Реферат

РПЗ n страниц, p рисунка, q таблиц, t источников.

Объектом исследования данной работы являются музыкальные произведения. Целью рабоы является разработка метода автоматизированного выделения голосовой составляющей из музыкальной композиции.

Задачи:

— анализ предметной области;
— анализ методов выделения вокальной партии из музыкальных композиций;
— разработка собственного метода на основании проанализированных;
— разработка программного комплекса, реализующего выбранный метод;
— анализ эффективности работы программного продукта.

Способы применения программного продукта: программный продукт может применяться для дальнейшей разработки выделения компонентов сведенного аудио сигнала, выделения голосовой составляющей из песен для автоматизированного создания текста, выделения голосовой составляющей из аудио дорожки видеофайла для улучшения автоматизированной генерации субтитров.

Содержание

Глоссарий

Дискретизация — Преобразование непрерывного информационного множества аналоговых сигналов в дискретное множесто.

Введение

Тут будет красиво введение.

1 Аналитический раздел

Выделение голосовой составляющей из аудио сигнала является частным случаем задачи разделения комбинации аудио сигналов на составляющие. Музыкальное произведение может быть записано как с использованием нескольких микрофонов, захватывающих разные источники, при этом изоляция источников будет ограничена, либо с использованием выделенной на каждый источник. Все записанные сигналы в последствии проходят процесс сведения для получения итоговой аудио записи. Тем самым конечный аудио сигнал можно представить в виде суммы отдельных источников с приминением фильтров к каждому источнику.

Математически это можно записать как:

$$x(t) = \sum_{j=1}^{J} \sum_{\tau=-\infty}^{\infty} a_j(t-\tau, \tau) s_j(t-\tau) \tag{1.1}$$

где

— x – итоговый аудио сигнал;
— s_j – сигнал источника;
— J – количество источников;
— a_j – фильтр, примененный к источнику в процессе сведения.

Если приминяемые фильтры являются линейными, то итоговый сигнал представляет собой линейную комбинацию. С другой стороны, к итоговому сигналу так же могут быть применены фильтры. С этой точки зрения сведенный сигнал перестает быть линейной суммой отдельных источников.

Тем самым, итоговая запись, в контексте задачи выделения аудио источников, может быть определена по следующим категориям:

а) Недостаточно или черезмерно определенные записи. Зависит от заранее известном числе источников. Например, запись с использованием нескольких микровонов считается черезмерно определенной, а запись с использованием одного микрофона считается недостаточно определенной.

б) Мгновенные или свернутые записи. Зависит от информации о эффектах, примененных к итоговому сигналу.

в) Записи, изменяющиеся или неизменающиеся во времени. Например запись, сделаная с использованием перемещающегося микрофона определяется как изменяющаяся во времени.

Эта работа будет сфокусирована на обработке черезмерно определенной записью, предполагающеся мгновенной и неизменной во времени.

1.1 Цифровое представление аудио сигналов

Звук имеет три основных характеритичесих параметров: амплитуда, частота и фаза. Две последнии характеристики являются функциями времени, в то время как амплитуда определяет динамический диапозон. По этому в цифровом представлении аудио сигнала записываются изменения аплитуды как функция времени. Тем самым процесс отцифровки аналогового аудио сигнала является записью мгновенных амплитудных значений (дискретизация по амплитуда) в постоянные моменты времени (дискретизация по времени). Графическое представление дискретизации аналогового сигнала представлено на рисунке 1.1.

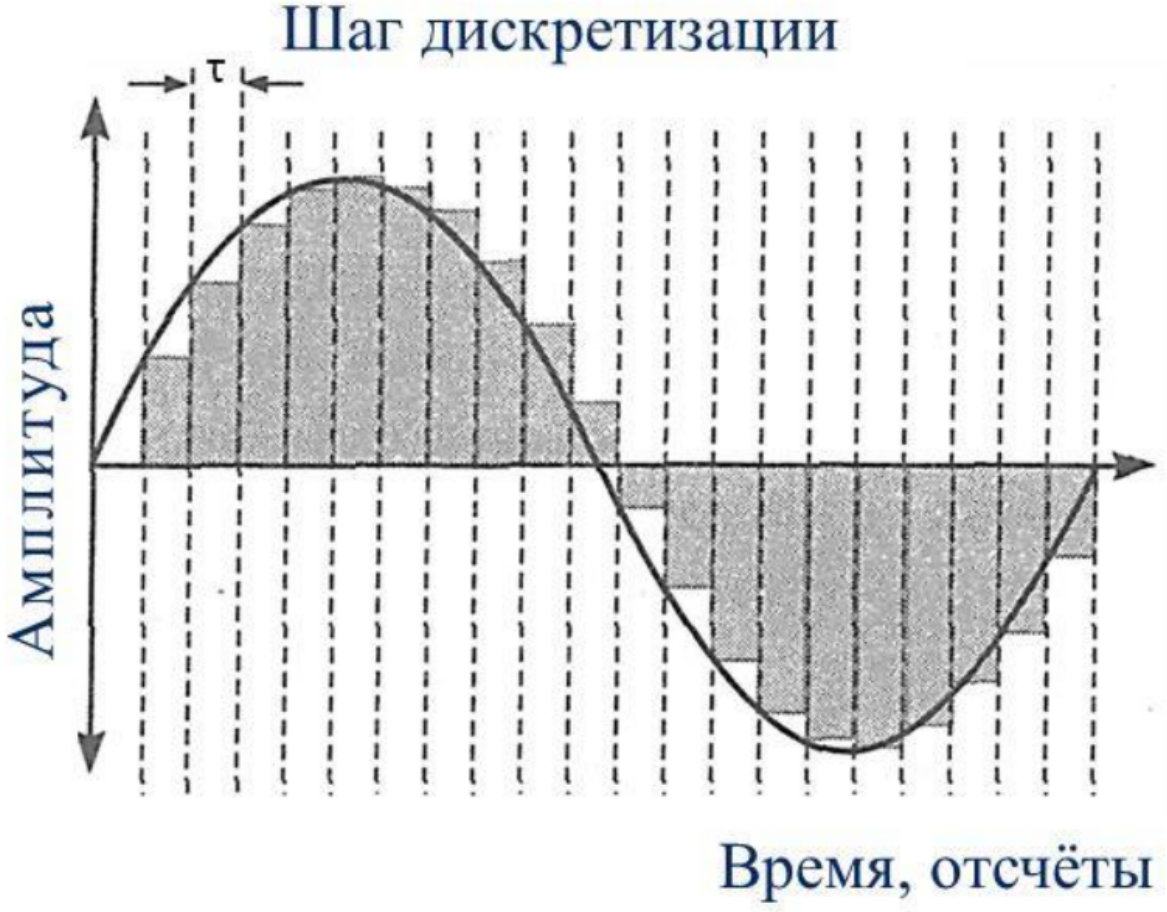

Рисунок 1.1 — Дискретизация аналогового сигнала

Для определения периода записи амплитудных значений задается частота дискретизации.

Существует теорема Котельникова[1], утверждающая, что «любую функцию $F(t)$, состоящую из частот от 0 до f_1, можно непрерывно передавать с любой точностью при помощи чисел, следующих друг за другом через $1/(2f_1)$ секунд».

При максимальной воспринимаемой человеческим ухом частоте в 20 кГц, по теореме Котельникова минимальная необходимая частота дискретизации должна быть 40 кГц.

Стандартная частота дискретизации аудио сигнала составляет 44,1 кГц, максимальная – 192 кГц.

Для определения максимального значения амплитуды используется значение квантования (или разрядность), задающаяся в битах. В зависимости от используемого формата разрядность может быть 1, 8, 16, 24 и 32 бит.

Сравнение существующих форматов хранения цифрового аудио сигнала по квантованию и частоте дискретизации представлено в таблице 1.1.

Таблица 1.1 — Сравнение цифровых аудиоформатов

Название формата	Квантование, бит	Частота дискретизации, кГц
WAVE (WAV)	8; 16; 24; 32	любая
AIFF	8; 16; 24; 32	11,025; 22,05; 24; 32; 44,1; 48; 96; 192
WavPack	8; 16; 24; 32	6 – 192
FLAC	4 – 32	до 192
Lossless Predictive Audio Coder (LPAC)	8; 16; 20; 24;	до 192
LosslessAudio (LA)	16	48
Windows Media Audio 9 Lossless	16; 24	8; 11.025; 16; 22.05; 32; 44.1; 48; 88.2; 96
Apple Lossless (ALAC, ALE)	16; 24	44.1; 48; 88.2; 96; 192
MP3	16; 24	до 48

1.2 Существующие методы выделения источников

1.2.1 Метод главных компонент

Метод главных компонент (англ. Principal Component Analysis) и анализ независимых компонент (англ. Independent Component Analysis) – это статистические способы уменьшения размерности данных, который являлся объектом для исследований в области выделения аудио источников [2] [3].

Основная идея этоих метода заключается в том, чтобы проецировать данные из временных рядов, таких как аудиозапись, в новые системы отсчета, которые основаны на некотором статистическом критерии. Эти оси являются статистически независимые в отличие от преобразования Фурье, где данные временной области проецируются на оси частот, которые могут перекрываться. Частотные оси в преобразовании Фурье остаются неизменными независимо от анализируемой части, тогда

как в методе главных компонент и анализе независимых компонент оси являются динамическими и различны для каждой анализируемой части. После нахождения, оси, на которые происходит проецирование, могут быть разделены и инвертированы для нахождения источников, представленных в исходном сигнале.

В методе главных компонент мерой разделения осей является дисперсия. Оси рекурсивно выбираются в качестве направления, в которых дисперсия сигнала максимальна, что приводик к декорреляции во втором порядке между осями. Тем самым, основные компоненты амплитудной характеристики сигнала находятся на первых осях. Данный метод может быть использован как для сжатия информации, так и для выделения источников.

В анализе независимых компонент, четвертый момент, называющийся коэффициент эксцесса, используется в качестве критерия для нахождения новых систем отсчета. В теории вероятностей коэффициент эксцесса является мерой остроты пика распределения случайной величины или показателем «негуссовости» сигнала. Негативное значение коэффициента означает, что функция ресрпежедения вероятностей шире Гауссовского распределения, в то время как положительное значение – уже. Тем самым в анализе независимых компонент сигнал разделяется на негауссовские независимые сигналы, а в методе главных компонент сигнал разделяется на гауссовские независимые сигналы.

Задача сводится к серии матричных произведений, представляющих свобой фильтры. В общем случае входной сигнал X размерности N из M образцов (представляется матрицей размерности NxM) может быть приведен к сигналу Y с использованием матрицы преобразования W размерности NxN как $Y^T = WX^T$. Такое преобразование проецирует сигнал на разные оси основываясь на матрице преобразования. Если размер полученного сигнала равен размер исходного сигнала, то преобразование называется ортогональным и оси перпендикулярны.

Эти операции называются преобразованиями без потерь из-за того, что исходный сигнал может быть восстановлен без потерь информации. Если один или несколько столбцов матрицы являются нулевыми, то такие операции называются преобразованиями с потерями и используются для фильтрации и сжатия данных. Преобразование считается биортогональным, когда исходная и результирующая оси перпендикулярны. Такое преобразование является без потерь. Метод главных компонент и анализ независимых компонент являются ортогональными и биортогональными соответственно.

Главной задачей яляеется получение матрицы преобразования.

1.2.2 Факторизация неотрицательных матриц

Факторизация неотризательных матриц (англ. Non-Negative Matrix Factorization) широко использовалась в области выделения источников в прошлом. Основная идея данного метода заключается в представлении матрицы Y в виде комбинации базиса B и активационного усиления G как $Y = BG$. Базовый вектор предстовляет частотную характеристику источника в заданный момент времени, а вектор G представляет усиление частот в любой момент времени. Таким временем G является горизонтальным вектором вдоль времени.

В контексте задачи выделения аудио источников, если исходный сигнал является объединением двух источников, S_1 и S_2, так, что $Y = S_1 + S_2$, и базисные вектора двух источников вычисляются как B_1 и B_2, то исходный сигнал можно представить как $Y = B_1G_1 + B_2G_2$, где G_1 и G_2 – соответсвующие активанционные усиления для двух источников, представленные в разные моменты времени.

Для применения факторизации неотрицательных матриц необходимо, чтобы сигнал представлял из себя линейную комбинацию базисных векторов. Для K источников:

$$X_{i,j} = \sum_{k=1}^{K} B_{i,k}G_{k,j} \tag{1.2}$$

Расхождение между X и BG должно быть минимизировано, чтобы гарантировать, что найденные аудио источники представляют в комбинации исходный сигнал:

$$B, G = argmin_{B,G>=0} D(X, BG) \tag{1.3}$$

где D является функцией расхождения, которая может быть:

а) евклидовым расстоянием:

$$D(A, B) = ||A - B||^2 = \sum_{ij}(A_{ij} - B_{ij})^2 \tag{1.4}$$

б) дивергенцией Кулбека-Лейблера:

$$D(A, B) = \sum_{ij}(A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}) \tag{1.5}$$

Для этой дивергенции применяется алгоритм мультипликативного обновления [4]:

а) Вектора B и G заполняются случайными значниями.

б) Вычисление нового значение B:

Для евклидова расстояния:

$$B \leftarrow B\frac{XG^T}{(BG)G^T} \tag{1.6}$$

Для дивергенции Кулбека-Лейблера:

$$B \leftarrow B\frac{(\frac{X}{BG})G^T}{1G^T} \tag{1.7}$$

в) Вычисление нового значения G:

Для евклидова расстояния:

$$G \leftarrow G\frac{B^T X}{B^T(BG)} \tag{1.8}$$

Для дивергенции Кулбека-Лейблера:

$$G \leftarrow G\frac{B^T\frac{X}{BG}}{B^T 1} \tag{1.9}$$

Эти действия посторяются предопределенное количество раз, либо до момента достижения дивергенции определенного минимума.

После нахождения магнитуды источника, сигнал можно получить при помощи фазовой характеристики исходного сигнала.

Работа данного метода зависит от задаваемых начальных условий, зависимых от числа источников и фильтров, примененных к итоговому сигналу.

1.2.3 Гибкие инструменты для выделения аудио источников

Гибкие инструменты для выделения аудио источников (Flexible Audio Source Separation Toolbox) являются модификацией метода факторизации неотрицательных матриц, основанный на обобщенном EM-алгоритме (англ. Generalized expectation-maximization (GEM) algorithm) [5].

Суть данного метода заключается в обобщении реализации, гибкой для разных случаев начальных условий. Каждый выделяемый источник представляется в виде модели возбуждающего фильтра:

$$V_j = V_j^{ex} \bigodot V_j^{ft} \tag{1.10}$$

где V_j^{ex} представляет спектральный возбудитель мощности и V_j^{ft} представляет спектральный фильтр мощности, $\bigodot$ означает поэлементное умножение матриц.

В дальнейшем V_j^{ex} определяется в качестве суммы спектральных шаблонов E_j^{ex}, модулированных временными коэффициентами P_j^{ex}. Эти параметрами являются аналогами базисного вектора и активационного усиления в методе факторизации неотрицательных матриц. Тем самым V_j^{ex} можно переписать как:

$$V_j^{ex} = E_j^{ex} P_j^{ex} \tag{1.11}$$

Спектральные шаблоны представляются в виде линейной комбинации произведения узкополосных спектральных шаблонов W_j^{ex} и неотрицательных весов U_j^{ex}.

Временные коэффициенты представляются в виде линейной комбинации произведения кратковременных шаблонов H_j^{ex} и G_j^{ex}.

В итоге можно записать:

$$V_j = (W_j^{ex} U_j^{ex} G_j^{ex} H_j^{ex}) \bigodot (W_j^{ft} U_j^{ft} G_j^{ft} H_j^{ft}) \tag{1.12}$$

Эти шаблоны оцениваются для каждого источника с помощью GEM алгоритма, с использованием алгоритма мультипликативного обновления [4].

1.2.4 Методы, использующие нейронные сети

Идея применения методов машинного обучения в задачах разделения аудио источников возникла после успешных использований нейронных сетей в других областях, в частности обработки изображений [6]. Для решения данной задачи наиболее широко рассматривались глубокие[7], сверточные[8] и рекуррентные[9] нейронные сети.

Список использованных источников

1. *Биккенин, Р. Р.* Теория электрической связи / Р. Р. Биккенин, М. Н. Чесноков. — Издательский центр «Академия», 2010. — P. 336.

2. Blind Source Separation of audio signals using independent component analysis and wavelets / P. Lopez, H. Molina Lozano, F. Sanchez, L. N. Oliva Moreno // *21st International Conference on Electrical Communications and Computers*. — 2011.

3. *Dadula, C. P.* A genetic algorithm for blind source separation based on independent component analysis / C. P. Dadula, E. P Dadios // *International Conference on Humanoid, Nanotechnology, Informa- tion Technology, Communication and Control, Environment and Management*. — 2014.

4. *Lee, D. D.* Algorithms for Non-negative Matrix Factorization / D. D. Lee, Seung H. S. // *Advances in Neural Information Processing Systems*. — 2001.

5. *Ozerov, Alexey.* A General Flexible Framework for the Handling of Prior Information in Audio Source Separation / Alexey Ozerov, Emmanuel Vincent, Frédéric Bimbot // *IEEE Transactions on Audio, Speech and Language Processing*. — 2012.

6. *Krizhevsky, Alexey.* ImageNet Classification with Deep Convolutional Neural Networks / Alexey Krizhevsky, Ilya Sutskever, Geoffrey E. Bimbot // *Advances in Neural Information Processing Systems*. — 2012.

7. *Grais, Emad M.* Deep neural networks for single channel source separation / Emad M. Grais, Mehmet Umut Sen, Hakan Erdogan // *Speech and Signal Processing*. — 2014.

8. *Chandna, Pritish.* Audio Source Separation Using Deep Neural Networks / Pritish Chandna // *TESI DOCTORAL UPF*. — 2016.

9. Singing-Voice Separation From Monaural Recordings Using Deep Recurrent Neural Networks / Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, Paris Smaragdis // *Ismir*. — 2014.