Московский технический университет связи и информатики

ТЕХНИКА ЗВУКОВОГО ВЕЩАНИЯ

УДК 577.4:621.396

А.А. Кудинов

АВТОМАТИЧЕСКАЯ СЕГМЕНТАЦИЯ МНОГОГОЛОСНЫХ МЕЛОДИЙ

Введение

Автоматическая сегментация является составной частью задачи распознавания звуковых сигналов. Сегментация всегда предшествует собственно распознаванию, т.е. принятию решения о принадлежности реализации объекта к тому или иному классу. Соответственно, достоверность распознавания зависит, в том числе, и от достоверности сегментации.

Из-за различий в свойствах звуковых сигналов, подлежащих распознаванию в современной технике, — а это не только речевые и музыкальные сигналы, но также шумы всевозможных машин, узлов машин (например, двигателей), — специфичны и подходы к сегментации сигнала. Выбор подхода к сегментации определяется, в первую очередь, выбором *элемента распознавания*, т.е. элементарного с точки зрения системы распознавания звукового объекта, подлежащего распознаванию. При распознавании слитной речи элементом распознавания может быть и фонема, и слог и целое слово. Задача алгоритма сегментации — определение границ элементов распознавания и при необходимости их разделение. В случае, если элементы распознавания перекрываются во временной области, необходимо определение границ в частотной области.

Случай распознавания одноголосной мелодии самый простой — звуковые объекты не перекрываются, время затухания одного звукового объекта совпадает со временем возникновения следующего. Следует отметить, что пауза также является звуковым объектом, не имеющим, правда, высоты тона, но имеющим длительность, и в нотной записи паузы обозначаются специальными символами. Иными словами, одноголосная мелодия представляет собой линейную последовательность звуковых объектов. Распознавание сводится к определению моментов смены высоты тона, сведения об интенсивности колебаний используются лишь в случае разделения двух последовательных нот с одинаковой высотой тона. Поэтому в иностранной литературе распознавание одноголосных мелодий общепринято называют «слежением за мелодией основного тона» (fundamental frequency tracking).

В речевом сигнале звуковые объекты также не перекрываются (если это монолог), однако определение их границ затруднено по сравнению с одноголосной мелодией. Дело в том, что для перехода от одной фонемы к другой необходима перестройка голосового аппарата: и связок, и горла, и ротовой полости. Звуковые сигналы, сформированные на стадии перестройки, не классифицируются как фонемы, поскольку свойства этих сигналов устойчиво не повторяются, являясь «побочным продуктом» речеобразования они служат для соединения фонем.

Основная сложность распознавания многоголосных музыкальных сигналов определяется тем, что звуковые объекты перекрываются. Перекрываются звуковые объекты во времени, т.е. перекрываются интервалы существования звуковых объектов, моменты их возникновения могут совпадать, за время существования одного звукового объекта могут возникнуть и затухнуть ещё несколько. Перекрываются звуковые объекты и в частотной области, т.е. звуковые объекты могут иметь спектральные компоненты с примерно совпадающими частотами. Сложнейший вариант — т.н. исполнение в унисон, когда несколько инструментов одновременно исполняют одну и ту же мелодию. Перекрытие почти стопроцентное. Тем не менее, в случае, если тембры инструментов различны, человеческое ухо в некоторых случаях способно определить их количество.

После выполнения задачи сегментации образуется ритмическая структура распознаваемой мелодии. Ритмическая структура мелодии не менее важна, чем её высотная структура. В этом состоит одно из отличий задачи распознавания музыкального сигнала от распознавания речевого сигнала. При распознавании речи ошибки сегментации могут привести лишь к снижению достоверности распознавания фонемы, возникшие ошибки можно исправить автоматически, имея словарь языка. Ошибки сегментации музыкального сигнала повлекут и неверное определение высоты тона (если на вход алгоритма определения частоты основного тона попадут фрагменты последовательных звуковых объектов с разной высотой тона), и искажение ритмической структуры, в результате чего мелодия может стать неузнаваемой.

1. Современные подходы к сегментации музыкальных сигналов

Перечисленные отличия полифонических музыкальных сигналов от одноголосных не позволяют применять для сегментации алгоритмы, разработанные для одноголосных мелодий. Для сегментации линейных последовательностей звуковых объектов первоначально применялся *последовательный* спектральный анализ. Т.е. распознаваемый сигнал считывается короткими выборками, на которых выполняется спектральный анализ. Анализируемые выборки могут перекрываться, но анализатор

спектра один, и анализ производится во всей полосе частот. Современные алгоритмы сегментации музыкальных сигналов строятся на использовании *параллельного* спектрального анализатора. Как показывает обзор публикаций зарубежных исследователей за последние 5-7 лет, работы по созданию систем распознавания музыкальных сигналов ведутся в рамках решения более общей задачи — компьютеризированного анализа звуковых сцен (CASA – Computational Auditory Scene Analysis). К решаемым CASA задачам относятся пространственная локализация и идентификация источников звука. Бурное развитие CASA, видимо, вызвано развитием многоканальных систем звукозаписи, потребовавшее новых исследований психофизиологии слухового восприятия. По сути, CASA строится и в первую очередь занимается созданием модели слухового восприятия человека. По современным представлениям слуховой аппарат человека снабжён именно *параллельным* анализатором спектра — волокна покровной мембраны восприимчивы лишь к колебаниям своей узкой полосы частот, при этом информация от каждого волокна передаются в мозг отдельным нейроном. Представляется очевидным разделение звуковых объектов, перекрывающихся во времени, именно в частотной области. Работа волокон покровной мембраны моделируется с помощью набора фильтров. Сами по себе отклики фильтров, как правило, не используются при вынесении решения о появлении или затухании звуковых объектов. Для принятия решения сопоставляются различные параметры откликов фильтров, различные алгоритмы сегментации отличаются выбором этих параметров и схемой принятия решения.

Существует, правда, подход к сегментации звуковых сигналов, явно не использующий набор фильтров, но основанный на параллельном спектральном анализе. Речь идёт об использовании для распознавания и цифровой обработки сигнала *сонограммы (спектрограммы)*, т.е. временной зависимости коэффициентов кратковременного дискретного преобразования Фурье (ДПФ). Сонограмма используется для визуализации результатов спектрального анализа звуковых фрагментов большой продолжительности. Именно возможность проследить во времени изменения амплитуд коэффициентов ДПФ позволяет выделить и отнести к одному звуковому объекту «траектории» спектральных компонент (см. рис. 1). При таком подходе распознавание звуковых образов заменяется распознаванием зрительных образов. Амплитуды коэффициентов ДПФ отражают распределение энергии по узким полосам частот. По сути, временную зависимость амплитуды отдельного члена ряда можно рассматривать как огибающую отклика простого БПФ фильтра, АЧХ которого описывается символом

Кронекера: $\delta(i-k)=\begin{cases}1, i=k\\0, i\neq k\end{cases}$. На рисунке 1 моменты появления звуковых объектов отмечены пунктирными линиями, параллельными оси ординат. Изображена последовательность 3 объектов: одиночная нота («ля»), созвучие «си» + «ре» и аккорд до мажор («до» + «ми» + «соль»).

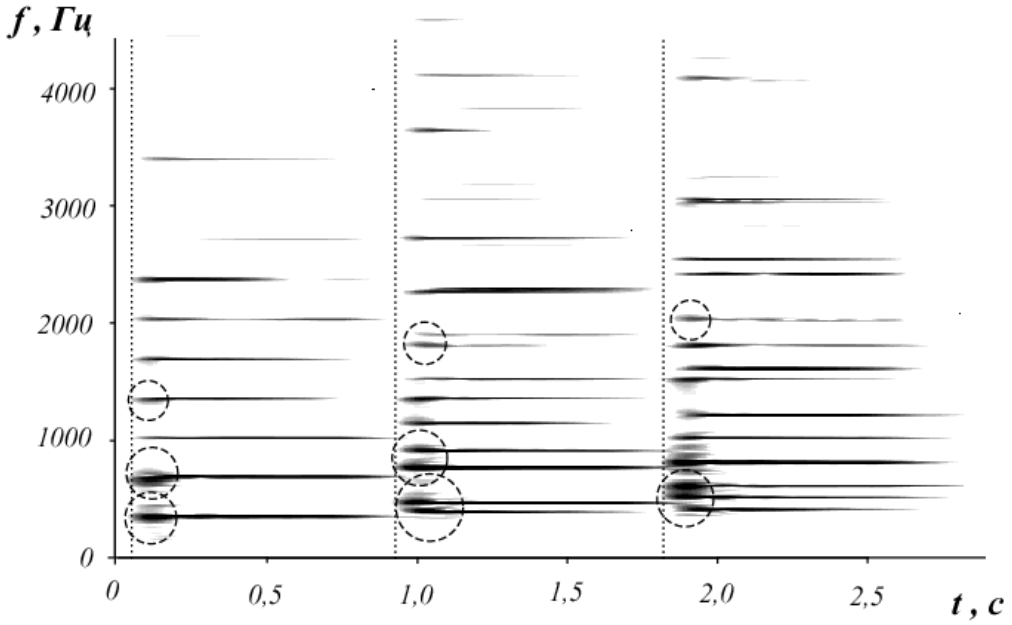


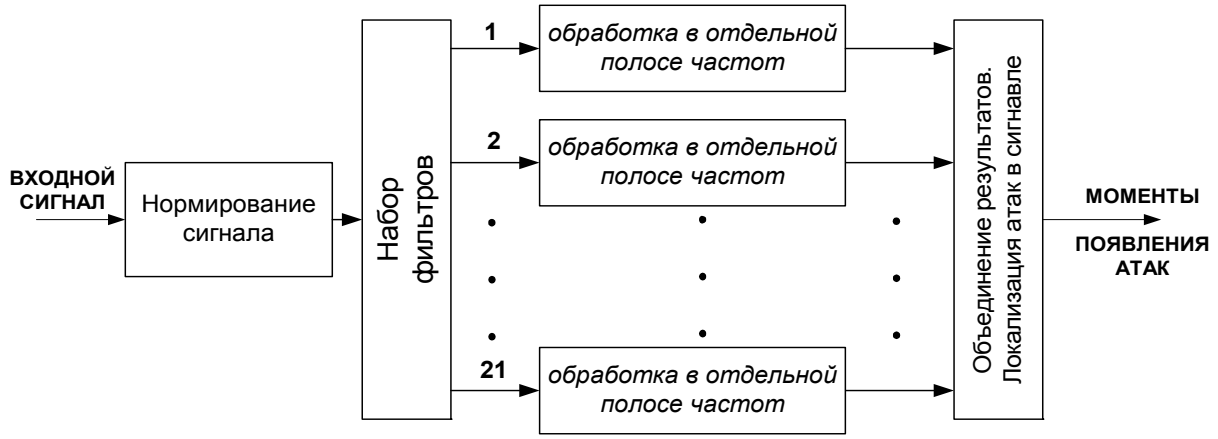
Рис 1. Спектрограмма музыкального сигнала

Штриховыми окружностями обозначены утолщения в началах траекторий гармонических составляющих. Стадия возбуждения спектральных компонент — стадия нестабильности параметров колебаний, имеют место частотная и амплитудная модуляции, что приводит к обогащению спектра дополнительными составляющими (боковые полосы по аналогии со спектром АМ- или ЧМ-сигнала). На сонограмме эти боковые полосы образуют утолщения траекторий, таким образом, образуется признак, важный при распознавании зрительных образов. На рисунке 1 хорошо видно, что моменты возбуждения и затухания обертонов и основной частоты не совпадают. Это создаёт дополнительные сложности при описываемом подходе, поскольку для распознавания необходимо установить принадлежность отдельной спектральной составляющей какому-либо звуковому объекту, иначе говоря, объединить траектории обертонов и основной частоты для каждой ноты.

Сегментация с использованием сонограммы не нашла пока широкого применения. Большего внимания заслуживает другой подход, позволивший добиться существенных результатов. Этот подход, уже упомянутый, основан на модели восприятии громкости человеческим ухом, подробное описание которой дано в [2]. Последовательность процессов обработки и анализа сигнала представлена на рисунке 2.

Все входные сигналы должны быть приведены к одному определённому значению относительной громкости, для этого перед обработкой сигнал нормируется. Нор-

а)



б)

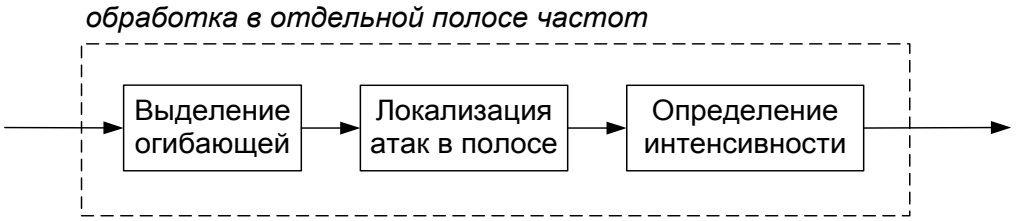


Рис. 2 Схема сегментации звукового сигнала моделированием слухового восприятия

мированный сигнал поступает на набор полосовых фильтров. Частотный диапазон от 40 Гц до 18 кГц делится на полосы таким образом, чтобы ширина каждой полосы была примерно равна ширине критической полосы слуховой системы. Первые три фильтра (в самой низкочастотной области) — октавные, остальные — третьоктавные. Для перекрытия указанного диапазона частот достаточно 21 фильтра. В системе сегментации, описанной в [1], применены рекурсивные фильтры 8-го порядка, в системе, описанной в [5] — цифровые аналоги эллиптических фильтров 6-го порядка, весь диапазон частот делится на 6 полос. После разделения все узкополосные сигналы (компоненты) одинаковым образом обрабатываются (см. рис. 2 б)). Прежде всего, выделяется огибающая узкополосного сигнала, сглаживаемая свёрткой с временной оконной функцией Хэмминга («косинус на пьедестале») длительностью 100 мс, чем моделируется интеграция энергии слуховой системой человека. Далее для каждого компонента определяются моменты появления атак. Моменты появления атак компонентов в дальнейшем рассматриваются как кандидаты при выборе моментов появления атак

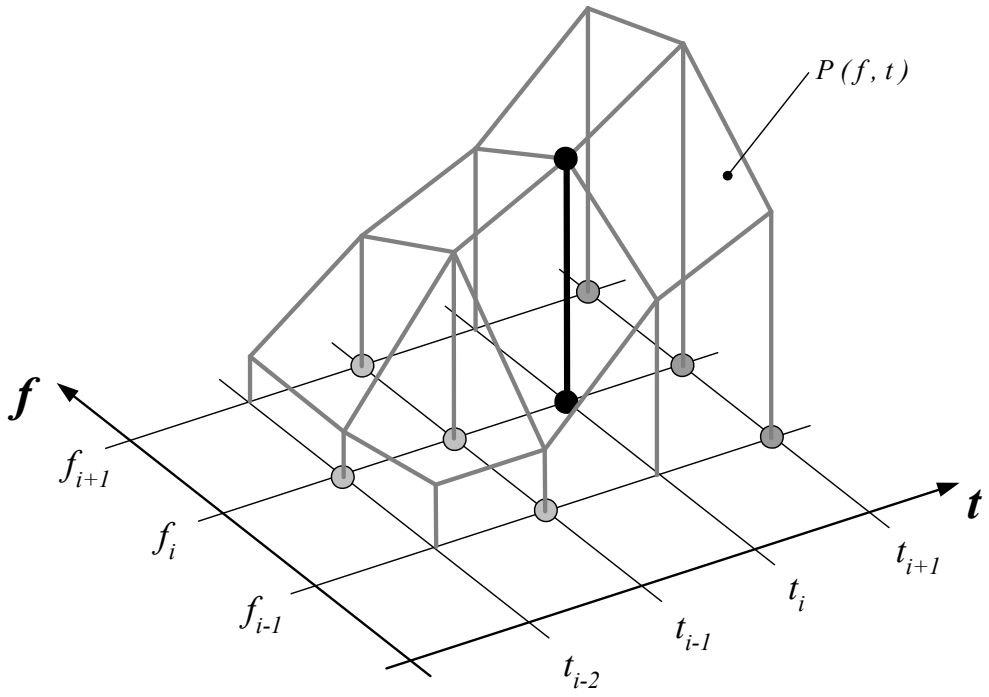


Рис. 3 Определение момента появления атаки по мощности сигнала (по [4])

сигнала как единого целого. Один из алгоритмов обнаружения атак в компонентах иллюстрируется рисунком 3 (подробно этот алгоритм описан в [3] и [4]). Решение о появлении атаки в сигнале-компоненте выносится при сравнении мощностей сигналов соседних частотных полос в соседние с текущим моменты времени. Так, решение о появлении атаки в момент времени t в компоненте с центральной частотой f_i принимается, если выполняются два условия:

$$\begin{cases} P(f_i, t_i) > PP \\ PP > NP \end{cases}, \text{ где}$$

$$PP = \max \left\{ P(f_i, t_{i-1}), P(f_{i\pm1}, t_{i-1}), P(f_i, t_{i-2}) \right\} \qquad (1)$$

$$NP = \max \left\{ P(f_i, t_{i+1}), P(f_{i\pm1}, t_{i+1}) \right\}$$

$P(f_i, t_i)$ — мощность сигнала полосы с центральной частотой f_i на выборке с начальным временем t_i. PP (“previous power”) — мощность на предыдущей временной выборке, NP (“next power”) — мощность на следующей временной выборке. В алгоритме сегментации, описанном в [5], как и во многих других алгоритмах, за моменты появления атак в узкополосных сигналах принимаются моменты времени, соответствующие максимальной скорости нарастания огибающей. Находя локальные максимумы первой производной огибающей $A(t)$, находят моменты появления атак.

Однако, у такого подхода есть важный недостаток. Как указывается в [1], функцию $\frac{dA(t)}{dt}$ удобно использовать для оценки интенсивности атаки, но локальные максимумы этой функции не всегда точно соответствуют истинному моменту атаки. Основных причины две:

1. низкочастотные компоненты нарастают медленнее, чем высокочастотные, в результате момент времени, соответствующий наибольшей скорости нарастания огибающей узкополосного сигнала, сильно отстаёт от субъективно воспринимаемого момента возникновения звукового объекта;
2. не все узкополосные компоненты имеют монотонно нарастающую и спадающую огибающую, вблизи наибольшего локального максимума функции $\frac{dA(t)}{dt}$ имеются другие локальные максимумы(см. рис. 4), что приводит к ошибочным решениям при объединении промежуточных результатов всех компонентов.

Для избежания последствий указанного недостатка в [1] предлагается искать ло-

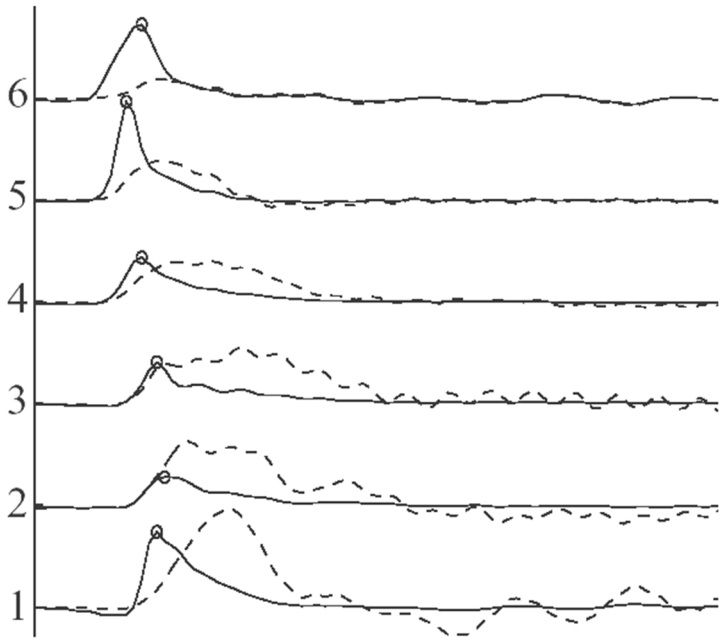


Рис. 4 Атака звука фортепиано: штриховые линии — производные огибающих узкополосных компонентов, сплошные линии — производные логарифмов огибающих ([1], рис. 2). Цифрами обозначены номера частотных полос

кальные максимумы функции $\frac{\frac{dA(t)}{dt}}{A(t)}$, т.е. относительного приращения огибающей. Такой подход согласуется с установленным свойством и слуховой, и зрительной систем

человека реагировать именно на относительные приращения интенсивностей воздействия. Известно, что $\frac{d}{dx}(\ln(F(x)) = \frac{1}{F(x)} \cdot \frac{dF(x)}{dx}$, поэтому $\frac{\frac{dA(t)}{dt}}{A(t)} = \frac{d(\ln(A(t))}{dt} = W(t)$. Функция $W(t)$ используется и при обнаружении атак узкополосных компонентов и при определении моментов появления этих атак, при этом достаточно простого нахождения локальных максимумов, поскольку функция $W(t)$ более гладкая, чем $A'(t)$. Рисунок 4 иллюстрирует преимущество использования при сегментации функции $W(t)$ перед использованием $\frac{dA(t)}{dt}$.

Определив моменты появления атак в компонентах можно переходить к локализации атак обрабатываемого сигнала. Атаки компонентов являются компонентами атаки сигнала, на основе соотношений моментов появления атак компонентов и их интенсивностей принимается решение о наличии атаки в сигнале. Интенсивности компонентов определяется следующим образом:

$$I_i(t) = \frac{dA_i(t)}{dt} \cdot \Delta f_i = D_i(t) \cdot \Delta f_i, \qquad (2)$$

т.е. производная огибающей i-того компонента $D_i(t)$ умножается на ширину спектра компонента Δf_i (т.е. ширину полосы пропускания i-того фильтра). Интенсивности атак, появляющихся одновременно в разных компонентах суммируются. Далее атаки всех компонентов сортируются по времени появления и рассматриваются в дальнейшем как кандидаты при выборе момента появления атаки в сигнале. Каждому такому кандидату ставится в соответствие величина *относительной громкости*. Напомним, что перед разделением на узкие частотные полосы входной сигнал нормировался, т.е. приводился к определённому значению относительной громкости. Такое нормирование необходимо по одной причине: при сегментации необходимо иметь точку отсчёта энергетических параметров сигнала и его компонентов, например для игнорирования атак с интенсивностями ниже условного порога слышимости. Вычисление относительной громкости кандидата производится в соответствии с моделью слухового восприятия, описанной в [2]. При вычислении громкости кандидата учитываются интенсивности атак компонентов, находящихся во временном отрезке 50 мс с серединой в моменте появления данной атаки-кандидата. Последним этапом является устранение «лишних» атак, т.е. атак с громкостями ниже условного порога слышимости и условно

маскируемых атак. За маскируемые атаки принимаются атаки с меньшей относительной громкостью, находящиеся к атаке с большей относительной громкостью ближе, чем 50 мс. При рассмотрении близко расположенных атак-кандидатов, имеющих равные относительные громкости, за истинную атаку принимается средний по времени кандидат, остальные игнорируются.

Описанный подход можно назвать «сегментацией по громкости». Если быть точным, то рассмотренные в [1]-[5] алгоритмы выполняют не сегментацию в том смысле, в котором договорились понимать её мы. В англоязычной литературе описанная операция называется *sound* ***onset*** *detection*, что дословно означает *определение момента* ***начала*** *звучания звука*. Т.е. Определяются лишь моменты появления звуковых объектов, но не моменты их затухания. Необходимо заметить, что описанный алгоритм, как и многие другие, разрабатывался в рамках построения систем выделения ритма, ритмической структуры сигнала. Действительно, для этого достаточно определения лишь моментов начала звучания звуковых объектов, однако при распознавании музыкального сигнала необходимо определение не только моментов появления нот, но и их длительностей, что требует определения моментов затухания звуковых объектов, выделения и определения длительностей пауз. Важной особенностью описанного подхода является его универсальность. Определение моментов появления звуковых объектов может проводиться в сигналах любого типа звучания: инструментальной, камерной, симфонической, эстрадной или рок-музыке, речевом сигнале. Однако типом звучания определяется и достоверность определения ритма. По данным автора [1], достоверность работы его системы колеблется от 95% для фортепианной музыки (Шопен) до 7% для симфонической музыки (Бетховен). Такой разброс и есть плата за отсутствие специализации, построение алгоритма, не учитывающего особенности конкретного типа звучания.

2. Сегментация музыкального сигнала с использованием априорной информации

Развитие систем автоматической сегментации утвердило два принципа, на которых и строятся все современные системы:

1. необходимым и основным параметром сигнала, по которому осуществляется сегментация, является его огибающая;
2. для сегментации полифонических мелодий необходимо разделение сигнала на узкополосные компоненты.

Как правило, огибающая не является единственным информативным параметром при сегментации. Очевидно, что при переходе от одного звукового объекта изме-

няются и энергетические, и спектральные параметры сигнала, нестабильны эти параметры и на длительности отдельного звукового объекта. Выделение огибающей сигнала не требует сложных операций и вычислений, но при этом огибающая не всегда позволяет однозначно принять решение о появлении, а тем более затухании звукового объекта. Огибающая широкополосного многокомпонентного сигнала может вообще оказаться непригодной для сегментации: например, тихие пассажи в высоком диапазоне высот тонов могут совершенно не оставить в огибающей следа на фоне мощного длительного низкочастотного звука или его затухающего отзвука. Огибающие узкополосных составляющих сигнала оказываются более информативными при сопоставлении и объединение результатов разных частотных полос. При разделении спектра сигнала на узкие полосы количество звуковых объектов и их составляющих, модулирующих энергию колебаний в данной полосе, уменьшается с уменьшением ширины полосы, поэтому становится легче обнаружить «следы» отдельных звуковых объектов, перекрывающихся во времени.

При сегментации одноголосных мелодий, как мы говорили, достаточно слежения за изменениями основного тона за исключением случая следования друг за другом двух одинаковых нот. Выделение основного тона в полифонических мелодиях значительно усложняется частичным перекрытием обертонов звуковых объектов. При сегментации музыкального сигнала желательно разделение не по частоте, а *по периоду*. Иными словами, желательно создание такой системы фильтров, чтобы интенсивность отклика каждого фильтра была тем больше, чем ближе период входного сигнала к значению периода, на которое настроен этот фильтр. Конечно, это идеализация, но подумаем, как к этой идеализации можно приблизиться. Для музыкального сигнала справедливы два утверждения:

1. созвучия, состоящие из нот, отстоящих на целое число октав, появляются в сигнале реже, чем терции, кварты или квинты, при этом пока не разработаны алгоритмы, позволяющие достоверно разделять ноты, отстоящие на октаву;
2. музыкальный сигнал имеет более строгую, чем речевой сигнала, организацию спектральной структуры: понятие музыкального строя подразумевает определённое соотношение высот тонов звуков, выраженность ощущения высоты тона означает определённое соотношение частоты основного тона и частот обертонов.

Втрое утверждение означает, что при распознавании любого музыкального сигнала мы всегда обладаем априорной информацией о нём. Например, если известен музыкальный строй инструментов, участвующих в исполнении, а также основной тон

эталона, по которому ансамбль (инструмент) настраивался, автоматически становятся известны примерные значения частот основных тонов и обертонов, соответствующие любой ноте, которую может исполнить данный ансамбль (инструмент).

У звуков, чьи высоты тонов отстоят на целое число октав k (т.е. частоты их основных тонов соотносятся как 2^k) в идеальном случае все 100% гармоник перекрываются. В случае реальных звуков это означает, что частоты основного тона и обертонов, например, ноты «ля» третьей октавы примерно равны частотам обертонов ноты «ля» второй октавы с чётными номерами, или частотам обертонов ноты «ля» первой октавы с номерами, кратными 4. Ряды, образованные основными частотами и гармониками нот, отличающихся на октаву, выглядят так:

Ок-	Частоты гармоник															
I	*F*	*2F*	*3F*	*4F*	*5F*	*6F*	*7F*	*8F*	*9F*	*10F*	*11F*	*12F*	*13F*	*14F*	*15F*	*16F*
II		*2F*		*4F*		*6F*		*8F*		*10F*		*12F*		*14F*		*16F*
III				*4F*				*8F*				*12F*				*16F*
IV								*8F*								*16F*

Известно, что если частоты основных тонов звуков З1 и З2 соотносятся как $\frac{m}{n}, \quad m,n \in Z, \quad m,n \geq 1$, то гармоники звука З1 с номерами $P(k) = n \cdot k, \quad \text{где } k = 1,2,3,...$ перекрывают гармоники звука З2 с номерами $Q(k) = m \cdot k, \quad \text{где } k = 1,2,3,...$. Для звуков с основными тонами, отличающимися на целое число октав, это означает, что большинство их гармоник, имеющих одинаковые частоты будут иметь номера, являющиеся степенями 2. Т.е. система полосовых фильтров с центральными частотами, образующими ряд $f_{ö_i} = f_0 \cdot 2^i$ (частоту f_0 будем называть *основной частотой системы*), будет иметь мощность отклика на тональный звук тем большую, чем ближе будет основной тон этого звука к центральной частоте любого из фильтров системы. Иными словами, можно построить систему фильтров, настроенную, например, на ноты «ля» определённого числа октав. Для этого центральные частоты полосовых фильтров системы должны быть равны основным тонам нот «ля» разных октав.

На рисунке 5 приведены огибающие испытательного сигнала и отклики трёх систем фильтров, настроенных на разные ноты. В качестве испытательного сигнала использована запись хроматической гаммы от «до» малой октавы до «си» третьей октавы, исполненной на скрипке. Три системы фильтров настроены соответственно на ноты «до», «до-диез» и «ре». Как и ожидалось, максимальные амплитуды откликов

соответствуют именно тем нотам, на которые фильтры настроены, значительные интенсивности откликов соответствуют малым секундам, квартам и квинтам от нот, на которые настроены системы фильтров (соответственно, 1, 5 и 7 отклики после каждого максимального). Такое поведение системы фильтров обсудим позже, а пока рассмотрим, что она из себя представляет.

Каждая система фильтров состоит из 8 полосовых фильтров. В качестве полосовых фильтров используются рекурсивные фильтры 2-го порядка, разностные уравнения которых получены методом билинейного z-преобразования из коэффициентов передачи аналоговых прототипов. Фильтры-прототипы построены по схеме на рис. 6, имеют одинаковую добротность и отличаются центральной частотой.

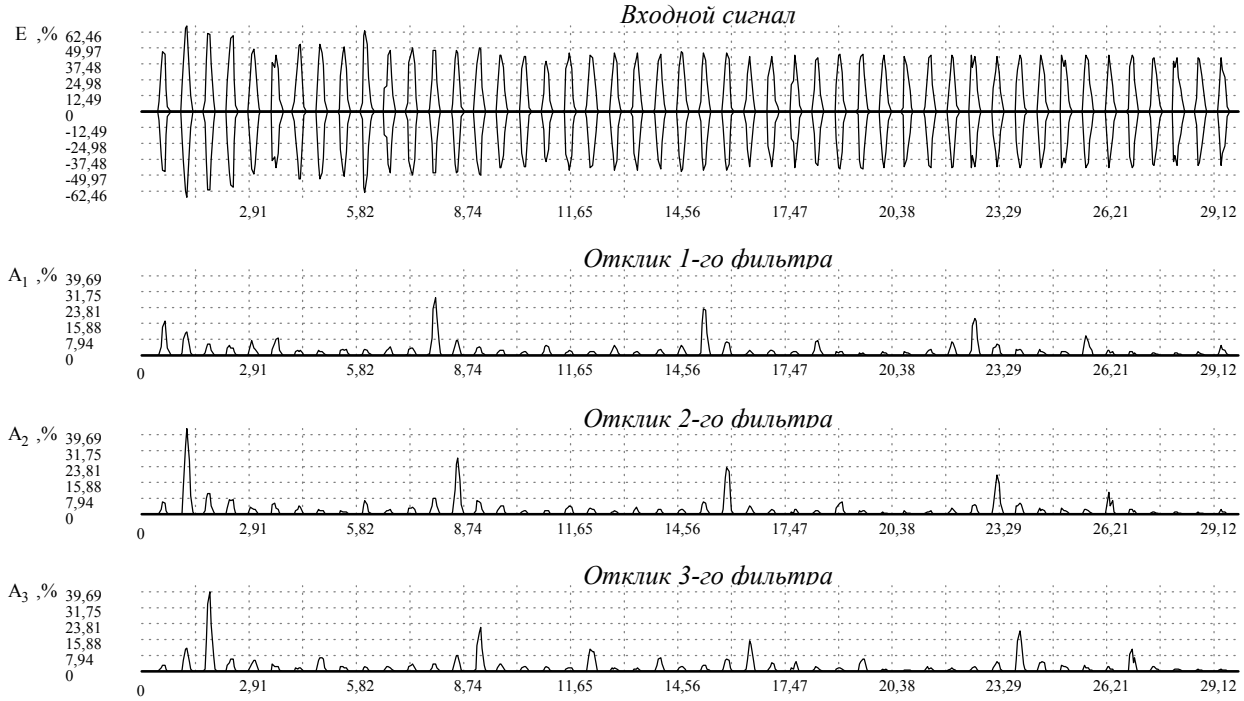


Рис. 5 Испытательный сигнал и отклики на него трёх систем фильтров

На рисунке 7 приведены семейства АЧХ, ФЧХ и характеристик групповой за-

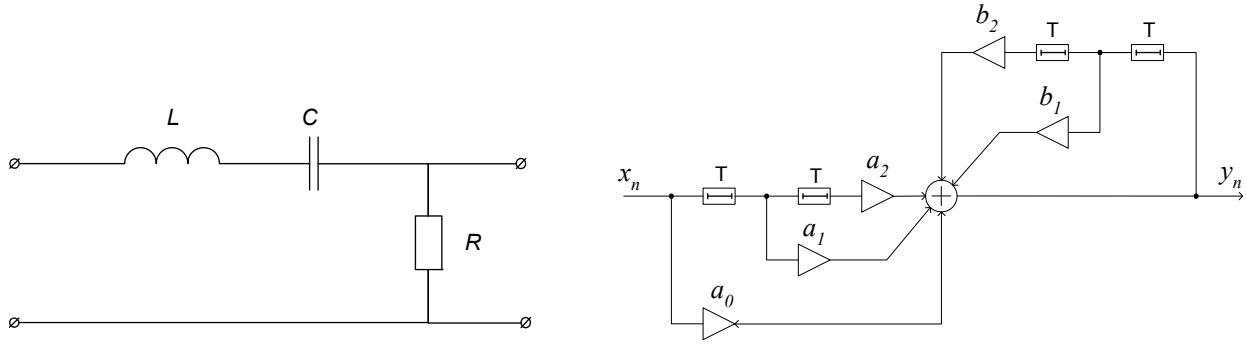
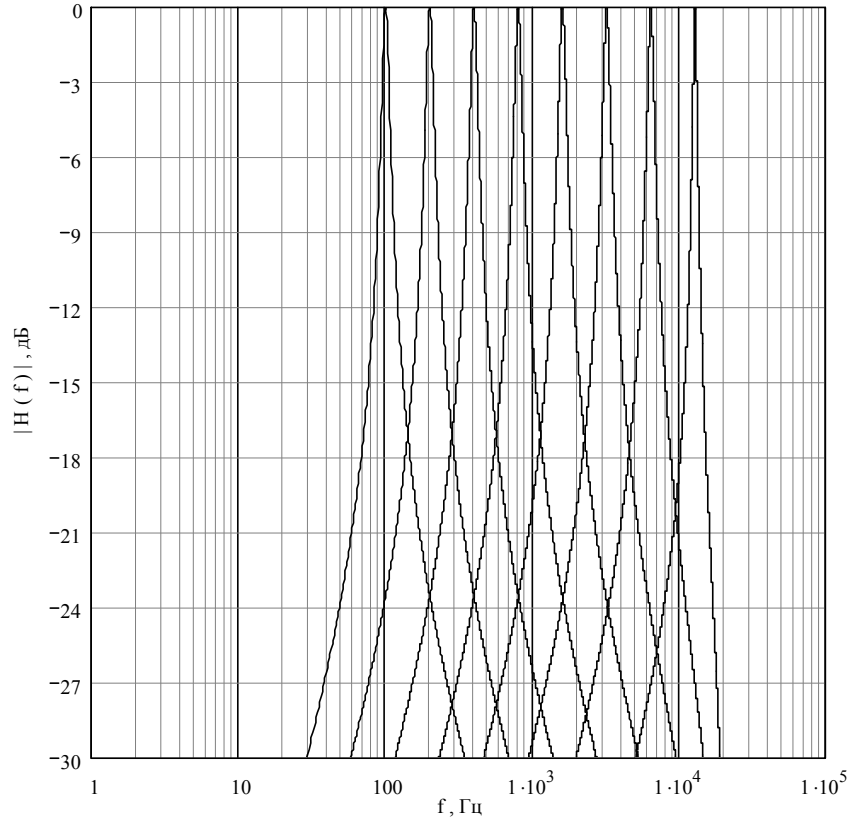


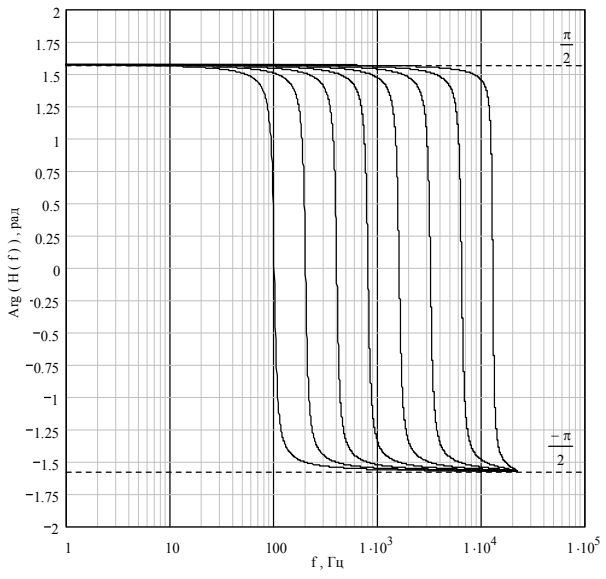
Рис. 6 Схема аналогового фильтра-прототипа и рекурсивного фильтра 2-го порядка

держки фильтров, составляющих систему с основной частотой 100 Гц.

а)



б)



в)

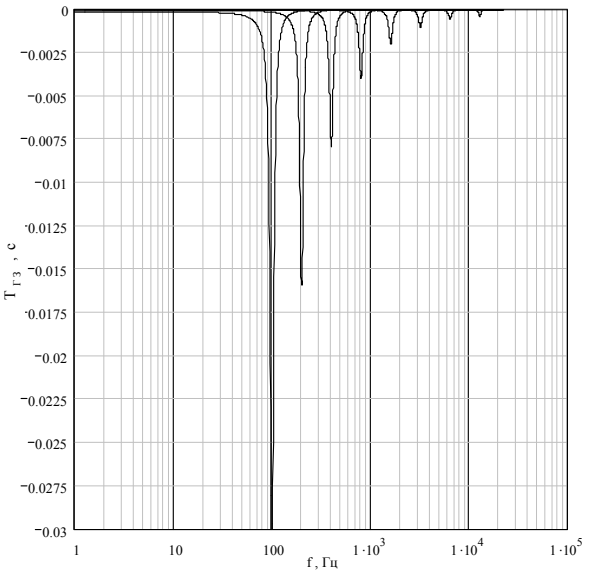
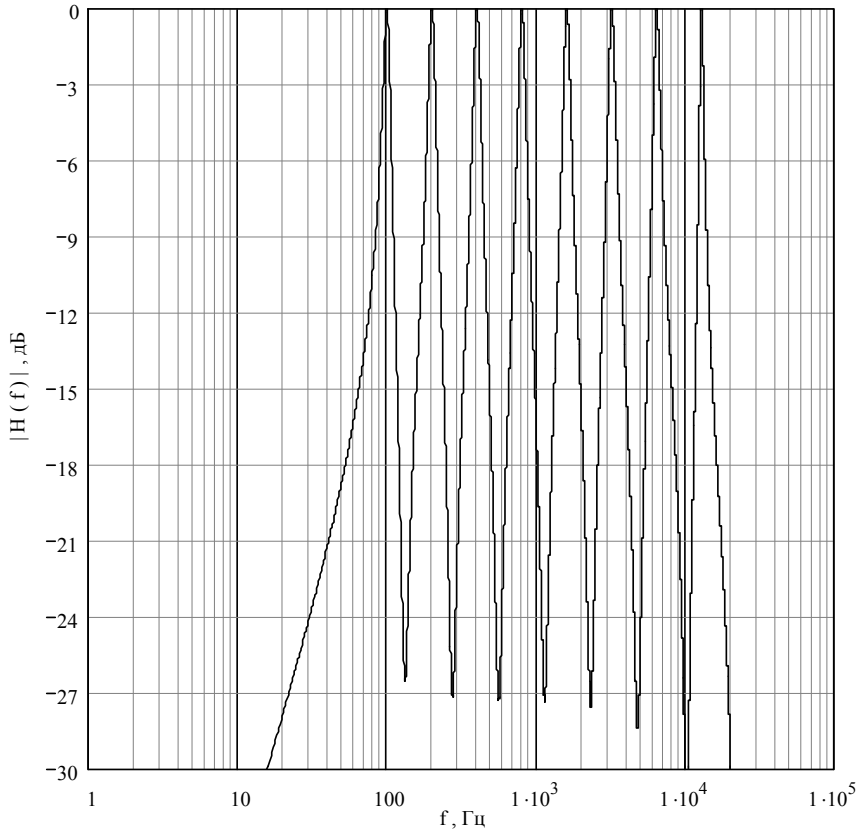


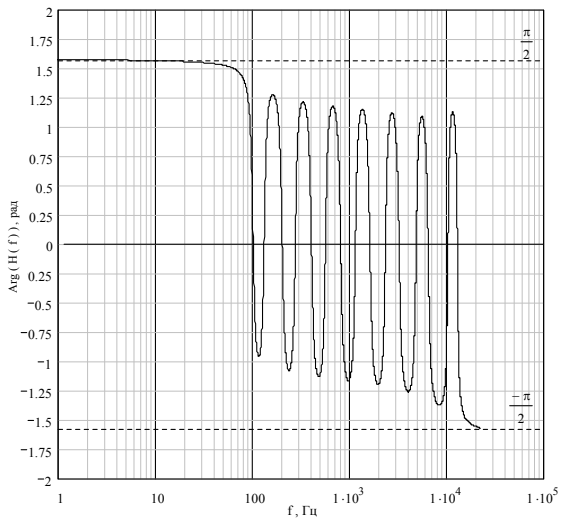
Рис. 7 Семейства характеристик цифровых фильтров, составляющих систему с основной частотой 100 Гц: а) АЧХ, б) ФЧХ, в) характеристика групповой задержки

Те же характеристики системы в целом приведены на рисунке 8. Сама система состоит из 8 включённых параллельно полосовых фильтров, выходной сигнал системы — сумма их выходных сигналов.

а)



б)



в)

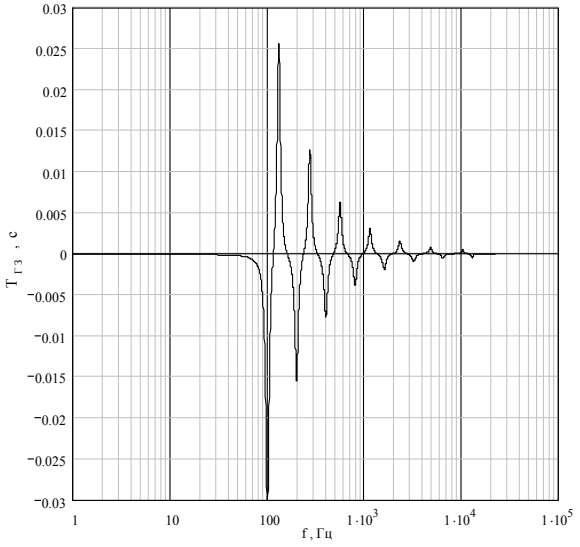


Рис. 8 Характеристики системы с основной частотой 100 Гц: а) АЧХ, б) ФЧХ, в) характеристика групповой задержки

Обсудим поведение построенной системы фильтров. В действительности система откликается не только на звуки с основным тоном, отличающимся от основной частоты системы на целое число октав. Звуки, отличающиеся по высоте тона на кварту и квинту, имеют примерное соотношение основных частот соответственно 4/3

и 3/2. В таком соотношении находятся «до» и «фа» (кварта) и «до» и «соль» (квинта). Тогда каждая 3-я гармоника «фа» имеет ту же частоту, что каждая 4-я гармоника «до», т.е. совпадают частоты гармоник: 3-ей и 4-ой, 6-ой и 8-ой, 12-ой и 16-ой. Это означает, что при звучании ноты «фа» её 3-я, 6-я и 12-я гармоники будут подавлены системой фильтров (будут пропущены 1-я, 2-я, 4-я, 8-я и т.д.), настроенных на «фа», но будут пропущены системой, настроенной на «до». Очевидно, что это приведёт к ошибочной сегментации в том смысле, что будет принято решение о появлении двух звуковых объектов вместо одного. Следовательно, необходим дополнительный параметр для принятия решения о появлении нового звукового объекта. В качестве такого параметра предлагается использовать основной тон отклика системы фильтров. Если основной тон выходного сигнала системы фильтров отличается на целое число октав от основной частоты системы, а значение огибающей превышает некий порог, принимается решение о появлении нового звукового объекта. Для иллюстрации работы алгоритма будем устанавливать значение огибающей отклика системы равным нулю, если не принято решение о появлении нового звукового объекта:

$$E_i^C(t) = \begin{cases} E_i(t) \ , f_{OT}(t) = f_0 \cdot 2^k \ , E_i(t) > E_{ПОР}(t) \\ 0 \qquad\qquad иначе \end{cases}, \text{ где} \qquad (3)$$

f_0 — основная частота системы фильтров,

$E_i(t)$ — огибающая выходного сигнала (отклика) системы фильтров,

$E_{ПОР}(t)$ — пороговое значение огибающей.

Схема системы автоматической сегментации на основе априорной информации и «фильтрации по периоду» приведена на рисунке 9. В основе системы — 12 фильтров с АЧХ вида рис. 8 а), разделяющих входной сигнал на сигналы широкополосные в отличие от ранее описанных систем. Схема обработки откликов всех 12-ти фильтров одинакова: последовательное подавление сигналов с интенсивностями ниже пороговой и сигналов с ОТ (ВОТ — выделитель основного тона), не соответствующими ожидаемым для данного фильтра значениям. Более жирными линиями на схеме показано прохождение основных (*информационных*) сигналов, более тонкими линиями — сигналов, управляющих работой схемы сегментации. Интенсивности откликов сравниваются для определения максимального для данной выборки значения. В блоке «сравнение с порогом» сигнал с интенсивностью меньше пороговой приравнивается нулю. В блоке «принятие решения» нулю приравниваются сигналы с несоответствующим ОТ.

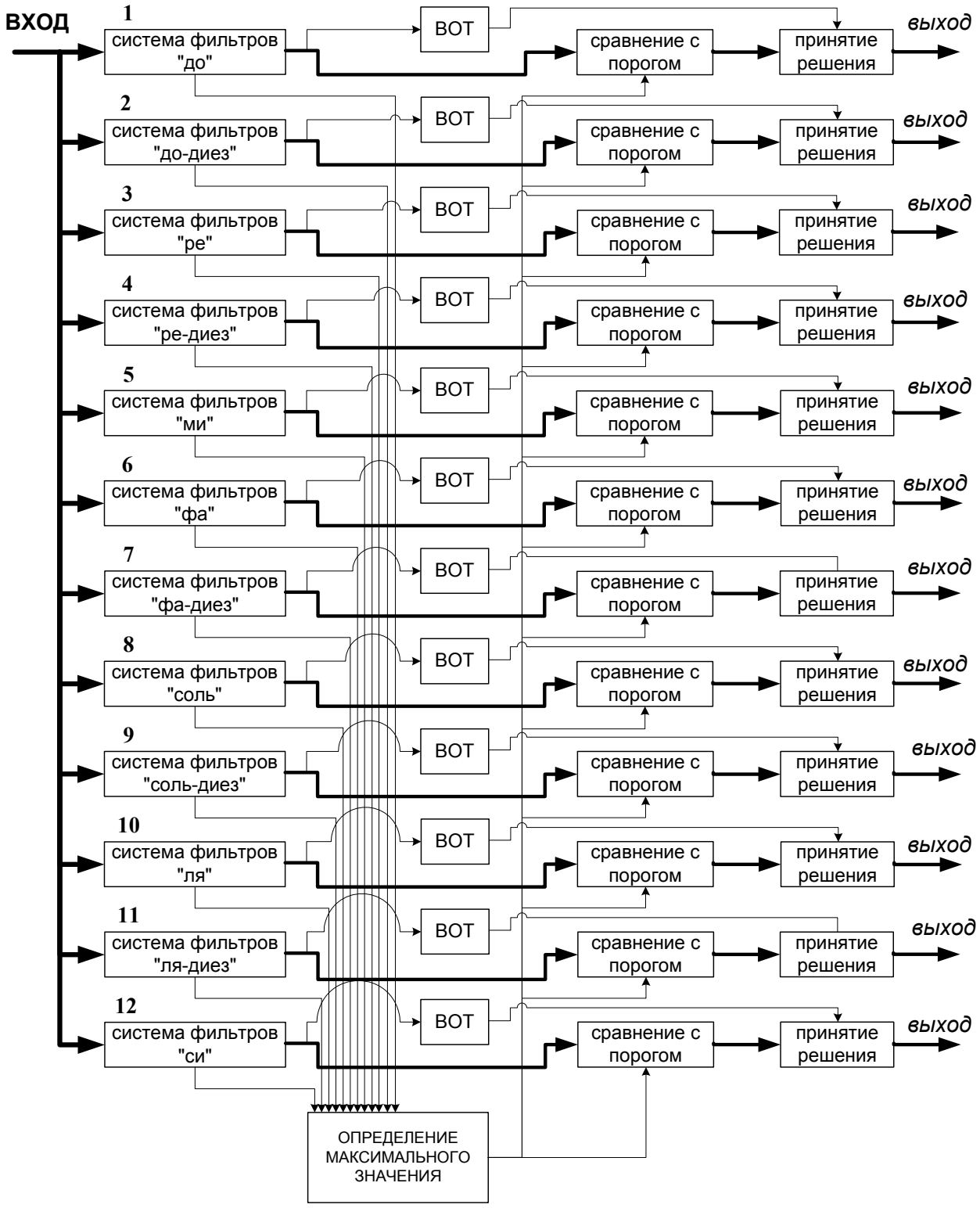


Рис. 9 Схема сегментации на основе фильтрации по периоду

Выходных сигналов 12 — по количеству фильтров. Сигналы эти названы *информационными*, поскольку несут информацию об обнаруженных звуковых объектах. И информация эта почти полная: мощности откликов пропорциональны мощностям звуковых объектов, номер отклика (по номеру фильтра) позволяет с точностью до октавы определить основной тон обнаруженного звукового объекта (т.е. утверждать, например, что звучит «ля», но не известно, какой октавы).

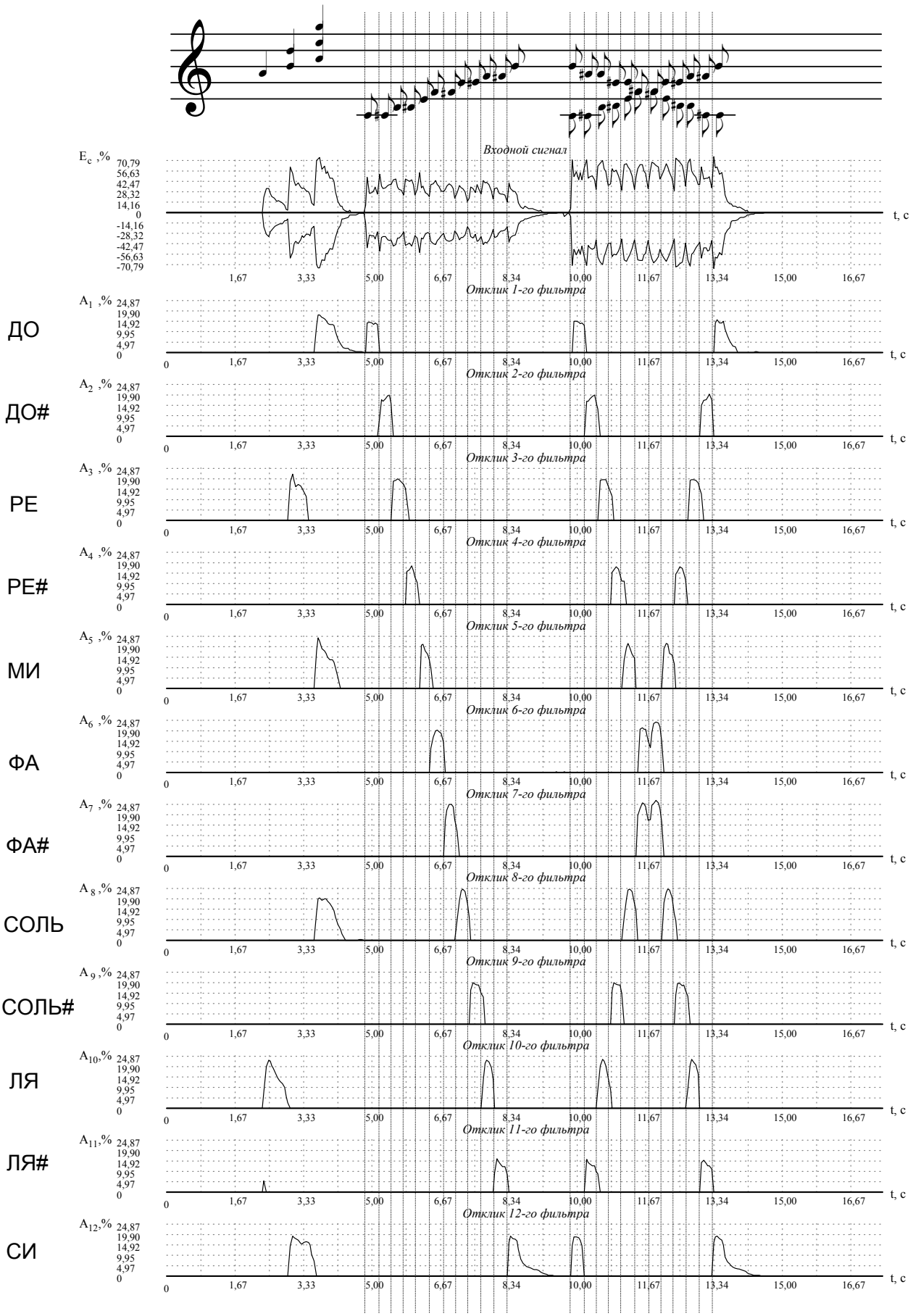


Рис. 10 Пример работы алгоритма сегментации

Важно отметить, что точное определение высоты тона звуковых объектов по откликам фильтров некорректно — информация в них неполная. Более того, при определении ОТ анализом входного сигнала можно исправить и ошибки сегментации. В то же время, предлагаемый алгоритм сегментации не только локализует во времени звуковые объекты, но и даёт алгоритму определения высоты тона набор *гипотез*, т.е. значений ОТ служащих наиболее вероятными кандидатами. Выделение ОТ многоголосных сигналов намного сложнее, чем одноголосных, и наличие гипотез, полученных ещё до начала анализа (на предмет определения ОТ) повышает достоверность распознавания. Поэтому предлагаемый алгоритм сегментации можно назвать *ориентированным на распознавание*.

На рисунке 10 приведены отклики 12 систем фильтров с основными частотами, составляющими хроматическую гамму от «до» до «си» малой октавы. Огибающая входного музыкального сигнала и соответствующая нотная запись приведены на том же рисунке. Музыкальный сигнал синтезирован программным синтезатором с помощью звуков фортепиано. Отклики фильтров «почищены» по алгоритму (3), в качестве $E_{ПОР}(t)$ использовано значение $0{,}4 \cdot E_{МАКС}(t)$, где $E_{МАКС}(t)$ — максимальное по всем 12-ти откликам значение огибающей на данном интервале времени. Результаты, можно сказать, близки к идеальным: правильно определено количество звучащих нот, моменты их появления и затухания, однако при синтезе той же мелодии с помощью звуков трубы правильно локализованы лишь 50% звуковых объектов. Такое различие вызвано различиями свойств музыкальных инструментов. Звуки трубы богаче обертонами, при этом наиболее интенсивны обертоны среднего диапазона спектра звука. Звуки фортепиано, напротив имеют наиболее интенсивные первые несколько обертонов, амплитуда обертонов спадает с частотой (см., например [7]). В результате поведение системы сегментации сильно зависит от свойств сигнала.

Скажем несколько слов о добротности системы фильтров. Уменьшение добротности ухудшает избирательность, увеличение добротности ведёт к «затягиванию» фронтов выходного сигнала. Стремиться к максимальному увеличению избирательности нецелесообразно: в реальных сигналах всегда имеет место отклонение частот обертонов и основных тонов от точных значений, определяемых музыкальным строем, эталоном для настройки и номером гармоники. Выбор добротности системы — отдельная проблема, требующая большого количества экспериментов. Целесообразным представляется построение *настраиваемых* систем фильтров для точной подстройки центральных частот полосовых фильтров при отклонении высот тонов от эталонных.

Подведём итоги. Прежде всего, предлагаемый алгоритм сегментации в определённых условиях оказался работоспособным. Основные достоинства и недостатки алгоритма определяются использованием априорной информации. Априорная информация может быть не всегда точной: в реальных сигналах возможны следующие отклонения:

- отклонения от ожидаемого строя (соотношения между высотами тонов)
- отклонения от ожидаемых значений высот тонов (отклонения от эталонных значений)
- отклонения частот обертонов от значений, кратных основной частоте.

С другой стороны, именно использование априорной информации позволяет ожидать и фиксировать появление звуковых объектов с определёнными параметрами: каждый из 12-ти фильтров настроен на свою ноту, хотя истинное значение высоты тона не определяется. Именно использование априорной информации позволяет ещё при сегментации составить набор гипотез для алгоритма определения ОТ звуковых объектов.

Исследования требует вопрос о виде АЧХ разделяющих фильтров. Прежде всего, очевидна необходимость высокой прямоугольности АЧХ фильтров, составляющих системы. Рассмотренный вариант — наиболее простой, использованы фильтры 2-го порядка, минимизированы объём требуемой памяти и вычислительная сложность. Неочевидно, что области пропускания систем фильтров, настроенных на разные ноты не должны перекрываться, возможно, такое перекрытие повысит устойчивость системы к описанным выше отклонениям частот компонентов звуковых объектов.

Литература

1. **Klapuri A.** “Sound onset detection by applying psychoacoustic knowledge” IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1999.

2. **Moore B., Glasberg B., Baer T.** “A model for the prediction of thresholds, loudness and partial loudness” Journal AES, Vol. 45, No. 4, pp. 224-240. April 1997.

3. **Goto M., Muraoka Y.** “Beat Tracking based on Multiple-agent Architecture - A Real-time Beat Tracking System for Audio Signals”. Proceedings of The Second International Conference on Multiagent Systems, pp.103–110, 1996.

4. **Goto M., Muraoka Y.** “A Real-time Beat Tracking System for Audio Signals”. Proceedings of the 1995 International Computer Music Conference, pp.171–174, September 1995.

5. **Scheirer Eric D.** “Music-Listening Systems”. PhD Thesis, Massachusetts Institute of Technology, 2000.

6. **Рабинер Л., Гоулд Б.** Теория и применение цифровой обработки сигналов, пер. с англ. — М.: Мир, 1978.

7. **Кудинов А. А.** «Исследование структуры звуковых объектов», М. : 2002, Депонировано в ЦНТИ «Информсвязь»