicating instrument activity.



(b) Estimated IAG using the proposed algorithm. $N_w = 5\ s$ and $N_{sh} = 100\ ms$.

Fig. 3: An instrument activity graph for a polyphonic signal with varying number of instruments across time. α values are indicated by inverted grey scale.

Category	String	Wind	Key
Instruments chosen	Guitar, Harp Mandolin, Violin	Clarinet Flute, Trombone	Piano

Table 1: Different categories of instruments chosen.

3. EXPERIMENTAL RESULTS

The performance of the proposed approach is evaluated on RWC database [13]. We choose 8 instruments from three categories as shown in Table 1.

Twelve dimensional Mel-frequency cepstral co-efficients (MFCCs) are used as feature vectors after silence removal. A closer examination of the chosen instruments reveals that a minimum of 200 ms captures the attack-sustain period. Hence, a frame length of 250 ms and a frame shift of 10 ms is used for obtaining the MFCCs using HTK [16]. An analysis segment, S_j, constitutes T consecutive frames.

3.1. tMMs as Instrument models

It is our belief that speech and music are similar in many aspects and hence some of the advances in speech analysis can be borrowed to music analysis and vice-versa. In [17], we have shown that tMMs are in general better models than GMMs not just in terms of parsimony, but also in terms of accurate functional approximation. In particular, we have shown that for speaker recognition task, tMMs outperform GMMs. Therefore, we choose tMMs as instrument models.

The suitability of tMMs as instrument models is verified based on its performance on instrument recognition task in monophonic data. In the training phase, a 32−component tMM is used to model each instrument. The parameters are learnt using an EM algorithm for tMM [18]. For each instrument, monophonic training data of at least 5 minutes is used. In the testing phase, the detected instrument in the monophonic test signal, corresponds to the instrument model yielding highest likelihood of the data, given the model. For each instrument, a set of 20 randomly selected files, excluding the training set, is chosen for testing. Test data length is around 10 s. Care has been taken to ensure that the test data and training data of any given instrument differ in either the artist or the instrument manufacturer [13]. Table 2 shows the confusion matrix for the solo instrument recognition task. $(i,j)^{th}$ entry in this matrix denotes the percentage of files containing the i^{th} instrument and detected as the j^{th} instrument.

	C	F	T	P	G	H	M	V
Clarinet	96	0	0	4	0	0	0	0
Flute	0	100	0	0	0	0	0	0
Trombone	0	0	100	0	0	0	0	0
Piano	0	0	0	90	10	0	0	0
Guitar	0	0	0	0	100	0	0	0
Harp	0	0	0	0	0	100	0	0
Mandolin	0	0	0	0	0	10	100	0
Violin	0	0	0	0	0	0	0	100

Table 2: Confusion matrix of solo instrument recognition using t-MM models on RWC dataset.

The diagonal entries of the confusion matrix are dominant indicating the ability of the detection approach to yield more true positives. The confusion across instrument categories is almost negligible. The above experiment shows that the choice of feature vectors and instrument models is able to detect and differentiate between the chosen instruments. The choice of the feature set and the instrument models for a finer recognition performance is beyond the scope of this paper. Therefore, we use the above feature set and instruments models in the proposed approach for analyzing polyphonic signals.

3.2. Multi-Instrument Recognition in Polyphonic Music

Multi-instrument polyphonic test signals are created by linearly adding the amplitude normalized monophonic test data of different instruments. Polyphonic signals containing two to five instruments are created. A $K-$polyphony test set comprises all $\binom{M}{K}$ combination of instruments. Each polyphony contains at least 15000 segments. The performance of the algorithm in detecting all instruments in each segment, s_j, is measured using F-ratio [2]. An instrument I_k in j^{th} segment is considered detected if $\alpha_k^j > \epsilon$, where the threshold $\epsilon \in (0,1)$. The F-ratio is defined as, $F \triangleq \frac{2RP}{R+P}$, where R and P are recall and precision of detecting the instruments in the multi-instrument polyphony.

The EM algorithm in the proposed approach requires an initial estimate i.e., $\underline{\alpha}^j(0)$. Since the number of instruments in the given polyphony is generally not known a priori, all the $M-$instrument models are assumed to be equally probable and hence we initialize $\alpha_k^j(0) = \frac{1}{M}; \forall\ 1 \leq k \leq M; \forall\ 1 \leq j \leq J$.

In general, performance of stochastic models improves with more data. However, for online instrument recognition applications, real time processing on short data segments becomes imperative. Hence, there in an inherent trade-off between performance and segment length. Figure 4 shows the

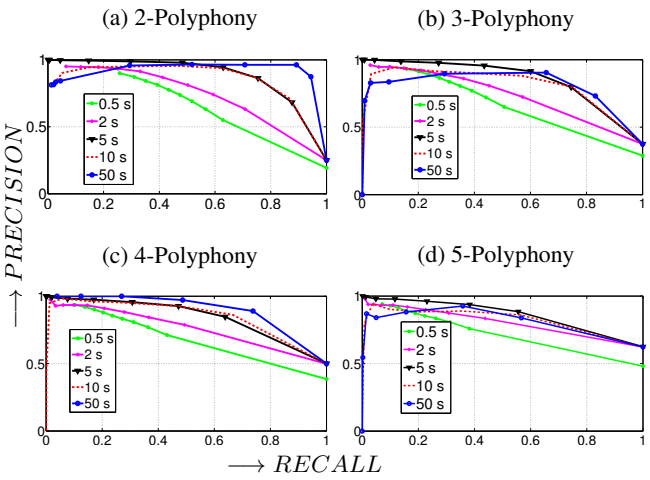


Fig. 4: *(Color Online):* PR curves for different segment lengths (N_w) and polyphony. $N_{sh} = 100\ ms$.

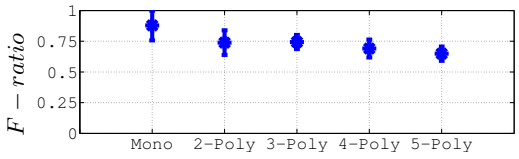


Fig. 5: Mean and s.d. of F-ratios for different polyphony. $N_w = 5\ s$, $N_{sh} = 100\ ms$ and $\epsilon = 0.1$

performance of the proposed approach for varying segment lengths (N_w) and different polyphony using Precision Recall (PR) curves.

For an ideal detection system, the operating point on the PR curve is at $(1, 1)$. For a practical detection system, the point on the PR curve closest to $(1, 1)$ is considered an operating point, as it gives a good trade-off between precision and recall. For a given polyphony, it is observed that PR curves corresponding to higher N_w dominate those corresponding to lower ones. Thus, there is an inherent trade-off between N_w and performance. Note that the PR curves corresponding to $N_w = 5\ s$ and $10\ s$ are close to each other across different polyphonies. Although $N_w = 50\ s$ offers the best performance, we choose $N_w = 5\ s$ for a good trade-off. The threshold (ϵ) corresponding to the operating point is found to be 0.1. Figure 5 shows mean and s.d. obtained by accumulating F-ratios of each $K-$Polyphony, computed over $\binom{M}{K}$ combinations of instruments. Acceptable degradation of performance can be observed with increasing number of instruments in the polyphony.

3.3. Computational Time

The proposed approach is implemented on a Intel(R) Core (TM) 2 Duo CPU T6600 @2.20 GHz and 4 GB RAM. From Figure 1a, the major computational blocks are feature extraction and LV analysis. Polyphonic music signal analysis is done on $N_w = 5\ s$ and $N_{sh} = 100\ ms$. The overall computational time of the algorithm for each segment, is found to be $\sim 85\ ms$ of which, the LV analysis block consumes about 99% of the time. Thus the processing time of the proposed algorithm is less than the data segment rate N_{sh}.

4. DISCUSSION

We propose a generic latent variable framework for multi-instrument recognition using monophonic instrument models. tMMs are used to model individual instruments from their monophonic signals. The polyphonic signal is modeled as a linear combination of LV conditioned individual instrument models. We propose to estimate the LV weights from the polyphonic data using an EM algorithm. Experimental results on eight instruments from RWC database, show monophonic instrument recognition accuracy of 98% indicating the efficacy of tMMs as instrument models. Two to five instrument polyphony are analyzed using PR curves for varying segment lengths. The proposed approach has an average F-ratio of 0.75 for a segment length $N_w = 5\ s$. With the choice of $N_w = 5\ s$, it is shown in Section (3.3) that the proposed algorithm has a processing time of $\sim 85\ ms$. Since the processing time for each segment is less than $N_{sh} = 100\ ms$, the proposed approach can be used for online applications. Instrument Activity Graph (IAG), a graphical display of the LV weights obtained, shows the activity of each of the instruments considered, over time, for a given polyphony.

In any LV decomposition, the definition of latent variable, choice of features, and instrument models depend on what hidden aspects of the mixture signal one wishes to unravel. Other LV decompositions [11, 19] can be seen as a particular case of Equation (1). In particular, the decomposition in [19–21] is obtained by choosing spectral vectors as features, normalized magnitude spectrogram as the mixture signal model and a collection of multinomial distributions in frequency (given the latent variable) as individual source models. In applications demanding signal reconstruction, for instance source separation, the choice of spectrogram as signal model becomes attractive. This is because, spectrogram viewed as normalized histogram, admits decomposition based on non-parametric source models, and signal reconstruction from such a decomposition, although non-trivial, is possible [22]. The choice of spectrogram as a signal model can also be seen in the application of music transcription [11], wherein multiple F0 estimation is essential. Our proposed formulation is generic in the sense that, the type of features or instrument-models are not restrictive. The individual instrument models $p\left(\underline{s}_t^j; \lambda_k\right)$ can be different for different k. Since we are not updating the model parameters in the EM algorithm, it would suffice if these models have a computable expression. Thus, the framework can accommodate heterogeneous models (different models for different instruments) without any modification to the model-weights update equation. This is quite relevant in case of music as melody-based instruments and rhythm-based instruments may require different models.

5. REFERENCES

[1] F. Fuhrmann, "Automatic musical instrument recognition from polyphonic music audio signals," in *PhD Thesis*. Univ. Pompeu Fabra, Barcelona, Spain, 2012.

[2] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. Int'l Symp. on Music Info. Retr.*, pp. 327–332.

[3] K. Kashino and H. Murase, "A sound source identification system for ensemble music based on template adaptation and music stream extraction," in *Speech Comm.*, 1999, vol. 27, pp. 337–349.

[4] P. Smaragdis, M. V. S. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Neural Info. Process. Sys. Conf.*, 2009, pp. 1705–1713.

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

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

[7] F. Fuhrmann, M. Haro, and P. Herrera, "Scalability, generality and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music," in *Proc. Int'l Symp. on Music Info. Retr.*, 2009, pp. 321–326.

[8] J. G. A. Barbedo and G. Tzanetakis, "Musical instrument classification using individual partials," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, no. 1, pp. 111–122, Jan. 2011.

[9] Jun 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. of Select. Topics in Sig. Process.*, vol. 5, pp. 1124–1132.

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

[12] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 16, no. 1, pp. 116–128, Jan. 2008.

[13] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database : Music genre database and musical instrument sound database," in *Proc. Int'l Symp. on Music Info. Retr.*, pp. 229–230.

[14] H. Sundar, T. V. Sreenivas, and W. Kellermann, "Identification of active sources in single-channel convolutive mixtures using known source models," *IEEE Sig. Process. Letters*, vol. 20, no. 2, pp. 153–156, Feb.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy. Stat. Soc., Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[16] 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.

[17] H. Sundar and T. V. Sreenivas, "Robust mixture modeling using t-distribution: Application to speaker ID," in *Proc. Interspeech*, 2010, pp. 2750–2753.

[18] D. Peel and G. J. McLachlan, "Robust mixture modelling using the t distribution," *Statistics and Computing*, vol. 10, no. 4, pp. 339–348, 2000.

[19] 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.

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

[21] P. Smaragdis, B. Raj, and M. V. S. Shashanka, "A Probabilistic Latent Variable Model for Acoustic Modeling," in *NIPS Workshop on Advances in Modeling for Acoustic Processing*, 2006.

[22] S. Nawab, T. Quatieri, and J. Lim, "Algorithms for signal reconstruction from short-time fourier transform magnitude," in *Proc. Intl. Conf. Acoust. Speech and Signal Process.*, Apr 1983, vol. 8, pp. 800– 803.