

ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ/INFORMATICS AND INFORMATION PROCESSES

DOI: https://doi.org/10.60797/IRJ.2026.166.13 EDN: QZQLBX

МЕТОД ИЗВЛЕЧЕНИЯ ИНСТРУМЕНТАЛЬНЫХ ПАРТИЙ ИЗ АУДИОФАЙЛОВ НА ОСНОВЕ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ

Научная статья

Мередова А.[1, *], Тропченко А.А.[2]
[1] ORCID : 0009-0001-1091-0556;
[2] ORCID : 0000-0003-2666-9522;
[1, 2] Университет ИТМО, Санкт-Петербург, Российская Федерация

* Корреспондирующий автор (ayjahanmeredova17[at]mail.ru)

Аннотация

В статье рассматривается задача извлечения инструментальных партий из полифонических аудиозаписей как задача сегментации музыкального потока на участки с устойчивыми статистическими характеристиками, интерпретируемые как тембровые текстуры. Предложен метод, в котором аудиосигнал разбивается на кратковременные кадры длительностью 30 мс с 50%-ным перекрытием, а каждый кадр описывается вектором акустических признаков на основе кратковременных спектральных представлений. Динамика смены текстур моделируется эргодической скрытой марковской моделью с NNN состояниями, при этом распределение наблюдений в каждом состоянии аппроксимируется смесью гауссовских распределений. Оценивание параметров выполняется методом максимального правдоподобия с использованием алгоритма Баума–Уэлша, восстановление последовательности скрытых состояний — методом Витерби.

Для формирования обучающей выборки предложен конвейер подготовки размеченных данных на основе MIDI-представления, обеспечивающий группировку по инструментам и генерацию целевых WAV-файлов. Экспериментальная проверка на многопартийных фрагментах показала согласованность суммарной спектральной структуры извлечённых компонентов со спектром исходной записи при локальных отклонениях в сегментах с неучтёнными источниками. Полученные результаты подтверждают применимость сегментации по тембровым текстурам на основе скрытых марковских моделей для задач анализа музыкального контента в рамках парадигмы Music Information Retrieval.

Ключевые слова: извлечение инструментальных партий, полифоническая музыка, сегментация аудиосигнала, тембровые текстуры, скрытые марковские модели, смесь гауссовских распределений, спектральные признаки, Music Information Retrieval.

A METHOD FOR EXTRACTING INSTRUMENTAL TRACKS FROM AUDIO FILES USING HIDDEN MARKOV MODELS

Research article

Meredova A.[1, *], Tropchenko A.A.[2]
[1] ORCID : 0009-0001-1091-0556;
[2] ORCID : 0000-0003-2666-9522;
[1, 2] ITMO National Research University, Saint-Petersburg, Russian Federation

* Corresponding author (ayjahanmeredova17[at]mail.ru)

Abstract

The article examines the task of extracting instrumental parts from polyphonic audio recordings as a problem of segmenting the musical stream into sections with stable statistical characteristics, which can be interpreted as timbral textures. A method is suggested in which the audio signal is divided into short-term frames of 30 ms duration with 50% overlap, and each frame is described by a vector of acoustic features based on short-term spectral representations. The dynamics of texture changes are modelled by an ergodic hidden Markov model with NNN states, where the distribution of observations in each state is approximated by a Gaussian moxture distributions. Parameter estimation is performed using the maximum likelihood method with the Baum–Welch algorithm, and the hidden state sequence is reconstructed using the Viterbi method.

To generate the training sample, a pipeline for processing annotated data based on MIDI representation has been proposed, which enables grouping by instrument and the generation of target WAV files. Experimental verification on multi-part fragments demonstrated the consistency of the overall spectral structure of the extracted components with the spectrum of the original recording, with local deviations in segments containing uncounted sources. The obtained results confirm the applicability of segmentation by timbral textures based on hidden Markov models for tasks of musical content analysis within the Music Information Retrieval paradigm.

Keywords: extraction of instrumental parts, polyphonic music, audio signal segmentation, timbral textures, hidden Markov models, Gaussian mixture distributions, spectral features, Music Information Retrieval.

Введение

Извлечение инструментальных партий из полифонических аудиозаписей относится к числу методически сложных задач обработки музыкального сигнала вследствие перекрытия гармонических составляющих, вариативности тембра и

существенной зависимости наблюдаемого спектра от динамики исполнения и аранжировки. В прикладном контексте решение данной задачи востребовано при подготовке учебных материалов, анализе исполнения, редактировании аранжировок, а также при автоматизированном индексировании и поиске музыкальных фрагментов в рамках Music Information Retrieval (MIR) [3], [4].

При этом значительная часть существующих подходов ориентирована на выделение мелодической линии или на восстановление доминирующего компонента сигнала, тогда как в ряде практических сценариев первичной становится корректная сегментация аудиопотока на участки с различающимися статистическими свойствами, соответствующими смене инструментальных сочетаний (тембровых текстур). Следовательно, требуется модель, способная одновременно:

а) опираться на информативные спектральные признаки тембра;

б) учитывать временную организацию текстаурных переходов.

Целью исследования является разработка и программная реализация метода извлечения инструментальных партий из аудиофайлов, основанного на сегментации по акустическим текстурам с использованием скрытых марковских моделей и последующим выделением компонент на основе полученной разметки. Научная новизна состоит в формализации полифонического фрагмента как последовательности скрытых текстурных состояний и в использовании эргодической HMM со смесью гауссовских распределений в пространстве спектральных признаков для сегментации, а также в применении конвейера подготовки обучающих данных через MIDI-группировку инструментов [5].

Методы и принципы исследования

Входной аудиосигнал дискретизируется с частотой 10 кГц и представляется в виде последовательности кадров длительностью 30 мс с 50%-ным перекрытием. Выбор оконного разбиения обусловлен стандартным для анализа аудио допущением квазистационарности сигнала на малых интервалах времени, что обеспечивает корректность вычисления кратковременных спектральных характеристик.

Для каждого кадра формируется вектор признаков O_t на основе кратковременного спектрального представления (STS) и связанных с ним тембровых индикаторов, используемых в задачах классификации и сегментации аудио. Признаковое описание ориентировано на фиксацию спектральной структуры тембра и должно быть по возможности менее зависимым от высоты тона, что принципиально для сегментации по текстурам, а не по отдельным нотным событиям.

Смена инструментальных сочетаний в музыкальном фрагменте рассматривается как стохастический процесс, наблюдаемый через последовательность признаковых векторов. Для формализации временной динамики используется эргодическая скрытая марковская модель с N состояниями $\{S_i\}_{i=1}^{N}$. *Каждое состояние интерпретируется как акустическая текстура (включая паузы и типовые сочетания партий), а переходы между текстурами описываются матрицей вероятностей* $A = \{a_{ij}\}$ и начальным распределением π.

Распределение наблюдений в каждом состоянии моделируется смесью гауссовских распределений в пространстве признаков:

$$\rho\left(Q_t \mid S_i\right) = \sum_{m=1}^{M} \mathrm{Cim}\, N\left(Q_t; \mu_{im}\right); \Sigma_{im}.$$

Где C_{im} — коэффициенты смеси, μ_{im} — векторы средних, Σ_{im} — ковариационные матрицы компонент. Применение гауссовых смесей позволяет аппроксимировать неоднородные и потенциально многомодальные распределения признаков, характерные для полифонических текстур.

Параметры HMM оцениваются по критерию максимального правдоподобия на основе наблюдаемой последовательности $\{O_t\}$. Для обучения применяется алгоритм Баума-Уэлша, являющийся EM-процедурой для скрытых марковских моделей [5]. После обучения для восстановления наиболее вероятной последовательности скрытых состояний $\{S_t\}$, породившей наблюдения, используется декодирование Витерби.

В результате каждому кадру сопоставляется состояние S_i, что формирует сегментацию аудиопотока и выделяет временные интервалы однородных текстур.

Для задачи последующей кластеризации/классификации по типу инструмента требуется размеченная выборка. При недостатке открытых наборов данных предлагается конвейер формирования индивидуального датасета, использующий MIDI-представление как носитель структурной информации об инструментах:

1) сбор моноинструментальных аудиозаписей;

2) конвертация WAV → MIDI;

3) выделение и группировка событий по инструментам;

4) генерация отдельных MIDI и их конвертация в WAV.

Такой подход обеспечивает получаемость «квази-эталонных» дорожек, пригодных для обучения и валидации моделей [6], [7], [8].

Метод реализован на языке Python с использованием библиотек обработки аудио и вероятностного моделирования (librosa, numpy, soundfile, wave, hmmlearn) и средств конвертации форматов на базе FFmpeg/ffmpeg-python. Экспериментальные расчёты выполнялись на вычислительной платформе под управлением Windows 10 (Intel Core i5-10210U, 8 ГБ ОЗУ).

Основные результаты

Экспериментальная проверка метода выполнена на музыкальных фрагментах с несколькими инструментальными партиями. Качество оценивалось сопоставлением амплитудных и спектральных характеристик исходного аудио и извлечённых компонент.

Данный подход к проверке обоснован тем, что в случае корректного разложения суммарная спектральная структура извлечённых партий должна воспроизводить ключевые элементы спектра исходной записи, тогда как

отклонения могут указывать на наличие дополнительных источников или на погрешности сегментации/моделирования.

Полученные графики демонстрируют согласованность спектральных характеристик извлечённых партий со спектром исходного аудио на основных участках записи. Локальные расхождения наблюдаются преимущественно в завершающих сегментах, что интерпретируется присутствием в исходном сигнале партий иных инструментов, не включённых в рассматриваемую конфигурацию модели и/или отсутствующих в обучающем наборе.

Обсуждение

Интерпретация результатов подтверждает, что HMM-сегментация по текстурам является адекватным инструментом для случаев, где распределения признаков различных текстур разделимы в признаковом пространстве, а переходы между ними обладают выраженной временной структурой (например, малые ансамбли и фрагменты с устойчивой оркестровкой). В таких условиях вероятностная динамика модели снижает риск фрагментарных ошибок, характерных для статической классификации кадров, и обеспечивает более устойчивую разметку последовательности во времени [8].

Одновременно выявляются ограничения: при усложнении тембровой структуры (оркестровая музыка, плотная современная аранжировка) возрастает внутрисостоянийная вариативность признаков, а спектральная огибающая становится более изменчивой во времени. Это может приводить к снижению согласованности сегментации и к эффектам «антикластеризации», когда признаки приобретают структуру преимущественно за счёт временного порядка, а не за счёт устойчивых статистических различий. В качестве направлений развития целесообразно рассматривать:

а) расширение признакового описания (включая устойчивые тембровые дескрипторы);

б) адаптивный выбор числа состояний (N) и числа компонент смеси (M);

в) введение количественных метрик качества (например, спектральная корреляция, SDR/SIR, либо метрики согласованности разметки по кадрам) для воспроизводимой валидации [9], [10].

Заключение

Разработан метод извлечения инструментальных партий из полифонических аудиофайлов, основанный на сегментации аудиопотока по акустическим текстурам с использованием эргодической скрытой марковской модели и гауссовых смесей в пространстве признаков. Реализованы процедуры обучения (Баум-Уэлш) и восстановления скрытой последовательности состояний (Витерби), предложен конвейер подготовки размеченных данных через MIDI-группировку инструментов. Экспериментальная проверка показала согласованность спектральных характеристик извлечённых партий со структурой исходного аудио при локальных отклонениях в сегментах с неучтёнными источниками. Полученные результаты позволяют рассматривать предложенный подход как модульное решение для задач анализа музыкального контента в MIR и как основу для дальнейшего развития в направлении расширения признаков и формализации количественных критериев качества.

Конфликт интересов

Не указан.

Рецензия

Колмогорова С.С., Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург Российская Федерация, Санкт-Петербургский государственный лесотехнический университет им. С.М. Кирова, Санкт-Петербург Российская Федерация
DOI: https://doi.org/10.60797/IRJ.2026.166.13.1

Conflict of Interest

None declared.

Review

Kolmogorova S.S., Saint Petersburg State Electrotechnical University "LETI" named after V.I. Ulyanov (Lenin), Saint-Petersburg Russian Federation, Saint Petersburg State Forestry University named after S.M. Kirov, Saint-Petersburg Russian Federation
DOI: https://doi.org/10.60797/IRJ.2026.166.13.1

Список литературы на английском языке / References in English

1. Salamon J. Melody Extraction from Polyphonic Music Signals: Approaches, Applications, and Challenges / J. Salamon, E. Gomez, D.P.W. Ellis [et al.] // IEEE Signal Processing Magazine. — 2014. — Vol. 31. — № 2. — P. 118–134. — DOI: 10.1109/MSP.2013.2271648.

2. Durrieu J.-L. Source/Filter Model for Unsupervised Main Melody Extraction from Polyphonic Audio Signals / J.-L. Durrieu, G. Richard, B. David [et al.] // IEEE Transactions on Audio, Speech, and Language Processing. — 2010. — Vol. 18. — № 3. — P. 564–575. — DOI: 10.1109/TASL.2010.2041114.

3. Lee K. A Unified System for Chord Transcription and Key Extraction Using Hidden Markov Models / K. Lee, M. Slaney // Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). — 2007. — P. 245–250.

4. Qian G. A Music Retrieval Approach Based on Hidden Markov Model / G. Qian // Proceedings of the 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). — 2019. — P. 721–725.

5. Chuan C.-H. Polyphonic Audio Key Finding Using the Spiral Array CEG Algorithm / C.-H. Chuan, E. Chew // Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). — 2005. — P. 21–24. — DOI: 10.1109/ICME.2005.1521350.

6. Lee K. Acoustic Chord Transcription and Key Extraction from Audio Using Key-Dependent Hidden Markov Models Trained on Synthesized Audio / K. Lee, M. Slaney // IEEE Transactions on Audio, Speech, and Language Processing. — 2008. — Vol. 16. — № 2. — P. 291–301.

7. Krishna A.S. Identification of Carnatic Raagas Using Hidden Markov Models / A.S. Krishna, V. Ishwar, H.A. Murthy // Proceedings of the IEEE 9th International Symposium on Applied Machine Intelligence and Informatics (SAMI). — 2011. — P. 107–110.

8. Song M. Audio-Visual Based Emotion Recognition Using Tripled Hidden Markov Model / M. Song, C. Chen, M. You // Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). — 2004. — Vol. 5. — P. V-877–V-880.

9. Kogan J.A. Automated Recognition of Birdsong Elements from Continuous Recordings Using Dynamic Time Warping and Hidden Markov Models: A Comparative Study / J.A. Kogan, D. Margolias // The Journal of the Acoustical Society of America. — 1998. — Vol. 103. — № 4. — P. 2185–2196.

10. Katahira K. Complex Sequencing Rules of Birdsong Can Be Explained by Simple Hidden Markov Processes / K. Katahira, K. Suzuki, K. Okanoya [et al.] // PLoS ONE. — 2011. — Vol. 6. — № 9. — e24516.