Введение в мультимедийные базы данных

12. Базы данных для аудио

- 12.1. Представление аудиоданных
- 12.2. Сжатие аудио
- 12.3. Извлечение аудиоданных
- 12.4. Стандарт MIDI

Представление аудиоданных

- Наступление «цифровой» эпохи: появление CD-дисков в 1982 году
- Значительное улучшение в общем качестве звука и отношении «сигнал-шум» относительно лучших аналоговых систем
- Для передачи по сетям данных необходима широкая полоса пропускания

Преобразование аналогового сигнала в цифровую форму (аналого-цифровое преобразование):

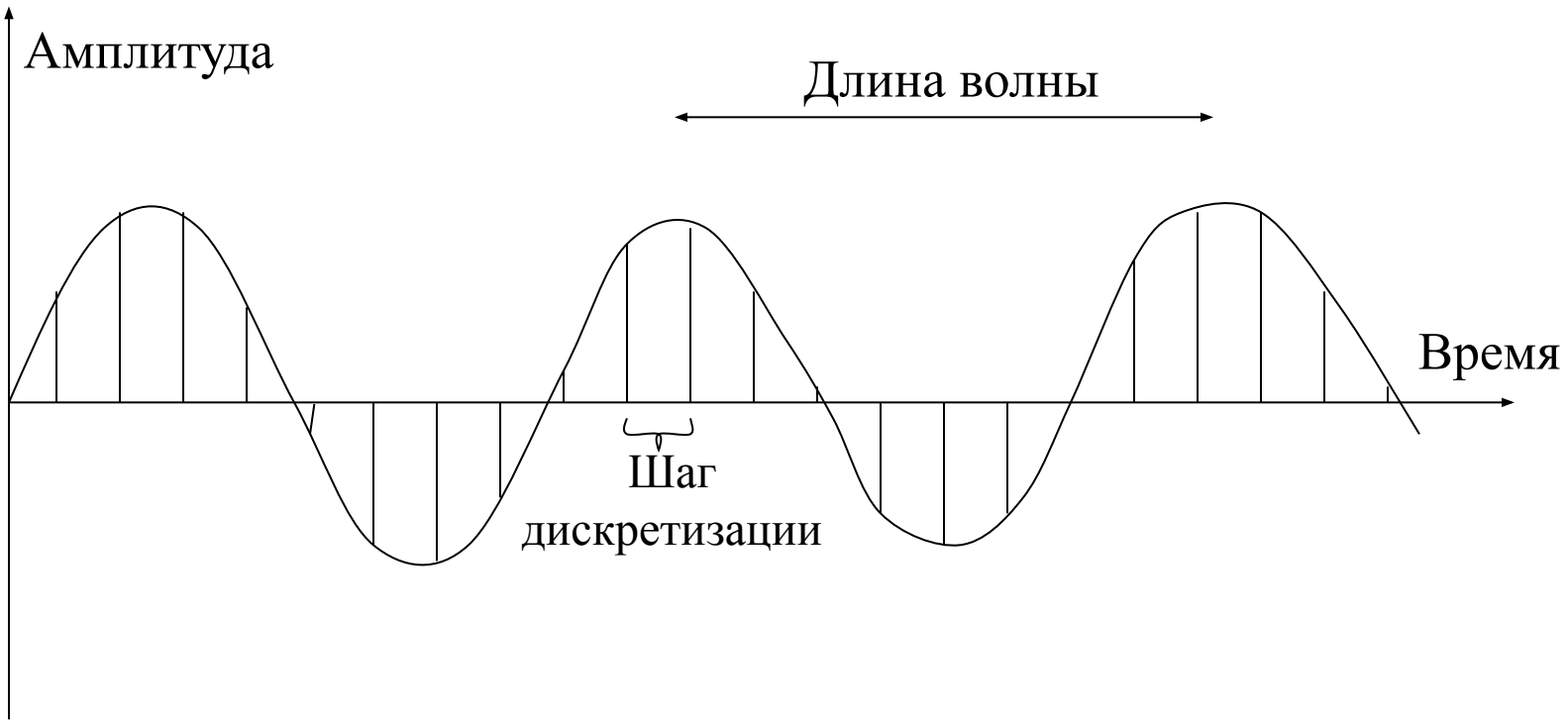
- Линейная импульсно-кодовая модуляция (ИКМ) (PCM)
- Двухэтапный процесс:

  а) Дискретизация (sampling): замер амплитуды сигнала через равные промежутки времени; типичные частоты дискретизации – 32, 44.1, 48кГц (или половины от них)

  Теорема Уиттакера-Найквиста-Котельникова-Шеннона *(или просто теорема Котельникова :)*: аналоговый сигнал со спектром, ограниченным частотой F_{max}, может быть однозначно и без потерь восстановлен по своим дискретным отсчётам, взятым с частотой $f_{дискр} = 2* F_{max}$; человеческое ухо ≈ 20...20 000Гц

Представление аудиоданных

Иллюстрация:



Представление аудиоданных

б) Квантование (quantization): дискретная шкала значений для наблюдаемых амплитуд

- Линейное квантование: одинаковые шаги квантования
- Адаптивное квантование: величина шага зависит от свойств сигнала
- Неравномерное квантование: неодинаковые величины шагов в зависимости от диапазона амплитуд на различных участках сигнала

Типичное квантование: 16 бит на значение, что дает 65536 различных значений

Вместе с частотой дискретизации 44.1кГц и двумя (стерео) каналами получим: 2 x 16 x 44 100 ≈ 1.4 Мбит/с

Цифро-аналоговое преобразование:

- Погрешность дискретизации значений амплитуд ведет к искажению восстанавливаемого сигнала (шум дискретизации по амплитуде)
- В целом, достаточно точное приближение к изначальному сигналу

Сжатие аудио

- В аналоговом сигнале как правило нет резких скачков интенсивности; поэтому если кодировать не саму амплитуду сигнала, а ее изменение по сравнению с предыдущим значением, то можно обойтись меньшим числом разрядов

А) Дельта-модуляция:

- Крайне простой подход, иногда используется для кодирования речи
- Одноразрядное квантование
- Следующее значение апроксимируется предыдущим значением ± Δ (Δ может быть фиксированной или адаптивно-настраиваемой)

Б) Адаптивная дифференциальная импульсно-кодовая модуляция (ADPCM):

- Используется преимущественно для сжатия речи
- Следующее значение предсказывается на основе предшествующих значений
- Шаг квантования может адаптивно-настраиваться
- Рекомендация ITU-T G.726 (кодирование речи):
  - 8000 значений в секунду; 5, 4, 3, или 2 бита на значение
  - 40, 32, 24 или 16 Кбит/с соответственно (PCM-сигнал (речь) – 8бит на 8000 значений в секунда, что дает 64 Кбит/с)

Сжатие аудио

B) Многополосное кодирование:

- Частотное разделение сигнала на поддиапозоны (полосы) частот
- Каждый поддиапазон частот кодируется независимо
- Рекомендация ITU-T G.722:
  - Речь с высоким качеством со скоростью передачи 64Кбит/с: может разделяться на два канала – основной и вспомогательный: 56 + 8 или 48 + 16 Кбит/с
  - Диапазон исходного сигнала – от 50 до 7000 Гц
  - 16000 значений в секунду
  - 14-битный квантизатор
  - Два поддиапазона: 0-4 кГц и 4-7 кГц
  - Окончательное кодирование с помощью ADPCM

Сжатие аудио

- MPEG-аудио:
  - Частоты дискретизации – 32, 44.1, 48кГц (или половины от них); значения помещаются во фреймы (384/576/1152 значения на фрейм) и далее обрабатываются фреймы
  - 32 фильтра, каждый с шириной полосы в 1/64 от частоты дискретизации
  - Изменяемые шаги квантования (переменная скорость потока): каждый фрейм может кодироваться разным числом бит
  - Скорости сжатого потока (MPEG-1 Layer 3) – от 32 до 320 Кбит/с (вспомним: скорость для CD - 1.4 Мбит/с)
  - Достаточно хорошее качество звука при скоростях от 128 Кбит/с
  - MPEG Layer I: устарел
  - MPEG Layer II (MP2): аудиовещание (цифровые радио и телевидение)
  - MPEG Layer III (MP3): компьютерные/интернет-приложения

Г) Кодирование с преобразованием:

- Одномерное дискретное косинус-преобразование (DCT)
- MPEG Layer III: модифицированное DCT к поддиапазонам частот

Сжатие аудио

Е) Психоакустическое кодирование:

- Возможное дополнение методов В) и Г)
- Психоакустика - изучение обработки звуков мозгом человека
- Используются знания о том какие свойства не имеют большого значения для человеческого уха
- Большие амплитудные значения (громкий звук) на одной частоте влияют на воспринимаемость соседних частот
- Определенные диапазоны частот более важны
- Акустически-маловажные части аудиосигнала могут не рассматриваться: использовать меньшее число бит (больший шаг квантования) для менее значимых поддиапазонов
- MPEG: психоакустические модели 1 и 2
  - Работают с Layer I-III
  - Обработка по 512/1024 значений
  - Более сложная модель 2 специально разработана для Layer III; используется Фурье-преобразование
  - С точки зрения человеческого восприятия - сжатие без потерь

Извлечение аудиоданных

а) На основе метаданных:

- К речевой информации могут быть добавлены дополнительные атрибуты (как к изображениям или видео), например: источник речи (диктор), дата, продолжительность, композитор, оркестр, инструмента и т.д.
- Атрибуты могут быть приписаны ко всей аудиопоследовательности или только к ее определенным частям
- Можно использовать стандартные методы извлечения документов

б) Распознавание речи:

- Пример приложения: распознавание голосовых команд пользовательским интерфейсом; «цифровой дом» - *расдвинуть шторы, включить свет*; распознавание путем нахождения ближайших волновых форм (нечёткая определённость)
- Более сложные приложения: грамматический разбор произнесенного и преобразование, например, в запрос к бд; может дополняться методами обработки естественного языка; обычно используется предопределенный набор образцов-шаблонов
- Продвинутые приложения: преобразование практически произвольной речи в текст, на основе образцов и фонетических правил

Извлечение аудиоданных

в) Распознавание говорящего:

- Сложнее чем распознавание речи
- Приложения: системы безопасности
- Чувствительны к физическому состоянию говорящего (например, при гриппе может искажаться тембр голоса)
- Вариации:
  - Текстозависимое распознавание (простейшее):
    Ограниченный набор возможных слов/предложений
    Сравнение волновых форм
  - Текстозависимое распознавание (более сложное):
    Может основываться, например, на распознавании основного тона голоса
    Должны храниться более сложные речевые образцы пользователей
    Сложные верификационные алгоритмы сверяют произнесенное с хранящимися образцами

Извлечение аудиоданных

г) Индексация аудиоданных:

- Индексация метаданных (внешних атрибутов):
  - Аналогично индексации текстовых документов: инвертированный индекс, сигнатурные файлы и т.д.
- Индексация аудиосигнала:
  - Сначала, разбить на сегменты (фреймы)
  - Преобразование (например, DCT)
  - Индекс (возможно многомерный) по группам наиболее важных коэффициентов; запросы по близости (ближайший сосед, k ближайших соседей)
  - Затруднение: выравнивание сегментов

Стандарт MIDI

- Экономичный способ кодирования информации о том как воспроизвести музыку
- Стандарт с 1983 года
- Коэффициент сжатия порядка 1:1000 (относительно оцифрованного аудио)
- Содержит только инструкции, необходимые синтезатору, для проигрывания музыки
- Инструкции – MIDI-сообщение
- Возможность редактировать музыку, менять скорость проигрывания и т.д.
- Специальное приложение: караоке
- MIDI-поток: асинхронный, 31.25Кбит/с, 8+2 бита на 1 байт
- Возможность объединения MIDI-устройств (в цепочку и звездой)
- MIDI-каналы: одновременное воспроизводение звуков от нескольких независимых инструментов
- Один синтезатор может воспроизводить несколько звуков (многотембровость)
- Структура MIDI-сообщения: 1 статусный байт (команда) плюс 1-2 байта данных (например, номер ноты, громкость)
- Преемник MIDI: MPEG-4 Structured Audio (MP4-SA)
  - Structured Audio Orchestra Language (SAOL) – «звуки»
  - Structured Audio Score Language (SASL) – «ноты»