Речевые технологии

4/2008

Содержание

Редакция:
Редактор — *Артём Ганькин*
Корректор — *Татьяна Денисьева*
Дизайн — *Анна Ладанюк*
Вёрстка — *Максим Буланов*

Адрес редакции: 109341, Москва, ул. Люблинская, д. 157, корп. 2.
Тел.: 8 (495) 979-54-27

Подписано в печать 24.09.2009. Формат 60×90⅛. Бумага офсетная. Печать офсетная.
Печ. л. 6. Заказ № 1002. Издательский дом «Народное образование».
Отпечатано в типографии НИИ школьных технологий. 143500, г. Истра-2, ул. Заводская, д. 2А.
Тел.: 8 (901) 513-97-64, (495) 792-59-62.

От составителя номера

Научные исследования в области разработки речевых технологий начаты в Белоруссии в середине 60-х годов XX века на основе тесной кооперации группы «речевиков» из Минского радиотехнического института (руководитель — Б.М. Лобанов) и лаборатории экспериментальной фонетики Минского института иностранных языков (руководитель — Е.Б. Карневская).

Затем долгие годы, вплоть до середины 80-х, основные исследования были сконцентрированы вначале в Лаборатории обработки речи Минского отделения Центрального НИИ связи, а затем в Лаборатории распознавания и синтеза речи (зав. лабораторией Б.М. Лобанов) Института технической кибернетики, ныне Объединённого института проблем информатики НАН Белоруссии (http://uiip.bas-net.by).

Главные направления научных исследований лаборатории:

- высококачественный синтез русской речи по тексту;
- компьютерное клонирование персонального голоса и дикции;
- многоязычный синтез речи;
- робастное распознавание дискретной и слитной речи;
- обнаружение ключевых слов;
- компьютерные системы реабилитации инвалидов слуха и зрения.

В 90-х годах исследования по речевой тематике начинают проводиться на кафедре электронных вычислительных средств (зав. кафедрой А.А. Петровский) Белорусского государственного университета информатики и радиоэлектроники (http://www.bsuir.by).

Главные направления научных исследований кафедры:

- кодирование сигналов звука и речи;
- распознавание речи в шумовой обстановке;
- редактирование шумов;
- психоакустика;
- подавление эффектов эхо и реверберации;
- модификация и конверсия голоса;
- проектирование проблемно-ориентированных мультимедиа-систем реального времени.

В 2000-х годах появляются первые коммерческие организации, занимающиеся разработкой и продажей речевых систем. К настоящему времени, кроме перечисленных выше организаций, в области разработки речевых технологий активно работают ООО «Сакрамент» (www.sakrament.com, научный руководитель И.Э. Хейдоров) и ООО «Речевые технологии» (www.speechtech.com, директор В.В. Киселёв).

Главные направления деятельности ООО «Сакрамент»:

- синтез речи;
- распознавание речи;
- идентификация голоса;
- индексация аудиофайлов.

Главные направления деятельности ООО «Речевые технологии»:

- распознавание речи;
- синтез русской речи;
- распознавание ключевых слов;
- идентификация диктора.

Данный номер журнала посвящён публикации работ, авторами (или соавторами) которых являются, в основном, аспиранты и соискатели, принадлежащие к белорусской школе учёных в области речевых технологий. Из представленных в журнале двенадцати статей первые шесть посвящены решению различных теоретических и практических вопросов распознавания, анализа и преобразования речевых сигналов:

1. *Янь Цзинбинь, И.Э. Хейдоров, А.А. Ткаченя* «Исследование характеристик системы поиска ключевых слов на основе минимального интервала редактирования и мер доверительности»;
2. *Янь Цзинбинь, у ши, А.М. Сорока, А.А. Трус* «Классификация аудиосигналов с использованием одноклассового метода опорных векторов для систем поиска информации в мультимедиа-архивах»;
3. *А.С. Рылов, В.В. Киселёв, А.Г. Давыдов, В.А. Чижденко* «Система оперативной модификации голоса диктора на основе полувокодера»;
4. *А.Н. Павловец, М.З. Лившиц, Д.С. Лихачёв, А.А. Петровский* «Конверсия голоса с использованием модели сепарации речевого сигнала на компоненты «гармоники+шум» и переходные фреймы»;
5. *А.Н. Павловец, А.А. Петровский* «Использование закономерностей психоакустики в процедуре квантования параметров гармонической модели речевого сигнала»;
6. *А.А. Петровский* «Построение психоакустической модели в области вейвлет-коэффициентов для перцептуальной обработки звуковых и речевых сигналов».

Следующие четыре работы посвящены различным практическим вопросам реализации речевых технологий:

7. *Е.А. Снюгина* «Выбор языковых средств пользователем для формулирования ответа в ходе диалога»;
8. *Т.В. Левковская* «Текстозависимая верификация диктора на основе коллектива решающих правил»;
9. *Л.И. Цирульник, Ю.С. Гецевич* «Алгоритмы преобразования сложноструктурированных объектов для синтеза речи по тексту»;
10. *Б.М. Лобанов, О.Г. Сизонов* «Квазиречевой видеонавигатор для слепых».

Последние две статьи — обзорные:

11. *С.Е. Лячканов* «Криминалистический учёт лиц по фонограммам их речи»;
12. *Е.Б. Никитин* «Технология VoiceXML и её приложения».

Безусловно, представленные здесь работы не могут отражать всё многообразие ведущихся в Белоруссии исследований в области создания новых речевых технологий. Однако они могут дать некоторое представление о тематике и научном уровне проводимых исследований и разработок.

Б.М. Лобанов

Исследование характеристик системы поиска ключевых слов на основе минимального интервала редактирования и мер доверительности

Янь Цзинбинь,
аспирант
И.Э. Хейдоров,
доцент, кандидат физико-математических наук
А.В. Ткаченя,
студент

В данной работе для поиска ключевых слов предлагается использовать двухэтапный алгоритм на основе решётки слогов и минимального интервала редактирования для детектирования временных координат возможных ключевых слов и верификации найденных слов на основе мер доверительности. Эксперимент показал, что использование решётки слогов обеспечивает большую точность распознавания, чем решётка на основе фонем. Точность поиска ключевых слов при этом составляет 88.2%. Использование метода опорных векторов (МОВ) для объединения мер доверительности позволило уменьшить вероятность ложной тревоги до 8.8%, что позволяет на этой основе создавать системы поиска ключевых слов с приемлемыми с практической точки зрения характеристиками.

Введение

Поиск ключевых слов в речевых файлах является одной из наиболее сложных задач в области обработки речи. С её помощью можно реализовать аудиоиндексацию и поиск информации, например: поиск информации в Интернете, контроль речевой связи, управление речевыми библиотеками [1, 2].

Наиболее простой метод поиска ключевых слов использует распознаватель с большим словарём для перевода непрерывной речи в текст. Для поиска ключевого слова осуществляется поиск в полученном тексте с использованием традиционных алгоритмов поиска текста. Проблема этого метода состоит в том, что из-за ограниченного множества слов в распознавателе невозможно распознать слова, отсутствующие в словаре, например: имена, акронимы и слова из иностранных языков.

Другой метод поиска ключевых слов основан на скрытых Марковских моделях (СММ). Он использует СММ для каждого ключевого слова и одну «модель мусора» для всех остальных слов [3]. Этот метод не имеет ограничений при условии, что определено множество ключевых слов, которые необходимо найти. Но для каждого нового ключевого слова необходимо не только обучать новую СММ-модель, но также нужно заново обучать «модель мусора». Поэтому использование этого метода при определённых условиях является сложным.

На сегодняшний день наиболее популярным решением задачи поиска ключевых слов в потоке слитной речи является использование акустико-фонетической СММ и вычисление апостериорных вероятностей фонемной решётки, каждый узел которой ассоциируется с моментом времени в рамках произнесённой речи. Данный метод обладает большой гибкостью, и результат поиска не зависит от словарного множества распознавателя, что позволяет осуществлять поиск для любого запрашиваемого ключевого слова без дополнительного обучения системы. Однако данный подход сопряжён со значительной вычислительной сложностью и, следовательно, требует введения некоторых дополнительных процедур для создания систем поиска ключевых слов в реальном масштабе времени. Кроме того, отсутствие в явном виде словаря требует введения дополнительной процедуры верификации найденных ключевых слов.

В связи с этим в данной работе для поиска ключевых слов предлагается использовать двухэтапный алгоритм на основе решётки слогов и минимального интервала редактирования (МИР) для детектирования временных координат возможных ключевых слов и верификации найденных слов на основе мер доверительности.

Структура системы поиска ключевых слов на основе решётки слогов

Для осуществления поиска и верификации ключевых слов была предложена следующая схема системы ***(рис. 1)***. После выделения признаков речевого сигнала последовательность наблюдений $O=\{o_1,o_2,...,o_T\}$, где T — количество векторов-признаков, поступает на вход системы распознавания,

созданной на основе СММ, которая позволяет преобразовать последовательность наблюдений в последовательность некоторых структурных единиц речи (фонем, аллофонов и т.д.). Для вероятностей акустических единиц речи, полученных на выходе СММ, строим решётку $L = (N, A, n_{start}, n_{end})$ — направленный непериодический граф, где N — множество узлов, A — множество связей между узлами и n_{start}, $n_{end} \in N$ — начальный и конечный узлы решётки соответственно ***(рис. 2)***. Связь представляется в виде $a = (S[a], E[a], I[a], w[a])$, где $S[a]$, $E[a] \in N$ — начальный и конечный узлы; $I[a]$ — фрагмент речи (слог или фонема); $w[a] = p_{ac}(a)^{1/\lambda}$ — весовой коэффициент связи, который представляет собой вероятность перехода между узлами; $p_{ac}(a)$ — акустическое сходство; λ — весовой коэффициент.

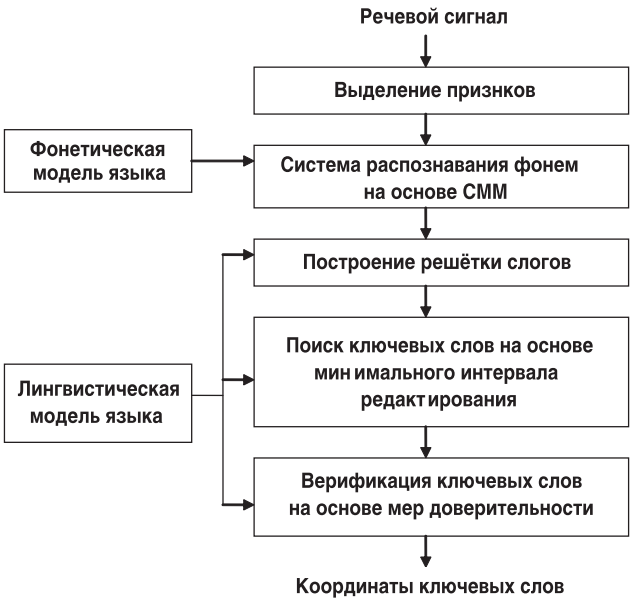

Рис. 1. Структура системы поиска ключевых слов

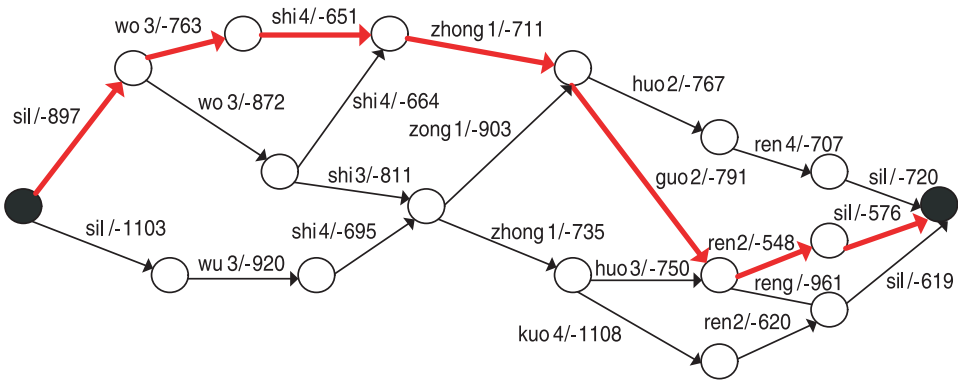

Рис. 2. Решетка слогов речи

Алгоритм минимального интервала редактирования (МИР) [4]

Для последовательности слогов, полученной на выходе решётки, характерно наличие трёх основных типов ошибок. Ошибка типа «замена» отвечает случаю, когда вместо одного слога система распознала другой. «Вставка» — это появление в потоке слогов, которых не было в исходной речи. И, наконец, ошибка типа «удаление» характерна случаю пропуска слога, присутствующего в реальной речи. Для повышения эффективности поиска ключевых слов в систему необходимо ввести дополнительный модуль, позволяющий учесть априорную информацию об ошибках такого рода.

Определим понятие минимального интервала редактирования между строками U и V как минимальные затраты на преобразование строки U в строку V с помощью трёх основных операций: вставки, удаления и замены. Для хранения значений минимального расстояния используем матрицу $M(0,...,p)(0,...,q)$, где p и q — это длины строк U и V соответственно. Элементы матрицы M для строк U и V вычисляются следующим образом:

$$M(0)(0) = 0;$$
$$M(i)(0) = i * I; i = 1,..., p;$$
$$M(0)(j) = j * D; j = 1,..., q;$$
$$M(i)(j) = \min\{M(i-1)(j-1) + S(U(i), V(j)), M(i-1)(j) + D, M(i)(j-1) + I\};$$

где S, I и D — затраты на операции замены, вставки и удаления соответственно.

Пусть $K=\{k_1, ..., k_N\}$ — это последовательность слогов ключевого слова, которое нужно найти в слоговой решётке; $Q=\{q_1, ..., q_M\}$ — последовательность слогов предполагаемого ключевого слова; C_S, C_I, C_D — функции затрат для замены, вставки и удаления; $MED(K,Q,C_S,C_I,C_D)$ — функция минимального расстояния, которая в качестве своего значения возвращает элемент $M(N,M)$; S_{max} — пороговое значение для функции MED.

Алгоритм минимального расстояния реализуется следующим образом.
1) Получение предполагаемых ключевых слов.
2) Для каждого предполагаемого ключевого слова:
a) получение последовательности слогов $Q=\{q_1, ..., q_M\}$;
b) вычисление минимальной корректировки $S=MED(K,Q,C_S,C_I,C_D)$;
c) если $S \le S_{max}$, тогда Q есть ключевое слово.

Использование данного алгоритма позволяет определить временные координаты возможных ключевых слов с учётом всех возможных ошибок, совершённых на уровне фонетического распознавания. Однако за счёт этого данный подход имеет большое количество ложных тревог, когда алгоритм принимает решение о наличии ключевого слова, хотя в данный момент времени оно отсутствует. В связи с этим в работе предлагается ввести дополнительный этап верификации ключевых слов, основанный на использовании мер доверительности [5].

Метод доверительного интервала

Представим речевой сигнал на входе системы поиска ключевых слов в виде последовательности наблюдений $O=\{o_1,o_2,...,o_T\}$. Обозначим модель ключевого слова как Λ и определим меру доверительности как количественную оценку совпадения O и Λ. Другими словами, мера доверительности определяется вероятностью генерации последовательности O на основе модели Λ. Для моделирования речевых сигналов на основе СММ компоненты вектора признаков (вектора наблюдений) предполагаются независимыми, а для каждого состояния плотность распределения вероятностей наблюдений представляется гауссовыми смесями (ГС), описывающими вероятность вектора наблюдений o_t в состоянии j как

$$b_j(o_t) = \prod_{s=1}^{S}\left[\sum_{m=1}^{M_s} c_{jsm} N\left(o_{st}; \mu_{jsm}; \sigma_{jsm}\right)\right],$$

где S — размерность вектора признаков речевого сигнала; M_s — размерность ГС для компоненты s; c_{jsm}, μ_{jsm}, σ_{jsm} — соответственно весовое значение, среднее значение и дисперсия для m-ой компоненты ГС, по предположению представляющей собой нормальное распределение:

$$N\left(o_{st}; \mu_{jsm}; \sigma_{jsm}\right) = \frac{1}{\sqrt{2\pi}\sigma_{jsm}} \exp\left(-\left(o_{st} - \mu_{jsm}\right)^2 / 2\sigma_{jsm}^{\ 2}\right).$$

Пусть j-ое состояние модели Λ представляется подпоследовательностью $O^k = \{o_k, o_{k+1}, ..., o_K\}$. Тогда определим значение меры доверительности CM_j следующим образом:

$$CM_j = \frac{\sum_{t=k}^{K}\sum_{m=1}^{M_s}\sum_{s=1}^{S} D\left(o_{st}; \mu_{jsm}, \sigma_{jsm}\right)}{(K-k)M_s S},$$

где

$$D(o_{ti}; \mu_{jsm}, \sigma_{jsm}) = \begin{cases} 1 & (o_{st} - \mu_{jsm}) \in [\mu_{jsm} - k\sigma_{jsm}, \mu_{jsm} + k\sigma_{jsm}] \\ 0 & \textit{иначе} \end{cases},$$

где k — управляющий параметр для доверительного интервала. Определим меру доверительности для каждого ключевого слова Λ путём нормализации значений для каждого состояния:

$$CM_1 = \frac{1}{N}\sum_{j=1}^{N} CM_j,$$

где N — количество состояний СММ. Верификация ключевого слова происходит путём сравнения полученной меры доверительности с некоторым порогом, выбранным, как правило, эмпирически.

Данная мера доверительности обеспечивает достаточно высокую точность верификации и имеет большой потенциал по улучшению характеристик. Алгоритм расчёта меры доверительности на основе использования доверительного интервала подразумевает получение множества промежуточных результатов, использование которых позволит повысить точность при незначительном увеличении вычислительной сложности.

Акустическая мера доверительности на основе нормализации длительности состояния

Одним из недостатков представленной выше меры доверительности является отсутствие её нормализации на длину состояния СММ, вследствие чего возможны ситуации, когда состояние с малой длительностью затеняет результаты для более длительной последовательности наблюдений.

Обозначим СММ как $\lambda = \{N, \pi, A, B\}$, где N — число состояний СММ $S = \{S_1, S_2, \ldots S_N\}$, π — матрица начальных вероятностей, A и B — матрицы переходных вероятностей и вероятностей наблюдения соответственно. Обозначим начальный и конечный моменты времени нахождения системы в состоянии i как $b[i]$ и $e[i]$. Тогда определим нормализованную меру доверительности следующим образом:

$$CM_2 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{e[i]-b[i]+1}\sum_{t=b[i]}^{e[i]}\log p(o_t \mid s_i)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{e[i]-b[i]+1}\sum_{t=b[i]}^{e[i]}\log b_i(o_t)\right]$$

Путём замены $\log p(o_t \mid s_i)$ на $\log(s_i \mid o_t)$ согласно формуле Байеса получим ещё одно выражение для меры доверительности:

$$CM_3 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{e[i]-b[i]+1}\sum_{t=b[i]}^{e[i]}\log p(s_i \mid o_t)\right],$$

где

$$p(s_i \mid o_t) = \frac{p(o_t \mid s_i)p(s_i)}{\sum_{j=1}^{N} p(o_t \mid s_j)p(s_j)} = \frac{b_i(o_t)\,p(s_i)}{\sum_{j=1}^{N} b_j(o_t)\,p(s_j)} \quad .$$

Обе эти меры доверительности представляют собой среднее значение акустической вероятности в рамках СММ.

Мера доверительности на основе динамического рейтинга

Описанные выше методы подтверждения ключевых слов в той или иной степени используют модели фонем. Они просты и эффективны, в значительной степени позволяют снизить вероятность ложной тревоги. Рассмотрим ещё один метод верификации ключевых слов, основанный на использовании апостериорной сети и динамического рейтинга.

В результате работы декодирующего алгоритма Витерби текущему вектору наблюдений o_t входящего речевого сигнала и каждой допустимой в данный момент модели слова Λ_j ставится в соответствие последовательность состояний модели $S(\Lambda_j, o_t)$. Определим меру соответствия (характеристическое значение) $L_j(o_t)$ последовательности состояний $S(\Lambda_j, o_t)$ модели Λ_j в момент времени t следующим образом:

$$L_j(o_t) = \ln P\left(o_t \mid S\left(\Lambda_j, o_t\right)\right) \quad j = 1,2,\ldots, N(o_t) \qquad (1)$$

где $N(o_t)$ — число активных моделей в момент времени t. Отсортируем набор характеристических значений по убыванию для всех моделей:

$$L_{j_1}(o_t) > L_{j_2}(o_t) > \cdots > L_{j_k}(o_t) > \cdots > L_{j_{N(o_t)}}(o_t) \ .$$

Пусть характеристическое значение $L_{Kw}(o_t) = L_{jk}(o_t)$ для модели ключевого слова Λ_{Kw} занимает в рейтинге k-ое место, тогда определим динамический рейтинг на o_t-ом фрейме как $k/N(o_t)$:

$$Q(o_t \mid \Lambda_{Kw}) = \frac{\sum_{k=1}^{N(o_t)} G(L_k(o_t) - L_{Kw}(o_t))}{N(o_t)} \ ,$$

где

$$G(L_k(o_t) - L_{Kw}(o_t)) = \begin{cases} 0 & \text{if } L_k(o_t) \le L_{Kw}(o_t) \\ 1 & \textit{иначе} \end{cases} \ .$$

На основе вышеприведённых рассуждений представим динамический рейтинг как обобщённое характеристическое значение в рамках всей анализируемой длительности сигнала:

$$CM_4(O \mid \Lambda_{Kw}) = \frac{1}{T}\sum_{t=1}^{T} Q(o_t \mid \Lambda_{Kw}) \ .$$

Для верификации ключевого слова необходимо полученное значение динамического рейтинга сравнить с порогом. Особенностью метода динамического рейтинга является то, что порог отторжения может быть установлен одинаковым для всех ключевых слов.

Следует отметить, что вычисление динамического рейтинга для ключевых слов не приводит к существенному увеличению вычислительной сложности по сравнению с алгоритмами на основе акустической меры доверительности. В процессе декодирования Витерби значения вероятностей для выражения (1) уже рассчитаны, поэтому вычисление $Q(o_t)$ и нормализация приводят к дополнительным KT сложениям и вычитаниям, а также T+1 делениям.

Однако данный алгоритм обладает рядом существенных преимуществ. Во-первых, верификация ключевых слов на основе динамического рейтинга имеет более стабильный характер, особенно в условиях шума. Имеющийся в речевых данных шум влияет на абсолютные значения акустических вероятностей, однако не оказывает практически никакого влияния на место в рейтинге, а соответственно, и на характеристическое значение.

Во-вторых, динамический рейтинг рассчитывается исключительно на основании значений акустической вероятности, поэтому изменение словаря ключевых слов не оказывает никакого влияния на работоспособность алгоритма.

В-третьих, как уже упоминалось ранее, порог отторжения может быть установлен единым для всех ключевых слов, а кроме того, наличие шума в исходных данных не влияет на абсолютное значение порога.

Представленные выше меры доверительности позволяют провести верификацию найденных ключевых слов. Особенно хороших результатов позволяет добиться использование комплексных критериев верификации, построенных на основе использования нескольких мер доверительности одновременно.

В данной статье для объединения мер доверительности и проведения экспериментов был использован МОВ [6].

Эксперимент

Для проведения эксперимента и определения сравнительных характеристик предложенной системы поиска ключевых слов была использована база данных, содержащая 124 часа речи. С использованием этой базы данных для создания акустико-фонетической модели речи были обучены СММ, в качестве вектора признаков был использован вектор из 39-ти мел-кепстральных характеристик и их производных. Тестирование системы производилось с использованием реальной слитной речи продолжительностью 3,54 часа, в которой обнаружению подлежали 13 часто повторяемых ключевых слов, появившихся в тестовых фрагментах 794 раза.

Для оценки характеристик системы поиска ключевых слов были использованы в качестве основных следующие величины: вероятность обнаружения Pd (процент правильно найденных ключевых слов), вероятность ложного отказа (процент неправильно отторгнутых ключевых слов), вероятность ложной тревоги FAR (процент принятия ложных слов в качестве ключевых). Для представления фонем была выбрана непрерывная СММ с пятью состояниями, каждое из которых моделировалось ГС.

Первый эксперимент был посвящён выбору базовой единицы распознавания и влиянию процедуры построения гипотез на основе МИР на результат правильного обнаружения ключевых слов ***(см. табл. 1)***.

Таблица 1

Вероятность правильного обнаружения ключевых слов в зависимости от структурной единицы решетки

Структурная единица решётки	Вероятность обнаружения Pd без использования МИР, %	Вероятность обнаружения Pd с использованием МИР, %
Фонема	80.7%	82.7%
Слог	85.2%	88.2%

Как видно из таблицы 1, использование слогов в качестве минимальных структурных единиц речи для задачи поиска ключевых слов является более предпочтительным. Кроме того, использование алгоритма МИР для формирования последовательности слогов с учётом возможных ошибок типа «замещение», «вставка» и «пропуск» позволило увеличить вероятность правильного обнаружения ключевых слов в среднем на 3%.

Второй эксперимент был посвящён тестированию точности верификации ключевых слов в зависимости от используемой меры доверительности ***(см. табл. 2)***.

Из таблицы 2 видно, что МОВ, объединяющий несколько различных мер доверительности, позволяет достичь намного лучших характеристик отторжения, чем каждая из представленных мер доверительности в отдельности. Изменение параметров МОВ позволяет в определённых пределах изменять вероятность ложной тревоги, настраивая систему под конкретные условия.

Таблица 2

Вероятность ложной тревоги системы верификации ключевых слов в зависимости от меры доверительности

FAR без верификации, %	FAR для меры динамического рейтинга, %	FAR для доверительного интервала, %	FAR для комплексного алгоритма МОВ, %
40.5	22.5	30.2	8.8

Эксперимент показал, что предложенный метод на основе МОВ, объединяющей различные меры доверительности, позволяет достичь точности верификации ключевых слов более высокой, чем при использовании этих мер доверительности по отдельности.

Использование МОВ для объединения мер доверительности позволило уменьшить вероятность ложной тревоги до 8.8% при сохранении той же точности правильного обнаружения, что позволяет на этой основе создавать системы поиска ключевых слов с приемлемыми с практической точки зрения характеристиками.

Заключение

В данной работе предложена двухэтапная структура системы поиска ключевых слов на основе решётки слогов с использованием алгоритма минимального расстояния и верификации найденных слов с использованием мер доверительности.

Результат эксперимента показал, что решётка слогов позволяет достичь большей точности, чем решётка фонем, и при этом точность поиска ключевых слов составляет 88,2%.

Использование МОВ для объединения мер доверительности позволило уменьшить вероятность ложной тревоги до 8,8% при сохранении той же точности правильного обнаружения, что позволяет на этой основе создавать системы поиска ключевых слов с приемлемыми с практической точки зрения характеристиками.

Литература

1. *J.Mamou, D.Carmel, R.Hoory.* Spoken Document Retrieval from Call-Center Conversations. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA. — 2006. — P. 51–58.

2. *Chung-Hsien Wu, Yeou-Jiunn Chen.* Multi-keyword Spotting of Telephone Speech Using a Fuzzy Search and Keyword-driven Two-level CBSM. Speech Communication. — 2001, — 33(3). — P. 197–212.

3. *Rose R.C.* Keyword detection in conversational speech utterance using hidden Markov model based continuous speech recognition. Computer Speech and Language. — 1995. — vol. 9. — P. 309–333.

4. *A. Kartik, V. Ashish.* Keyword Search Using Modified Minimum Edit Distance Measure. IEEE International Conference on Acoustics, Speech and Signal Processing. — 2007. — vol. 4. — P. 929–932.

5. *Mingxing Xu, Fang Zheng, Wenhu Wu, Ditang Fang.* Research on rejection method for continuous speech keyword spotting system. Journal of Tsinghua University. — 1998. — vol. 38(1). — P. 89–91.

6. *Vapnik, V.N.* Statistical Learning Theory. New York, Wiley, 1998.

Янь Цзинбинь —

аспирант БГУ, научные интересы — методы и алгоритмы обработки речевых сигналов, теория метода опорных векторов, алгоритмы обнаружения ключевых слов в потоке речи.

Хейдоров Игорь Эдуардович —

Окончил с отличием БГУ (Минск) в 1996 году, с 1998 года работает на кафедре радиофизики БГУ, к.ф.-м.н. (2000 г), доцент. Сфера научных интересов — методы и алгоритмы распознавания и синтеза речи, автоматическая индексация аудиодокументов. Автор 40 работ.

Ткаченя А.В. —

студент БГУ, факультет радиофизики и электроники . Область научных интересов — системы анализа и индексирования аудиосигналов, скрытые Марковские модели в задачах распознавания речи.

Классификация аудиосигналов с использованием одноклассового метода опорных векторов для систем поиска информации в мультимедиа-архивах

Янь Цзинбинь,
аспирант
У Ши,
аспирант
А.М. Сорока,
магистрант
А.А. Трус,
магистрант

В данной статье представлен способ классификации аудиосигналов на основе одноклассового метода опорных векторов (МОВ). Анализ полученных в ходе эксперимента результатов показывает, что данный метод имеет хорошую точность классификации, и его эффективность выше, чем у классификаторов на основе метода Байеса, скрытых Марковских моделей (СММ) и нейронной сети.

Abctract

This paper proposes an audio classification method based on one-class support vector machine (SVM). Experiment results show that SVM has good classification accuracy, and performs better than other classification systems using Bayes method, Hidden Markov Model (HMM) and Neural Network (NN).

Введение

Классификация аудиосигналов — это задача разделения непрерывного потока акустических данных на однородные участки (речь, музыка, звуки окружающей среды,

тишина и т.д.). С одной стороны, решение данной задачи можно использовать для удаления неречевых фрагментов из аудиосигналов, что приведет к увеличению скорости и точности распознавания речи. С другой стороны, сформулированная выше задача классификации является шагом предварительной обработки аудиосигналов в задачах классификации музыки по жанрам [1, 2]. Задача классификации является неотъемлемой частью алгоритмов индексации аудиосигналов и на сегодняшний день выполняется вручную, что приводит к огромным затратам человеческого труда и материальных ресурсов. Таким образом, построение системы автоматической классификации аудиосигналов является на сегодняшний день одним из наиболее актуальных нерешённых вопросов мультимедиа-технологий.

Традиционно алгоритм классификации аудиосигналов использует методы на основе решающего правила, учитывающие один или несколько признаков и принимающие решение путём сравнения с некоторым порогом, значение которого определяется обычно эмпирическим способом [3]. Такой подход имеет вполне очевидные недостатки. Во-первых, выбор решающего правила и последовательности классификации на группы не обязательно оптимальны. Во-вторых, ошибки верхнего уровня классификации переходят на следующий уровень и постепенно накапливаются, что значительно снижает общую точность системы. В-третьих, выбор порогового значения для решающего правила в значительной степени зависит от условий эксперимента; незначительное изменение качества речевых сигналов может привести к необходимости повторного обучения системы.

На протяжении последних лет ведутся поиски новых подходов и алгоритмов для классификации речевых сигналов: в частности, для этой задачи использовались классификаторы на основе метода K-ближайших соседей [4], нейронных сетей, моделей гауссовых смесей и алгоритма K-ближайших соседей. Однако каждому из этих методов в той или иной степени свойственны недостатки, не позволяющие строить на их основе высокоточные системы автоматической классификации аудиоданных.

В данной статье предлагается использовать новый подход к классификации аудиосигналов. Метод опорных векторов (МОВ) позволяет найти оптимальную гиперплоскость, разделяющую классы, в некотором модифицированном пространстве признаков, что позволяет преодолеть недостатки алгоритмов на основе решающих правил и пороговых значений. В качестве основной задачи рассматривается разделение аудиосигналов на пять основных классов: речь, речь с окружающими звуками, музыка, тишина и шум.

1. Построение вектора признаков аудиосигнала

При классификации аудиосигналов особое значение приобретает выбор способа построения для вектора признаков сигнала, которые обладают достаточной различающей способностью и являются устойчивыми.

Основные признаки, используемые для решения задачи классификации аудиосигналов, можно разделить на три большие группы. Первая группа признаков — это спектральные характеристики, отражающие свойства одного фрейма.

Признаки второй группы — комплексные спектральные характеристики, отражающие динамические свойства сигнала путём анализа нескольких соседних или всех фреймов сигнала в анализируемой области. К третьей группе признаков относятся характеристики сигнала, построенные на основе анализа последовательности временных отсчётов сигнала.

Для проведения классификации аудиосигналов важное значение имеют характеристики, в той или иной степени отражающие форму спектра. Частота основного тона (ЧОТ) является характерной и важной величиной для речевых сигналов и характеризует, в первую очередь, частоту колебаний голосовых связок человека. Формы ЧОТ для речи и музыки принципиально различаются в силу разной природы образования этих звуков. ЧОТ речи имеет сложную, пересечённую форму и включает лишь небольшой пропорциональный гармоничный состав. А для музыки характерна более гладкая форма ЧОТ.

Другой из таких характеристик является распределение энергии по спектральным диапазонам. Область частоты разделяется на четыре поддиапазона $\mathrm{sbi}(\mathrm{i}=0,1,2,3)$, соответственно $[0, w_0/8],[w_0/8, w_0/4],[w_0/4, w_0/2],[w_0/2, w_0]$, и для каждого вычисляются значения SW_1, SW_2, SW_3, SW_4 согласно выражению (1):

$$SW_i = \frac{1}{E}\sum_{i=L_j}^{H_j}\left|f(i)\right|^2, \qquad (1)$$

где $f(i)$ — коэффициенты БПФ-преобразования данного фрейма, $w_0 = \frac{1}{2} f_s$ f — частота дискретизации, L_j и N_j — нижняя и верхняя частоты диапазона. Энергия аудиосигналов разных типов по-разному распределяется по диапазону. Музыка имеет более равномерное распределение энергии по диапазонам, речь практически целиком фокусируется в первом диапазоне (около 80%).

Помимо распределения энергии по диапазонам, для моделирования формы спектра фрейма важную роль играет центроид частоты, обозначаемый FC и определяемый согласно (2), а также ширина спектра (3).

$$FC = \frac{\sum_{i=0}^{w_0}\left|f(i)^2\right|\cdot i}{\sum_{i=0}^{w_0}\left|f(i)^2\right|} \qquad (2)$$

$$BW = \sqrt{\frac{\sum_{i-0}^{w_0}(i-FC)^2\left|f(i)^2\right|}{\sum_{i=0}^{w_0}\left|f(i)\right|^2}} \qquad (3)$$

В частности, для речевых сигналов, согласно (3), частоты находятся в диапазоне 0.1 кГц — 3.4 кГц, а для музыки характерен диапазон 0.02 кГц — 22.05 кГц.

Помимо характеристик, отражающих характеристики отдельных фреймов, для анализа аудиосигналов очень важно использовать комплексные характеристики, отвечающие

за поведение некоторых участков сигнала в целом. В частности, в качестве компоненты характеристического вектора можно использовать не точное значение кратковременной энергии, а некоторое модифицированное значение отклонения кратковременной энергии (МОКЭ), обозначаемое как $LSTER$ и определяемое следующей формулой:

$$\mathrm{LSTER} = \frac{1}{2N}\sum_{n=0}^{N-1}\left[\mathrm{sgn}(0.5avSTE - STE(n)) + 1\right] \quad , \qquad (4)$$

где $STE = \frac{1}{L}\sum_{i=0}^{L-1} x^2(i)$ — кратковременная энергия фрейма, L — длина анализируемого фрейма, $avSTE = \frac{1}{N}\sum_{n=0}^{N-1} STE(n)$ — усреднённая кратковременная энергия, N — общее число фреймов. МОКЭ является эффективной характеристикой для различения речевых и музыкальных сигналов. В общем случае, поскольку для речи характерно большое число фреймов, содержащих тишину, МОКЭ для речевых сигналов выше, чем для музыки, как показано на ***рис. 1***. На ***рис. 2*** представлено распределение вероятностей для МОКЭ речевого и музыкального сигнала. Если в качестве вектора признаков для разделения музыкальных и речевых сигналов использовать только МОКЭ, а в качестве порога принятия решения использовать точку пересечения двух кривых (рис. 2), то ошибка классификации составит всего порядка 8%.

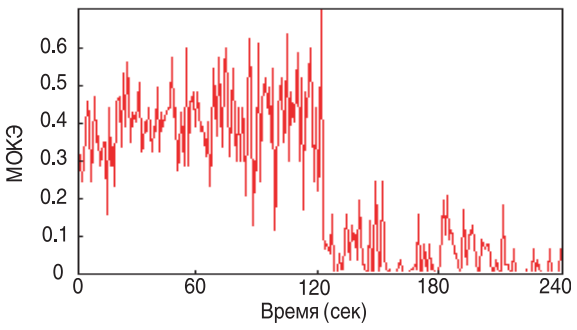

Рис. 1. Модифицированное значение отклонения кратковременной энергии (0–120 сек: речь, 121–240 сек: музыка)

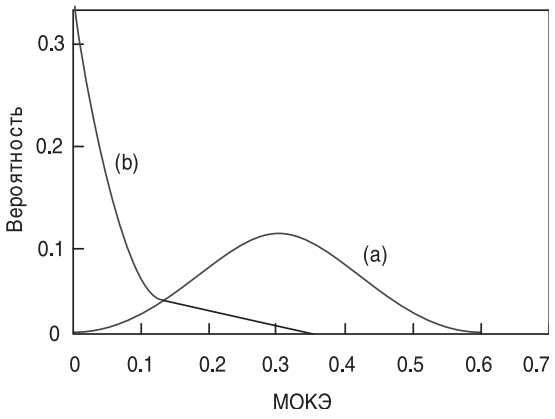

Рис. 2. Распределение вероятности МОКЭ: речь (а), музыка (б)

Другой комплексной характеристикой сигнала является спектральный поток, обозначаемый SF и определяемый как значение средней вариации спектра между двумя соседними фреймами в рамках одной секунды:

$$SF = \frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}\left[\log(A(n,k)+\delta - \log(A(n-1,k)+\delta\right]^2 \quad ,$$

где $A(n, k)$ — дискретное преобразование Фурье (ДПФ) от n-го фрейма входного сигнала:

$$A(n,k) = \left| \sum_{m=-\infty}^{\infty} x(m) w(nL-m) e^{j\frac{2\pi}{L}km} \right| ,$$

где $x(m)$ — исходные входные данные, $w(m)$ — функция окна, L — длина окна, K — порядок ДПФ, N — общее число фреймов и S — очень малое значение, нужное, чтобы избежать переполнения разрядной сетки при вычислениях. В ходе экспериментов было установлено, что спектральный поток для речевых сигналов выше, чем для музыкальных ***(рис. 3)***. Речевой сегмент представлен с 0 по 120 секунду, музыка — с 121 по 240 секунду.

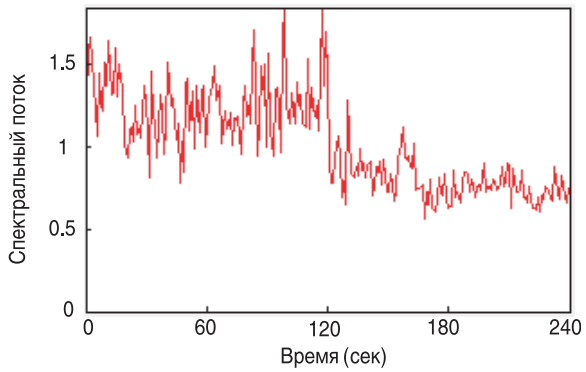

Рис. 3. Спектральный поток для речи (0–120 сек) и музыки (120–240 сек)

О временной структуре сигнала позволяет судить частота переходов через ноль (ЧПН), обозначаемая как ZCR и определяемая следующей формулой:

$$ZCR = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} \left| \mathrm{sgn}[x(m+1)] - \mathrm{sgn}[x(m)] \right| ,$$

где $x(m)$ — дискретный сигнал.

Соотношение тишины определяется как отношение количества фреймов с тишиной к суммарному количеству фреймов в сегменте и обозначается SR:

$$SR = \frac{количество \quad фреймов \quad тишины}{общее \quad число \quad фреймов}$$

В речи часто встречаются паузы, поэтому соответствующее соотношение тишины будет больше для речи по сравнению с музыкой.

ЧПН является важным параметром для описания различных аудиосигналов, однако в некоторых случаях более устойчивым параметром является модификация этой величины (МЧПН), определяемая следующим образом:

$$HZCR = \frac{1}{2N} \sum_{n=0}^{N-1} \left[\mathrm{sgn}\left(ZCR(n) - 1.5 \frac{1}{N} \sum_{n=0}^{N-1} ZCR(n) \right) + 1 \right] ,$$

где n — индекс фрагмента, N — общее количество фрагментов в окне длительностью в 1 секунду. В общем случае речевые сигналы содержат чередующиеся вокальные и невокальные звуки, в то время как такая структура не характерна для музыкальных сигналов. Таким образом, МЧПН будет выше для речевых сигналов по сравнению с музыкальными ***(рис. 4)***.

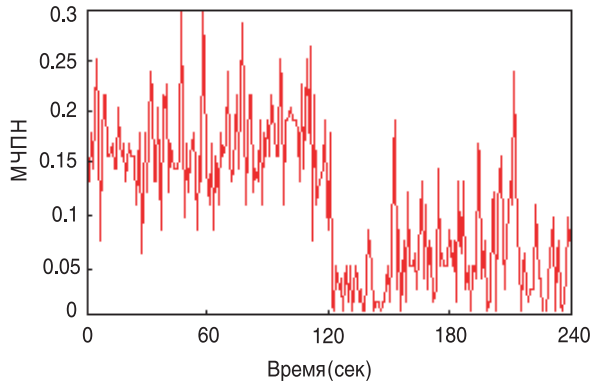

Рис. 4. Модифицированная частота переходов через ноль для речи (0–120 сек) и музыки (121–240 сек)

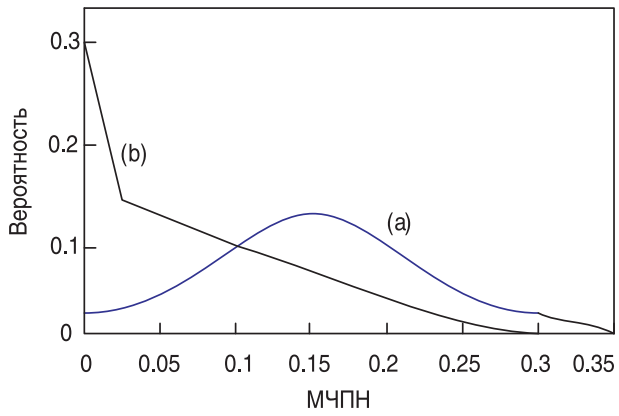

Рис. 5. Распределение вероятностей МЧПН для речи (a) и музыки (b) ноль для речи (0–120 сек) и музыки (121–240 сек)

На ***рис. 5*** представлены кривые распределения вероятностей МЧПН для речи и музыки. При использовании в качестве вектора признаков только МЧПН ошибка сегментации составит порядка 20%.

2. Частота основного тона (ЧОТ)

Традиционный способ определения частоты основного тона с использованием кепстра имеет ряд недостатков: высокая вычислительная сложность, значительное снижение точности определения при наличии фонового шума. Исследования показывают, что алгоритмы извлечения ЧОТ на основе вейвлет-анализа менее чувствительны к наличию фоновых шумов в сравнении с традиционными методами [6]. В качестве базовой функции непрерывного вейвлет-преобразования наиболее часто используется функция Morlet. Аналитическое выражение базовой функции непрерывного вейвлет-преобразования описывается следующей формулой:

$$\psi(x) = Ce^{-x^2/2}\cos(5x)$$

Поскольку большая часть энергии рассматриваемых сигналов лежит в низкочастотном диапазоне, используется логарифмическая шкала с целью улучшить частотное разрешение преобразования. Масштаб вейвлет-функций выбирается в пределах от 10 до 1024 и рассчитывается следующим образом:

$$scale_{k\in[1,num]} = MaxWvLng \cdot \exp\left[-\frac{k}{num}\cdot\log\left(\frac{MaxWvLng}{MinWvLng}\right)\right],$$

где $scale_k$ — масштаб вейвлета с индексом k, $MaxWvLng$ — наибольший масштаб вейвлета, MinWvLng — наименьший масштаб вейвлета, num — число масштабов.

В данном случае каждую компоненту вейвлет-вектора необходимо вычислять, используя массив с вейвлетом соответствующего масштаба. Компоненты вычисляются по следующей формуле:

$$CWT(pos,k) = \sqrt{rscale[k]}\cdot\sum_{i=0}^{MaxWvLng}\left(wave\left[pos-\frac{MaxWvLng}{2}+i\right]\cdot wvlt[k][i]\right), \quad (5)$$

где pos — текущая позиция (во времени), k — номер вейвлета, $rscale[k]$ — число, обратно пропорциональное масштабу вейвлета, $wave[]$ — массив со значениями выборок анализируемого сигнала, $wvlt[k][i]$ — i-ая выборка k-ого вейвлета, масштаб которого обратно пропорционален значению $rscale[k]$.

По формуле (5) рассчитывается двумерный массив вейвлет-коэффициентов в координатах времени-частоты (рис. 6). Максимальное значение вейвлет-коэффициентов является ЧОТ.

Как показано на ***рис. 6***, ЧОТ музыки изменяется более плавно в сравнении с ЧОТ речи. Дополнительно в качестве компонентов векторов признаков используются: дисперсия ЧОТ; степень гармоничности, определяемая как отношение количества фреймов в сегменте, ЧОТ которых не равна нулю, к общему их числу; процентное соотношение гладких сегментов к их общему количеству.

Описанные выше признаки включены в состав вектора признаков аудиосигнала и совместно с 12 мел-частотными кепстральными коэффициентами составили 25-компонентный вектор, который и использовался для классификации.

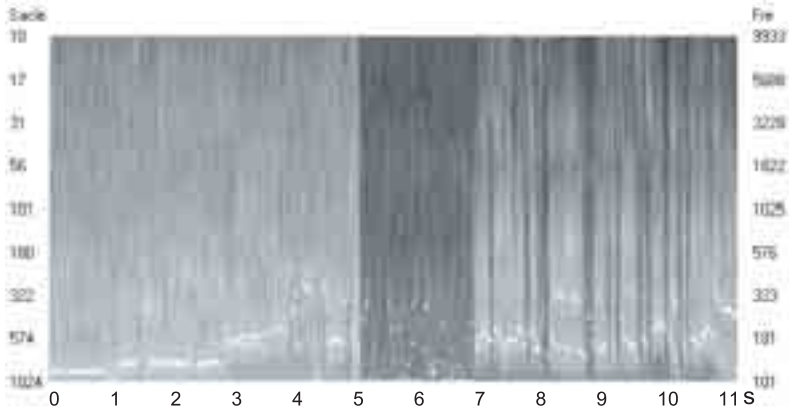

Рис. 6. Вейвлет-преобразование аудиосигнала (0-5 сек: музыка, 5.1-7 сек: тишина, 7.1-11 сек: речь)

3. Использование одноклассового метода опорных векторов в задачах многоклассовой классификации

Классический МОВ представляет собой бинарный классификатор. Для решения задач многоклассовой классификации конструируются каскады бинарных классификаторов. Это, в свою очередь, приводит к увеличению вычислительной сложности алгоритма, увеличению сложности настройки каждого классификатора, уменьшению точности итоговой модели классификации. Использование одноклассового МОВ позволяет увеличить точность и устойчивость итоговой модели классификации, а также снизить вычислительную сложность алгоритма обучения модели.

3.1. Одноклассовый метод опорных векторов

Предположим, что у нас имеется обучающая выборка $(x_1, y_1),...,(x_l, y_l)$, $x \in R^n$, $y \in \{+1,-1\}$, l — количество прецедентов в выборке, n — размерность пространства. В одноклассовом

методе опорных векторов задача обучения представляет собой задачу поиска гиперсферы минимального радиуса, включающей минимальное количество прецедентов из обучающей выборки. Задача обучения сводится к следующей оптимизационной задаче:

$$\min(\frac{1}{2}R^2+\frac{1}{vl}\sum_{i=1}^{l}\xi_i),,$$

где R — радиус гиперсферы, $\xi_1 > 0$ — штраф для соответствующего прецедента.

Набор параметров $\nu\,(0<\nu\le 1)$ представляет собой набор штрафов для каждого прецедента, и, таким образом, в гиперсферу может попадать некоторая часть прецедентов. При небольших значениях ν все прецеденты лежат на поверхности гиперсферы. Увеличение ν приводит к уменьшению радиуса гиперсферы. Решение задачи оптимизации может быть получено с использованием метода Лагранжа:

$$\min \quad W(\alpha)=\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K(x_i,x_j)\ ,$$

где $\sum_{i=1}^{l}\alpha_i=1$, $0\le\alpha_i\le\frac{1}{\nu m}$, $i=1,...,l$, $K(x_i,x_j)$ — ядерная функция.

Наконец, итоговая функция решения может быть представлена так:

$$f(x)=sign(\sum_{i=1}^{l}\alpha_i K(x_i,x)-\rho)\ .$$

Для любых $0<\alpha_i<\frac{1}{vm}$ и x_i рассчитывается параметр $\rho=\sum_{j=1}^{l}\alpha_j K(x_i,x_j)$.

3.2. Создание многоклассового классификатора на основе одноклассового метода опорных векторов

В задачах многоклассовой классификации может быть использован следующий подход: в процессе обучения для каждого класса строится решающее правило с использованием одноклассового МОВ. В процессе классификации проводится испытание тестовой выборки на полученных в процессе обучения функциях решения и в зависимости от максимального значения функции решения делается вывод о принадлежности выборки к какому-либо классу.

4. Экспериментальные результаты и анализ

4.1. База данных эксперимента

База данных эксперимента была представлена следующими аудиосигналами: чистая речь, музыка, звуки окружающей среды, речь с фоновым шумом и тишина.

Выборка речевых сигналов представлена аудиозаписями речи на четырёх различных языках. Выборка речевых сигналов с фоновыми шумами представлена аудиозаписями речи на фоне музыки, речи и городских звуков. Музыкальная выборка включает классическую детскую музыку, музыку, поп, джаз и другие виды.

Все данные представлены в виде аудиозаписей в формате PCM WAV 16kHz, 16Bit. Общий объём аудиоданных составил 500 минут, число сегментов равно 30000. Из каждого класса в качестве обучающей выборки использовалась 1/3 записей, в качестве тестовой выборки использовались 2/3 записей.

Точность классификации в ходе эксперимента определялась следующим образом:

$$точность = \frac{количество\ верно\ классифицированных\ выборок}{общее\ количество\ выборок}$$

4.2. Результаты

Результаты тестирования различных алгоритмов классификации представлены в таблице 1. Анализ полученных результатов показывает, что предложенный в данной статье подход позволяет получить более высокую точность классификации в сравнении с другими алгоритмами классификации.

Таблица 1

Результаты эксперимента

Тип аудиосигнала	Классификаторы			
	Метод Байеса	Скрытая Марковская модель	Нейронная сеть	Одноклассовый метод МОВ
Чистая речь	90.3%	94.7%	98.2%	98.7%
Музыка	82.5%	86.0%	91.7%	93.6%
Звуки окружающей среды	65.6%	68.1%	71.6%	75.5%
Речь с фоновым шумом	73.1%	78.3%	80.4%	85.0%
Тишина	91.1%	93.2%	94.5%	96.7%

5. Заключение

В данной статье рассмотрен подход к решению задачи классификации аудиосигналов с использованием одноклассового метода опорных векторов. Рассмотренный подход применён для классификации аудиосигналов, разделённых на пять классов: чистая речь, музыка, звуки окружающей среды, речь с фоновым шумом и тишина.

Анализ полученных результатов показывает, что использование одноклассового МОВ позволяет решить ряд проблем, присущих традиционным способам классификации, и получить модель классификации с более высокой точностью в сравнении с традиционными подходами.

Литература

1. *Foote J.* //Content-base retrieval of music and audio.In: Kuo C C J, et al(eds). Multimedia Storage and Archiving Systems II. Proc of SPIE, volume 3229, 1997. 138~147.

2. *Foote J.* //An overview of audio information retrieval. ACM-Springer Multimedia Systems, 1998.

3. *Srinivasan S., Petkovic D., Ponceleon D.* // Towards robust features for classifying audio in the cude Video system. In:Proc. of the 7 th ACM Intl. Conf. on Multimedia, Orlando, 1999. 393~400.

4. *Wold E., Blum T., Keislar D., Wheaton J.* // Content-based classification, search and retrieval of audio. IEEE Multimedia Magazine, 1996. 3(3): 27~36.

5. *Liu Z., Huang J., Wang Y., Chen T.* // Audio feature extraction and analysis for scene classification. In:IEEE Single Processing Society 1997 Workshop on Multimedia signal Processing.

6. *Dong Jing, Zhao Xiaohui, Ying Na.* Pitch detection algorithm based on dyadic wavelet transforms, Journal of Jilin University, 2006. 36(6)978-981.

Янь Цзинбинь —

аспирант БГУ, научные интересы — методы и алгоритмы обработки речевых сигналов, теория метода опорных векторов, алгоритмы обнаружения ключевых слов в потоке речи.

У Ши —

аспирант БГУ, научные интересы- методы и алгоритмы обработки речевых сигналов, теория метода опорных векторов, распознавание болезней голосового тракта по голосу

Сорока Александр Михайлович —

магистрант БГУ, научные интересы- методы и алгоритмы обработки цифровых сигналов, теория метода опорных векторов, смешанные гауссовы модели

Трус Александр Александрович —

магистрант БГУ, научные интересы- методы и алгоритмы обработки речевых сигналов, скрытые марковские модели, теория метода опорных векторов

Система оперативной модификации голоса диктора на основе полувокодера

А.С. Рылов,
доктор технических наук

В.В. Киселёв,
директор ООО «Речевые технологии»

А.Г. Давыдов,
кандидат технических наук, научный сотрудник

В.А. Чижденко,
старший научный сотрудник

В работе рассматривается вопрос модификации голоса диктора на основе полувокодера. Предлагается два варианта системы модификации голоса диктора, позволяющие выполнять нелинейное изменение характеристик голоса диктора таким образом, что последующая идентификация инструментальными средствами становится невозможной.

Abctract
The paper deals with voice modification on the basis of a semi-vocoder. We suggest two variants of a voice modification system which provide non-linear modification of the voice characteristics, so that subsequent identification using instrumental means becomes impossible.

1. Введение

Модификация голоса диктора является специфической задачей, востребованной, например, системами синтеза речи по тексту. В подобных системах процедура модификации применяется для обеспечения бо̀льшего диапазона настройки голоса диктора — если необходимо из подготовленного голоса одного диктора получить голос другого диктора. В англоязычной литературе данная процедура называется voice morphing.

Другой областью применения систем модификации голоса диктора (СМГ) является защита голоса свидетеля при даче показаний в суде. В этом случае модификация голоса доктора должна проводиться в реальном времени с обеспечением минимальной задержки преобразования и невозможности обратного преобразования.

Таким образом, СМГ могут использоваться в самых различных областях, при этом к ним могут предъявляться различные требования. Классификация СМГ по их назначению представлена на ***рисунке 1***.

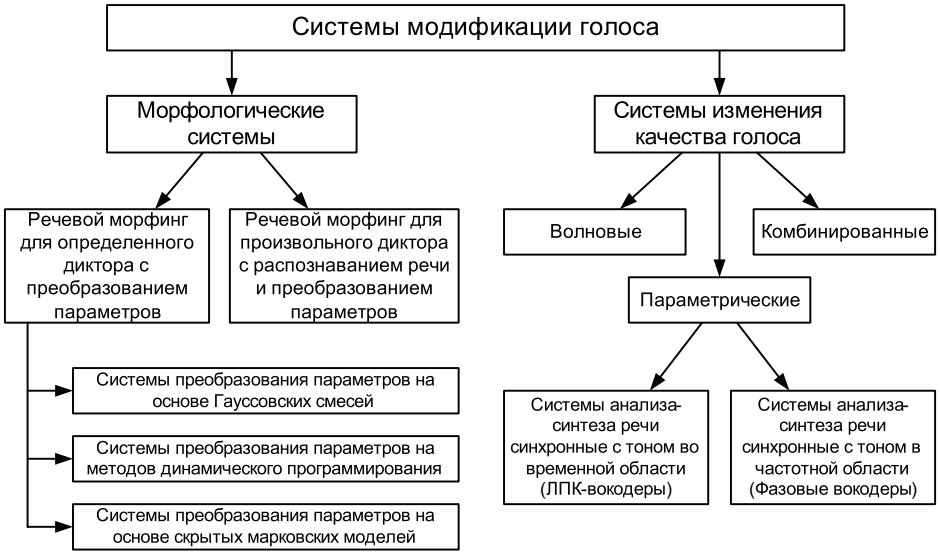

Рис. 1. Классификация систем модификации голоса диктора

Как видно из приведенного рисунка, все СМГ можно разбить на два типа: морфологические и системы для изменения качества голоса.

К первому типу относятся системы, которые позволяют трансформировать голос исходного диктора в голос определённого диктора (диктора-»мишени»). При этом осуществляется адаптация параметров речи исходного диктора к параметрам речи диктора-»мишени» для определённого речевого фрагмента (предложения, фразы), который был произнесён обоими дикторами. К наиболее сложным морфологическим системам относятся системы с распознаванием речи произвольного диктора.

Неморфологические СМГ предназначены для того, чтобы изменять качество голоса диктора, не добиваясь внешней похожести на голос какого-либо определённого диктора. По своему устройству эти СМГ можно разделить на три категории: волновые, параметрические и комбинированные, т.е. такие, в которых используются манипуляции с формой речевой волны и какими-либо акустическими параметрами речи. К параметрическим системам относятся системы анализа-синтеза речи (вокодеры), работающие во временной либо частотной областях, в которых осуществляется преобразование параметров, отвечающих за формантную структуру сигнала и за просодику речи.

Методы достижения поставленной цели — модификации голоса диктора — для морфологических и неморфологических СМГ существенно различаются. Это связано с тем, что исходные условия решения задачи являются различными. При модификации голоса диктора для систем синтеза речи по тексту доступным является большой объём подготовленных записей исходного голоса диктора (записанных с большим соотношением сигнал/шум, размеченных на фонемы и периоды основного тона). Запись голоса целевого диктора, как правило, также является качественной. Время решения задачи преобразования голоса одного диктора в голос другого диктора (время вычислений) для систем синтеза речи по тексту не является критическим и может равняться нескольким суткам. Дополнительным требованием, предъявляемым при решении такой задачи, является высокое качество полученного результата, а задача использования необратимого преобразования исходного голоса диктора в целевой голос диктора не ставится.

При модификации голоса диктора в оперативных условиях голос исходного диктора, как правило, является неизвестным, однако и голос целевого диктора обозначается в значительной степени условно. Например, требуется понизить тембр голоса или повысить тембр голоса. Дополнительным требованием в этом случае является невозможность обратимого преобразования для получения голоса исходного диктора. Вне зависимости от области применения системы модификации голоса диктора, обязательной задачей является сохранение высокой разборчивости произнесённого текста.

Данная статья посвящена рассмотрению неморфологической системы необратимой высококачественной модификации голоса диктора в оперативных условиях.

2. Система модификации голоса диктора на основе искажения линейных спектральных частот

Структурная схема системы модификации голоса диктора на основе искажения линейных спектральных частот (ЛСЧ, LSF в литературе на английском языке [1]) приведена на ***рисунке 2***. На рисунке 2 использованы следующие обозначения:

x, x' — последовательность отсчётов входного и выходного сигналов;

F_x — частота дискретизации входного и выходного сигналов;

F_s — частота дискретизации сигнала, на которой работает блок анализа;

$[s]$ — кадр сигнала;

v — мера тона для каждого кадра сигнала;

$\vec{a} = \{a_i\}$, $\vec{a}' = \{a'_i\}$ — исходный и модифицированный векторы коэффициентов линейного предсказания, где $i = \overline{1,M}$, M — порядок модели предсказания;

$[e]$, $[e']$ — исходный и модифицированный кадры ошибки предсказания;

T_e, T'_e — исходная и модифицированная длительность кадра ошибки предсказания;

$\vec{\omega} = \{\omega_1\}$, $\vec{\omega}' = \{\omega'_1\}$ — исходный и модифицированный векторы ЛСЧ.

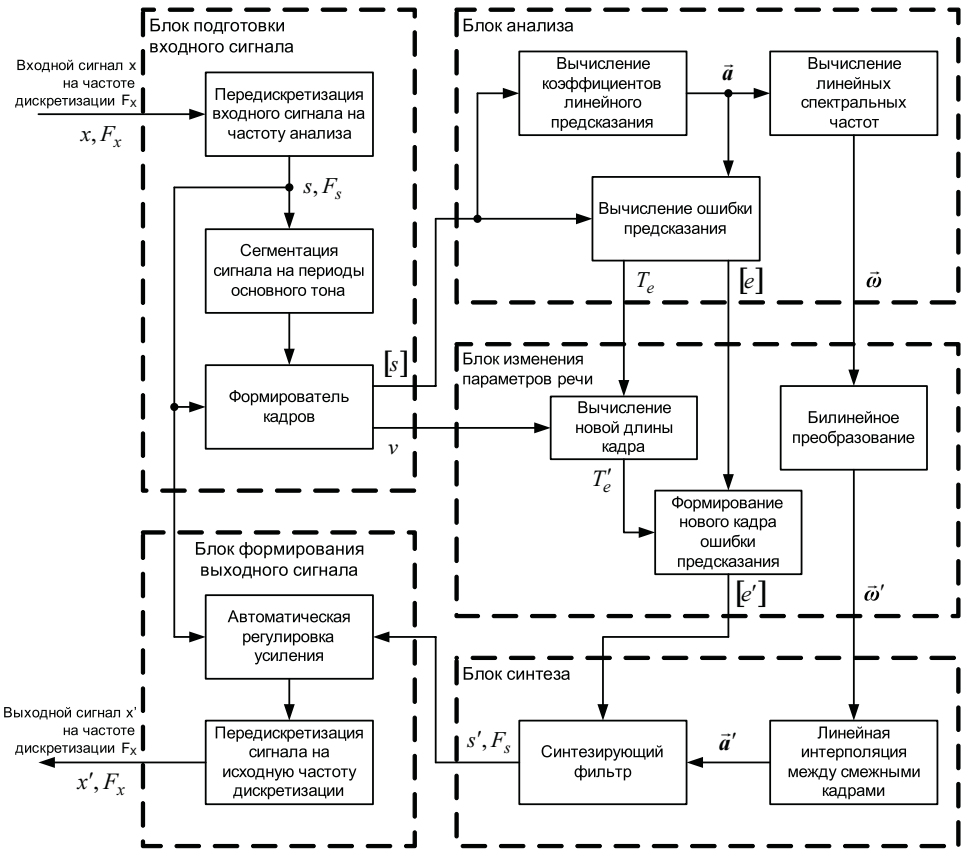

Рис. 2. Структурная схема системы модификации голоса диктора на основе искажения линейных спектральных частот

Как видно из структурной схемы, система включает блок предварительной подготовки входного сигнала, блок анализа (вычисления параметров речевого сигнала), блок модификации параметров, блок синтеза (формирования изменённого речевого сигнала), блок формирования выходного сигнала.

2.1. Блок подготовки входного сигнала предназначен для передискретизации входного сигнала x на частоту дискретизации F_s, формирования кадров $[s]$ речевого сигнала (РС) и вычисления меры тона v (степени вокализованности) каждого кадра.

Оптимальная частота анализа, определённая в ходе экспериментальных исследований, оказалась близка к 11025 Гц.

Формирование кадров анализа $[s]$ выполняется таким образом, что на вокализованных участках речи каждый кадр содержит один период основного тона. На невокализованных участках речи сигнал разделяется на кадры размером 20 мс.

Параллельно с каждым формируемым кадром сигнала $[s]$ формируется признак вокализованности кадра — v. Данная операция выполняется в соответствии

с алгоритмом, рассмотренным в [2] и модифицированным для применения в условиях реального масштаба времени.

2.2. Блок анализа служит для вычисления вектора коэффициентов линейного предсказания речи $\vec{a}$ [3, 4] для каждого кадра РС $[s]$. В соответствии с найденными коэффициентами линейного предсказания вычисляется вектор ЛСЧ $\vec{\omega}$ [1]. Вычисление кадра ошибки предсказания $[e]$ для найденных коэффициентов линейного предсказания $\vec{a}$ выполняется фильтром с конечной импульсной характеристикой, линия задержки которого не сбрасывается при переходе от одного кадра анализа к другому. Кадр ошибки предсказания $[e]$, его длительность T_e и вектор ЛСЧ $\vec{\omega}$ передаются в блок изменения параметров речи.

2.3. Блок изменения параметров речи предназначен для формирования модифицированных кадра ошибки предсказания $[e']$ и вектора ЛСЧ $\vec{\omega}'$.

2.3.1. Изменение кадра ошибки предсказания для вокализованных и невокализованных кадров речевого сигнала выполняется по немного различным формулам. Для вокализованных кадров новая длительность кадра T_e равна:

$$\hat{T}_e = T_e \cdot k_{T_0 mul} \cdot k_{\Delta T_0 mul} + \overline{T}_e \cdot k_{T_0 mul} \cdot \left(1 - k_{\Delta T_0 mul}\right),$$

где $k_{T_0 mul}$ — коэффициент изменения длительности периода основного тона; $k_{\Delta T_0 mul}$ — коэффициент изменения производной длительности периода основного тона; $\overline{T}_e$ — среднее значение длительности периода основного тона диктора, уточняемое в процессе анализа. Коэффициенты $k_{T_0 mul}$ и $k_{\Delta T_0 mul}$ задаются оператором.

Вычисление длительности нового кадра ошибки предсказания для невокализованного кадра выполняется в соответствии со следующей формулой:

$$\hat{T}_e = T_e \cdot k_{last\_mul},$$

где k_{last_mul} — коэффициент изменения длительности последнего вокализованного кадра.

Значение k_{last_mul} при запуске программы устанавливается равным $k_{T_0 mul}$, а в процессе работы на каждом вокализованном кадре постоянно обновляется в соответствии с формулой $k_{last\_mul} = \hat{T}_e / T_e$. Это позволяет маскировать ошибки классификации вокализованных кадров как невокализованных (что иногда наблюдается при малом отношении сигнал/шум на краях вокализованных участков РС).

Для более надёжного маскирования истинного поведения кривой периода основного тона длительность как вокализованных, так и невокализованных кадров ошибки предсказания может (при установке соответствующих настроек оператором) изменяться в соответствии со следующими формулами:

$$\hat{F}_0 = \frac{1}{\hat{T}_e};\ F_0' == \begin{cases} \hat{F}_0 \cdot \left[1 + \dfrac{k_{T_0 Vm}}{100} \cdot sin\left(\dfrac{\hat{F}_0}{10} + k_{T_0 Vf} \cdot 2\pi t\right)\right], & если\ \hat{F}_0 < 100; \\ \hat{F}_0 + k_{T_0 Vm} \cdot sin\left(\dfrac{\hat{F}_0}{10} + k_{T_0 Vf} \cdot 2\pi t\right), & если\ \hat{F}_0 \geq 100; \end{cases} \quad T_e' = \frac{1}{F_0'}.$$

Как видно из приведённых формул, данный вид маскирования кривой частоты основного тона напоминает её частотную модуляцию, а задаваемые оператором коэффициенты k_{T_0Vm} и k_{T_0Vf} определяют величину девиации частоты и частоту модулирующего сигнала.

Значение длительности кадра ошибки предсказания T'_e дополнительно ограничивается таким образом, чтобы получившееся значение не отличалось более чем в два раза от исходного значения длительности кадра ошибки предсказания T_e.

Формирование нового кадра ошибки предсказания в соответствии с найденным новым значением длительности выполняется: при увеличении длины кадра — линейной интерполяцией, при уменьшении длины кадра — передискретизацией.

Пример изменения частоты основного тона для задания различных значений управляющих коэффициентов приведён на ***рисунке 3***. Значение частоты основного тона оценивалось сторонним программным средством [5].

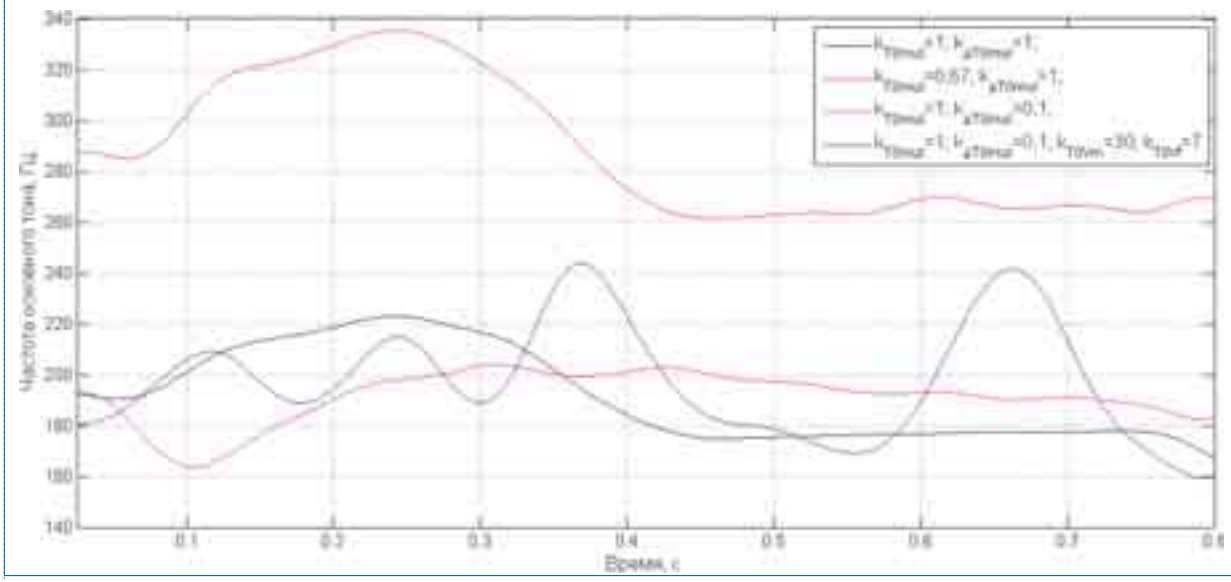

Рис. 3. Модификация частоты основного тона для различных управляющих коэффициентов

2.3.2. Выбор способа модификации огибающей спектра PC в значительной степени определяет качественные характеристики всей системы искажения голоса диктора. Очевидно, что наибольшую свободу в модификации голоса диктора можно получить при работе с фонограммами речи. При работе в реальном времени большое значение играет допустимая величина задержки между подачей исходного голоса и формированием изменённого. Чем эта задержка больше, тем более сложные алгоритмы можно использовать для модификации голоса, однако при этом система становится всё менее пригодной для использования в режиме диалога. Для модификации огибающей спектра PC в процессе разработки системы были исследованы три способа:

I. Модификация огибающей спектра РС посредством искажения индексов разностей ЛСЧ на основе билинейного преобразования [6].
II. Искажение коэффициентов ЛСЧ на основе билинейного преобразования.
III. Искажение спектра посредством изменения угловых координат полюсов.

Краткое описание первых двух способов модификации спектра и полученных результатов, а также анализ достоинств и недостатков приводится в данном подпункте описания системы модификации голоса диктора. Более подробное рассмотрение способа модификации огибающей спектра РС посредством изменения угловых координат полюсов приводится в следующем разделе.

I. Алгоритм модификации огибающей спектра РС посредством искажения индексов разностей ЛСЧ включает следующие этапы:

1. Вычисление вектора коэффициентов ЛСЧ текущего кадра РС $\vec{\omega} = \{\omega_i\}$ и вектора коэффициентов ЛСЧ, равномерно расположенных по частоте $\grave{\omega} = \{\grave{\omega}_i\} = \frac{i}{M+1}\pi$ (что соответствует сигналу с равномерным распределением спектральной плотности мощности).

2. Вычисление разности $\Delta\vec{\omega} = \vec{\omega} - \grave{\omega} \equiv \{\Delta\omega_i\} = \{\omega_i\} - \frac{i}{M+1}\pi$.

3. Искажение индексов вектора $\Delta\vec{\omega}$ как своеобразного представления шкалы частот по формуле $i' = \frac{M-1}{\pi} \cdot \left[\phi_\alpha\left(\frac{i-1}{M-1}\pi\right)+1\right]$, где $\phi_\alpha(\omega) = \omega - 2 \cdot arctg\left[\frac{\alpha \cdot sin\ \omega}{1+\alpha \cdot cos\ \omega}\right]$ — билинейное преобразование, α — управляющий коэффициент преобразования, задаваемый оператором и определяющий вид и степень модификации огибающей спектра РС. Пример функции билинейного преобразования для нескольких значений управляющих коэффициентов приведён на ***рисунке 4а***.

4. Вычисление значений последовательности $\Delta\omega_i$ с абсциссами i' в точках i посредством квадратичной интерполяции (получение модифицированного вектора разностей $\Delta\vec{\omega}' = \{\Delta\omega'_i\}$).

5. Вычисление модифицированного вектора ЛСЧ как суммы модифицированного вектора разностей $\Delta\vec{\omega}'$ и вектора $\grave{\omega}$: $\vec{\omega}' = \Delta\vec{\omega}' + \grave{\omega} \equiv \{\omega'_i\} = \{\Delta\omega'_i\} + \frac{i}{M+1}\pi$.

Пример искажения огибающей спектра РС в соответствии с этим алгоритмом приведён на ***рисунке 4б***.

При практической реализации обнаружилось, что рассмотренный алгоритм должен быть дополнен этапом проверки и коррекции значений модифицированного вектора ЛСЧ $\vec{\omega}'$. Это связано с тем, что в результирующем векторе иногда наблюдается значительное сближение двух смежных коэффициентов (что в огибающей спектра РС отражается как резкое увеличение мощности на частоте этих коэффициентов). Особенно это характерно для коэффициентов, отвечающих за первую форманту. Также при использовании рассмотренного выше алгоритма иногда получается вектор $\vec{\omega}'$, компоненты которого не следуют по возрастанию. Чаще всего наблюдается случай, когда первый коэффициент оказывается больше второго, а остальные следуют по возрастанию. Это соответствует нестабильному фильтру, полюс которого вышел

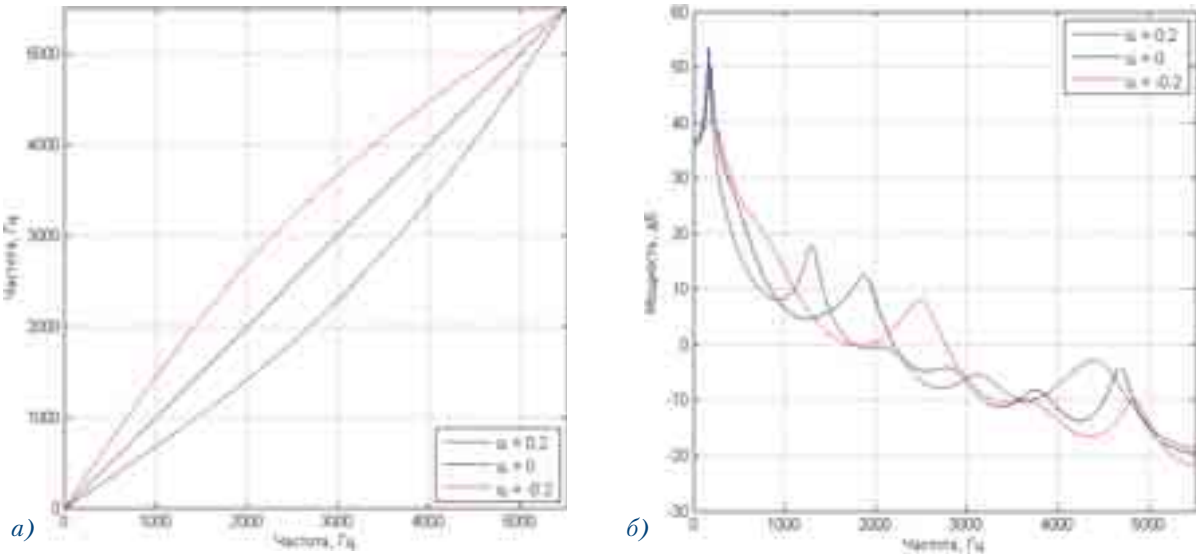

Рис. 4. Функция билинейного преобразования (а) и пример искажения огибающей спектра РС (б) для трёх значений управляющего коэффициента α

за единичную окружность. Для коррекции этих особенностей алгоритм целесообразно дополнить следующим этапом.

6. Все смежные пары коэффициентов вектора $\vec{\omega}'$, расстояние между которыми отвечает условию $\omega'_{i+1} - \omega'_i < \omega_{\Delta\min}$, заменяются на $\omega''_i = \frac{\omega'_i + \omega'_{i+1}}{2} - \frac{\omega_{\Delta\min}}{2}$ и $\omega''_{i+1} = \frac{\omega'_i + \omega'_{i+1}}{2} + \frac{\omega_{\Delta\min}}{2}$. Для тестовых записей, сделанных с частотой дискретизации 11025 Гц, и порядка модели предсказания M = 14 минимально допустимое расстояние $\omega_{\Delta min}$ было оценено как приблизительно равное 0,05.

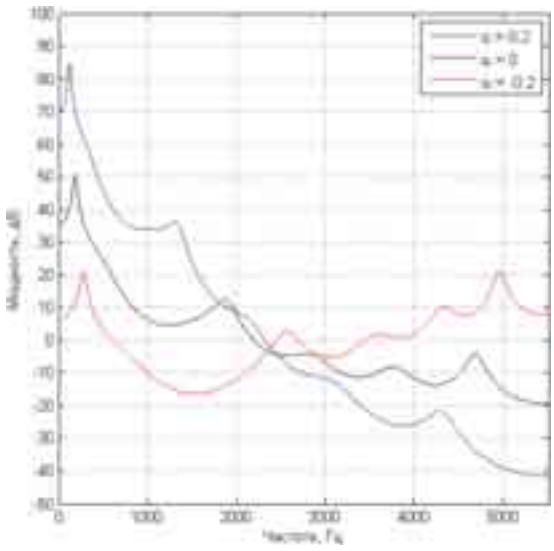

Рис. 5. Модификация огибающей спектра РС путём билинейного преобразования коэффициентов ЛСЧ

II. Второй способ модификации голоса диктора заключается в искажении непосредственно коэффициентов ЛСЧ по формуле билинейного преобразования $\phi_\alpha(\omega)$. Этот способ, в отличие от предыдущего, не требует дополнительного этапа проверки и коррекции. Пример результата искажения огибающей спектра РС данным способом приведен на ***рисунке 5***.

Как видно из приведённого рисунка, при таком способе модификации подвергается не только расположение максимумов огибающей спектра, но и наклон спектра.

2.4. Блок синтеза служит для выполнения линейной интерполяции векторов ЛСЧ между смежными кадрами, которая используется для сглаживания пере-

ходов на границах синтезируемых кадров. При вычислении каждого отсчёта синтезируемого сигнала путём линейной интерполяции определяется текущее значение вектора ЛСЧ, которое преобразуется в коэффициенты синтезирующего фильтра.

Синтезирующий фильтр, в отличие от анализирующего, применяемого при вычислении кадра ошибки предсказания, является фильтром с бесконечной импульсной характеристикой, который возбуждается модифицированным вектором ошибки предсказания $[e']$.

2.5. Блок формирования выходного сигнала служит для автоматической регулировки усиления синтезированного сигнала (необходимой при втором способе модификации огибающей спектра РС) и его передискретизации на частоту дискретизации входного сигнала.

3. Модификация голоса диктора на основе искажения полюсов передаточной характеристики

Искажение огибающей спектра РС при помощи билинейного преобразования не позволяет выполнять независимую модификацию некоторых формант без искажения остальных формант. Преодолеть это ограничение можно при использовании модификации голоса диктора на основе искажения угловых координат полюсов передаточной характеристики.

Отличие системы модификации голоса диктора на основе искажения полюсов передаточной характеристики от схемы, приведённой на рисунке 2, заключается в блоке изменения параметров речи, в котором вместо билинейного преобразования используется искажение угловых координат полюсов в соответствии со структурной схемой, приведённой на ***рисунке 6***.

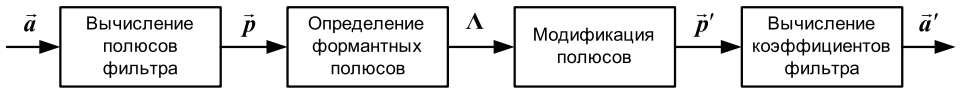

Рис. 6. Схема модификации полюсов передаточной характеристики

Как видно из данного рисунка, вектор коэффициентов предсказания $\vec{a}$ преобразуется в вектор полюсов $\vec{p} = \{p_k\}$, $k = \overline{1,K}$, из которых рассматриваются только те, которые находятся в верхней полуплоскости. На множестве найденных полюсов p_k определяются те, которые определяют частоты и амплитуды формант РС (определение формантных полюсов). Результатом этой процедуры является матрица соответствия формант и полюсов Λ. В соответствии с матрицей Λ угловые координаты полюсов подвергаются модификации для получения нового вектора полюсов $\vec{p}' = \{p'_k\}$, который затем преобразуется в новый вектор коэффициентов фильтра $\vec{a}'$. Процедуры вычисления полюсов фильтра по его коэффициентам и коэффициентов по полюсам достаточно подробно рассмотрены в литературе [7].

Процедура определения формантных полюсов включает следующие этапы:

1. Вычисление матрицы расстояний D между полюсами p_k и полюсами p_{Fn} ($n = \overline{1,N}$, $N = 5$), имеющими единичную амплитуду и угловые координаты, соответствующие середине частотных диапазонов формант. Для удобства дальнейшего рассмотрения будем считать,

что полюса p_k отсортированы в порядке возрастания их аргумента: $arg(p_{k-1}) < arg(p_k) < arg(p_{k+1})$.

Частотные диапазоны изменения формант хорошо изучены и могут быть найдены, например, в [8, 9]. При разработке СМГ были использованы частотные диапазоны формант, приведённые в таблице 1.

Таблица 1

Частотные диапазоны формант

Номер форманты	F_1	F_2	F_3	F_4	F_5
Частотный диапазон, Гц	100–1000	500–3000	1500–4000	2500–5300	3500–5500

Как видно из таблицы 1, частотные диапазоны формант были несколько увеличены (по сравнению с данными, приводимыми в [8, 9]) для исключения ошибок. Частотный диапазон пятой форманты был ограничен 5500 Гц в связи с использованием для анализа РС с частотой дискретизации, равной 11025 Гц. Таким образом, матрица расстояний D вычисляется в соответствии со следующей формулой:

$$\mathbf{D} = [d_{nk}] = \left[\left| p_{Fn} - p_k \right| \cdot \left(1 - \left| p_k \right|\right)\right].$$

Для полюсов, лежащих за границами частотных диапазонов соответствующей форманты, расстояние d_{nk} принималось равным бесконечности. Результирующая матрица расстояний D может принять, например, следующий вид:

$$\mathbf{D} = \begin{vmatrix} d_{11} & \infty & \infty & \infty & \infty & \infty \\ d_{21} & d_{22} & d_{23} & \infty & \infty & \infty \\ \infty & \infty & d_{33} & \infty & \infty & \infty \\ \infty & \infty & d_{43} & d_{44} & d_{45} & \infty \\ \infty & \infty & \infty & d_{54} & d_{55} & d_{56} \end{vmatrix}.$$

С учётом вышеизложенного, задача определения формантных полюсов сводится к выбору такого множества Λ элементов матрицы расстояний D, которое удовлетворяет следующим условиям.

1. Из каждой строки n матрицы расстояний D в конечное множество выбирается один элемент d_{nk}.

2. Так как полюса p_k отсортированы в порядке возрастания их аргумента, из каждой строки матрицы расстояний D может быть выбран только тот элемент d_{nk}, индекс столбца k которого лежит в диапазоне $[n; n + K - N]$.

3. Индексы столбцов выбранных элементов должны быть отсортированы в порядке возрастания: для любых двух различных элементов d_{nk} и d_{ml}, если $n < m$, то $k < l$; обратное также верно.

4. Сумма элементов d_{nk} множества Λ должна быть минимальной: $\rho = \sum_{n=1}^{N} d_{nk} \to \min$.

Перечисленные условия являются достаточно жёсткими и, как правило, оставляют очень мало различных вариантов для определения формантных полюсов, вследствие чего даже применение алгоритма полного перебора является вычислительно приемлемым.

Для модификации голоса диктора используются полюса вектора $\vec{p}_{\Lambda} = \{p_{\lambda}\}$, индексы λ которых совпадают со вторыми индексами элементов d_{nk} множества Λ.

Использование приведённых выше условий позволяет приблизительно в 90% случаев правильно определить формантные полюса. Однако на переходных участках РС, при малом отношении сигнал/шум, встречаются случаи, когда формантные полюса определяются неверно. Признаком таких случаев является резкое изменение угловой координаты формантного полюса от предыдущего кадра к текущему. Расстояние между формантными полюсами Θ предыдущего кадра $\vec{p}_{\Lambda}^{\,t-1}$ и текущего кадра $\vec{p}_{\Lambda}^{\,t}$ определяется так:

$$\Theta\left(\vec{p}_{\Lambda}^{\,t-1}, \vec{p}_{\Lambda}^{\,t}\right) = \sum_{\lambda=1}^{N} \ln\left(\left|p_{\lambda}^{t} - p_{\lambda}^{t-1}\right| + 1\right).$$

Если расстояние оказывается больше порогового значения (в ходе экспериментов определённого как приблизительно равное единице), то путём последовательного удаления из рассмотрения полюсов p_{Fb}, $b = \overline{2,N}$ выбирается тот вариант определения формантных полюсов, при котором достигается минимальное значение суммы ρ.

Для модификации форматных полюсов можно использовать самые различные алгоритмы, однако даже модификация голоса диктора путём задания смещений угловых координат формантных полюсов позволяет получить вполне удовлетворительные результаты.

4. Заключение

Разработанная система модификации голоса диктора позволяет оперативно решать поставленную задачу таким образом, что восстановление исходного голоса диктора становится невозможным. Это достигается за счёт использования нелинейного изменения длительности периода основного тона, которое управляется коэффициентами, задающими среднюю длительность, дисперсию и модуляцию периода основного тона. Таким образом, голос диктора может быть с различной степенью преобразован в низкий либо высокий, монотонный либо эмоциональный. Для дополнительной защиты от возможности восстановления исходного голоса диктора коэффициенты могут задаваться как функции времени.

Проведённые предварительные эксперименты показали, что даже при значительной модификации система обеспечивает высокое качество и разборчивость модифицированного голоса.

Билинейное преобразование линейных спектральных частот, использованное в рассмотренной системе для модификации формантных частот, показывает хорошие результаты искажения голоса.

Модификация голоса диктора на основе искажения полюсов передаточной характеристики хотя и является более сложной для реализации, чем билинейное преобразование, но обеспечивает бо̀льшие возможности искажения голоса.

Литература

1. *Huang X., Acero A., Hon H.-W.* Spoken Language Processing: A Guide to theory, algorithm, and system development. — New Jersey: Prentice Hall. — 2001. — 1008 p.
2. *Лобанов Б.М., Давыдов А.Г.* Алгоритм высокоточной разметки на питчи элементов компиляции для синтеза речи по тексту // Компьютерная лингвистика и интеллектуальные технологии (Диалог 2007): труды Международной конференции, Бекасово, 30 мая-3 июня 2007г. / Институт проблем информатики РАН; редкол.: Л.Л. Иомдин [и др.]. М.: Издательский центр РГГУ, 2007. С.388-392.
3. *Маркел Дж.Д., Грэй А.Х.* Линейное предсказание речи: Пер. с англ./Под ред. Ю.Н. Прохорова, В.С. Звездина. М.: Связь, 1980. 308 с., ил.
4. *Марпл-мл. С.Л.* Цифровой спектральный анализ и его приложения: Пер. с англ. М.: Мир, 1990. 584 с., ил.
5. *Camacho A., Harris J.G.* A sawtooth waveform inspired pitch estimator for speech and music // The Journal of the Acoustical Society of America. — 2008. — Vol. 124, issue 3. — P. 1638–1652.
6. *G. Strecha, O. Jokisch, M. Eichner, R. Hoffmann* Codec Integrated Voice Conversion for Embedded Speech Synthesis // Interspeech 2005, Lisbon, Portugal, September 4–8, 2005. P. 2589–2592.
7. *Самарский А.А., Гулин А.В.* Численные методы: Учебн. пособие для вузов. М.: Наука. Гл. ред. физ.-мат. лит., 1989. 432 с.
8. *Фант Г.* Акустическая теория речеобразования: Пер. с англ. Л.А. Варшавского и В.И. Медведева / Под ред. В.С. Григорьева. М.: Наука, 1964. 284 с., ил.
9. *Михайлов В.Г., Златоустова Л.В.* Измерение параметров речи / Под. ред. М.А. Сапожкова. М.: Радио и связь, 1987. 168 с., ил.

Рылов Александр Александрович —

доктор технических наук, Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники».

Киселёв Виталий Владимирович —

директор ООО «Речевые технологии», г. Минск. С 1999 г. профессионально занимается системами синтеза и распознавания речи, диалоговыми речевыми системами. Автор более 25 научных публикаций в области речевых технологий. Основные научные интересы связаны с системами обработкой и анализом текста и речи, системами синтеза, распознавания речи, поиска ключевых слов.

Давыдов Андрей Геннадьевич —

кандидат технических наук, старший научный сотрудник Академии управления при Президенте Республики Беларусь. Основные научные интересы связаны с областью цифровой обработки сигналов, анализом, сжатием и распознаванием речи. Автор более 20 научных публикаций и 3 патентов.

Чижденко Виктор Анатольевич —

старший научный сотрудник ООО «Речевые технологии», г. Минск. Основные научные интересы связаны с областью цифровой обработки сигналов, анализом, сжатием и распознаванием речи. Автор более 10 научных публикаций и 1 патента.

Обращаем Ваше внимание на факт безвременной кончины Рылова Александра Александровича.

Конверсия голоса с использованием модели сепарации речевого сигнала на компоненты «гармоники+шум» и переходные фреймы

А.Н. Павловец,
аспирант
М.З. Ливщиц,
кандидат технических наук, доцент
Д.С. Лихачёв,
кандидат технических наук, доцент
А.А. Петровский,
доктор технических наук, профессор

В статье представлена система конверсии голоса, основанная на модели сепарации речевого сигнала на «гармоники+шум» и переходные фреймы с раздельной конверсией для каждой компоненты модели. Преимущество системы конверсии голоса в данном случае складывается из достоинств анализа-синтеза гармонической модели с достоинствами анализа и конверсии переходных сегментов во временной области. Неформальные тесты прослушивания показали, что узнаваемость диктора соответствует приблизительно 70%, реконструированная речь характеризуется достаточно высокой разборчивостью.

Abctract
The voice conversion system based on the harmonic+noise speech signal model is presented in the given paper. One of the critical tasks in voice conversion framework is speaker parameter estimation. Here the method based on the Harmonic-Noise-Transient (HNT) decomposition of speech is proposed with the idea of processing each of the components separately and further converting them separately.

1. Введение

Проблема конверсии голоса становится очень популярной в мире. Сущность конверсии заключается в модификации голоса диктора, являющегося в данном случае источником, в голос другого (целевого) диктора. Актуальность темы исследований обусловлена широким применением устройств конверсии голоса в мультимедиа-системах реального времени: синтез речи по тексту (устранение «компьютерного акцента»); виртуальное дублирование (восстановление звуковых дорожек кинофильмов); защита свидетелей (применение в судебной практике); оперативная смена диктора в коммуникационных системах (озвучивание SMS-сообщений в мобильных телефонах) [1].

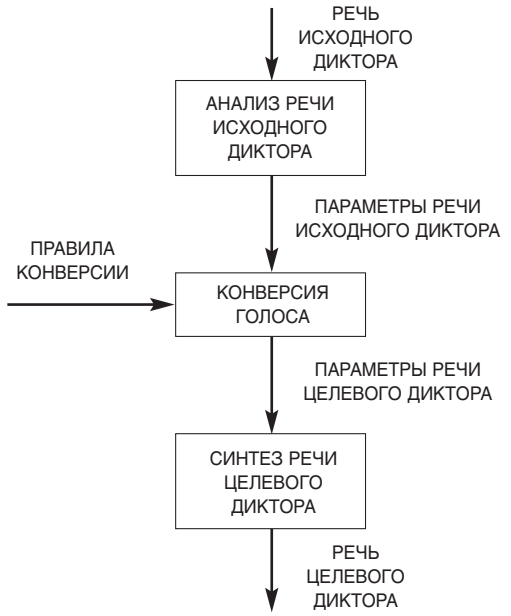

Рис. 1. Типовая схема процесса конверсии голоса

Общий алгоритм процесса конверсии показан на ***рис. 1***.

Процесс конверсии голоса можно разбить на два этапа: обучения и конверсии. На первом из них (этапе обучения) выделяется множество характеристических параметров исходного и целевого дикторов и определяются правила конверсии, посредством которых параметры исходного диктора будут преобразовываться в параметры целевого диктора. На втором этапе (этапе конверсии) характеристики речи исходного диктора преобразуются с использованием правил, определённых на первом этапе.

При реализации системы конверсии голоса требуется решить два основных вопроса: как и какие параметры извлекать из речевого сигнала, подлежащего преобразованию, и как модифицировать эти параметры таким образом, чтобы преобразованная речь была похожа на речь целевого диктора. В работе [2] улучшен подход [3], который был основан на модели «гармоники+шум», путём использования декомпозиции анализируемой речи на вокализованную и шумоподобную компоненты. Ранее подобный метод был применён в области кодирования речи [4, 5]. Дальнейшие исследования показали, что модель [2] не является достаточной в полной мере, поскольку с её помощью нельзя корректно анализировать переходные сегменты речи. Это послужило причиной дополнения модели [2] режимом анализа переходных сегментов.

Определение вокализованных областей в речевом сигнале является довольно сложной задачей. Вокализованность может определяться для анализируемого сегмента речи в целом [6], также можно находить максимальную частоту вокализованности [3] или принимать решение по вокализованности в какой-либо полосе частот [7]. Проблема заключается в том, что принимаемое решение обычно имеет два варианта: область либо вокализована, либо нет. С точки зрения процесса речеобразования более точным было бы рассматривать вокализованную речь как сумму вокализованной и шумоподобной составляющих. В [8] был предложен метод декомпозиции на вокализованную и шумоподобную компоненты, который применяется к сигналу —

остатку линейного предсказания. Идея заключается в использовании итеративного алгоритма, основанного на последовательном применении ДПФ/ОДПФ для определения шумовой компоненты.

Ещё один метод декомпозиции заключается в использовании гармонического фильтра, параметры которого изменяются в зависимости от частоты основного тона [9]: речевой сигнал взвешивается окном, при этом в окно должно укладываться некоторое целое количество периодов основного тона. Так же, как в работах [8] и [9], в подходе, представляемом в данной статье, считается, что вокализованная и шумоподобная составляющие присутствуют во всём диапазоне частот. Спектральный анализ проводится в области гармоник фундаментальной частоты речи, для этого ДПФ было модифицировано таким образом, чтобы учитывать изменение её контура. Точность определения параметров модели, а именно частоты основного тона, амплитуд и фаз гармоник, повышена за счёт использования метода анализа через синтез. Предполагается, что гармоническая компонента определяется суммой гармоник фундаментальной частоты с изменяющимися во времени амплитудами и фазами. Декомпозиция выполняется во временной области, шумоподобная компонента определяется разностью между оригинальной речью и синтезированной гармонической компонентой.

Для модификации параметров речи было предложено большое количество статистических подходов. Популярность приобрели методы, основанные на векторном квантовании. В этом случае определение правил конверсии представляет собой установление соответствия между кодовыми книгами, представляющими акустические классы дикторов. В [10] был описан метод, основанный на жёсткой кластеризации и дискретном соответствии между кодовыми книгами. Получаемый характеристический вектор y'_t в момент времени t определяется путём квантования исходного характеристического вектора и подстановки вместо него соответствующей центроиды из кодовой книги целевого диктора.

Однако жёсткая кластеризация подразумевает большую ошибку квантования. В данной работе представлена модификация метода [10]. В зависимости от типа сегмента речи используются различные кодовые книги. Целью работы является обеспечение лучшего качества конверсии голоса с использованием модели сепарации речевого сигнала на компоненты «гармоники+шум» и переходные фреймы.

2. Модель сепарации речевого сигнала на компоненты «гармоники+шум» и переходные фреймы

2.1. Гармонический анализ

В предлагаемой модели речевой сигнал представляется так:

$$s(i) = h(i) + r(i) + t(i)\,, \qquad (1)$$

где $h(i)$ — вокализованная (гармоническая) компонента, $r(i)$ — сигнал-остаток вокализованной компоненты (шум), $t(i)$ — переходный фрейм. Режим работы анализатора речевого сигнала полностью определяется наличием либо отсутствием основного тона (***рис. 2***).

Данная модель была успешно апробирована в области кодирования речи и аудиосигнала, например, [11, 12]. Первая из упомянутых работ представляет гибридный кодер речи, который сочетает параметрический кодер, работающий в частотной области (для

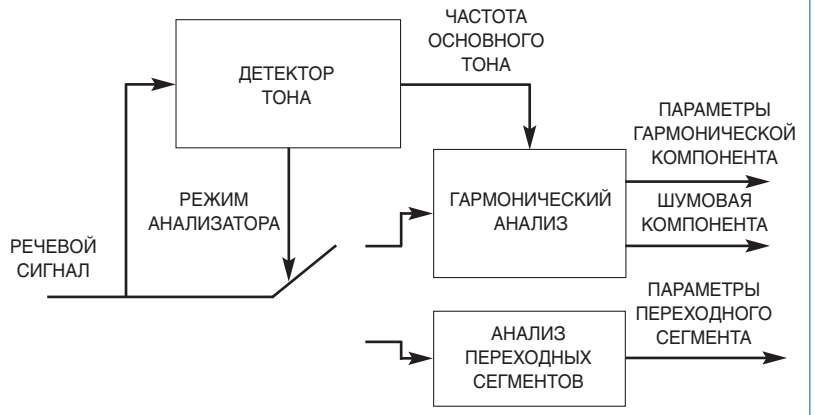

Рис. 2. Схема декомпозиции речевого сигнала

случаев стационарной вокализованной и стационарной невокализованной речи), с кодером формы сигнала, работающим во временной области (для переходных сегментов). Вторая работа использует сегментацию аудиосигнала на три различных сигнала: сигнал, моделирующий синусоидальную составляющую, сигнал, который моделирует все переходные сегменты, и шумовой сигнал.

Гармоническую составляющую речевого сигнала можно определить следующим образом:

$$h(i) = \sum_{k=1}^{K} A_k \cos\left(k \sum_{i=0}^{N-1} \frac{F_0(i)}{F_s} + \theta_k\right) \qquad (2)$$

где A_k — амплитуда k-ой гармоники, K — количество гармоник, $F_0(i)$ — мгновенная частота основного тона, θ_k — начальная фаза k-ой гармоники, F_s — частота дискретизации, N — длина сегмента.

Ядром гармонического анализа является процедура ДПФ, согласованного с изменением контура частоты основного тона (Pitch-Tracking Discrete Fourier Transform — PTDFT) [5]. Модифицированное ДПФ для анализа в области гармоник определяется так:

$$H_n(k) = \sum_{i=0}^{K-1} s_n(i) \exp\left(j \frac{2\pi k i}{F_s}\left(F_0 + \frac{\Delta F_0 i}{2N}\right)\right) w_n(i), \quad j = \sqrt{-1}, \qquad (3)$$

где $s_n(i)$ — i-й отсчёт n-го сегмента, F_0 — фундаментальная частота, ΔF_0 — приращение фундаментальной частоты, $w_n(i)$ — временное окно.

Таким образом, можно рассчитать амплитуды и фазы гармоник:

$$A_n(k) = \frac{\sqrt{\mathrm{Re}^2(H_n(k)) + \mathrm{Im}^2(H_n(k))}}{\sum_{i=0}^{L-1} w_n(i)},$$

$$\theta_n(k) = -arctg \frac{\mathrm{Im}(H_n(k))}{\mathrm{Re}(H_n(k))}.$$

Неортогональное ядро преобразования (3) может вызывать просачивание энергии в соседние спектральные отсчёты. Для устранения данного недостатка предлагается использовать времязависимое временное окно, форма которого пересчитывается синхронно с контуром изменения частоты основного

тона. Хорошие результаты получаются, если использовать в качестве прототипа окно Кайзера [13]:

$$w_n(i) = \frac{I_0\left(\beta\sqrt{1-\left[(2x-L_n+1)/(L_n-1)\right]^2}\right)}{I_0(\beta)},$$

где $i=0...N\text{-}1$, N — длина окна, β — параметр окна, $I_0(\ )$ — функция Бесселя нулевого порядка, x — функция, отражающая времязависимые характеристики:

$$x = \frac{a_{2,n}(N-1-i)^2 + a_{1,n}(N-1-i)}{a_{2,n}(N-1)+a_{1,n}},$$

где $a_{2,n}$ и $a_{1,n}$ — параметры, обеспечивающие линейное изменение фундаментальной частоты:

$$a_{2,n} = \frac{2\pi\Delta F_0}{F_s N},\ a_{1,n} = \frac{2\pi F_0}{F_s}.$$

2.2. Алгоритм определения параметров гармоник и частоты основного тона в цикле с обратной связью

Было показано [4, 5], что гармонический анализ с использованием PTDFT, а следовательно, и декомпозиция речевого сигнала являются достаточно корректными в случае точного определения значения частоты основного тона. Решение, которое предлагается здесь, подразумевает одновременное определение фундаментальной частоты и параметров (амплитуд и фаз) гармоник в цикле с обратной связью. Алгоритм грубой оценки частоты основного тона работает во временной области и основан на расчёте нормализованной автокорреляционной функции в комбинации с постобработкой на базе динамического программирования. В ходе предварительной оценки контура частоты основного тона производится низкочастотная фильтрация с частотой среза 1 кГц.

Информация о контуре фундаментальной частоты получается путём поиска максимумов нормализованной автокорреляционной функции (НАКФ):

$$\psi(k) = \frac{\sum_{j=1}^{N} s_j s_{j+k}}{\sqrt{\sum_{j=1}^{N} s_j^2 \sum_{j=1}^{N} s_{j+k}^2}},$$

где k — номер отсчёта НАКФ, соответствующий периоду основного тона. Максимумы, расположенные в пределах допустимых значений периода тона (в данном случае $16 \le k \le 160$), рассматриваются как кандидаты. Для того чтобы отбросить ложные пики, не рассматриваются кандидаты со значением автокорреляционной функции менее 30% от максимального на этом фрейме.

Следующий шаг алгоритма — отслеживание контура фундаментальной частоты, основанное на динамическом программировании (ДП) [14]. Так, для каждого кандидата рассчитывается

функция стоимости с учётом прошлой информации о контуре частоты основ ного тона: $D_{i,j} = d_{i,j} + \min_{k \in I_{i-1}} \{D_{i-1,k} + \delta_{i,j,k}\}$, где $d_{i,j}$ — локальная стоимость j-го кандидата в момент времени i, $\delta_{i,j,k}$ — стоимость перехода от k-го кандидата в момент времени i-1 к j-му кандидату в момент времени i ($1 \le j \le I_i$), I — количество кандидатов.

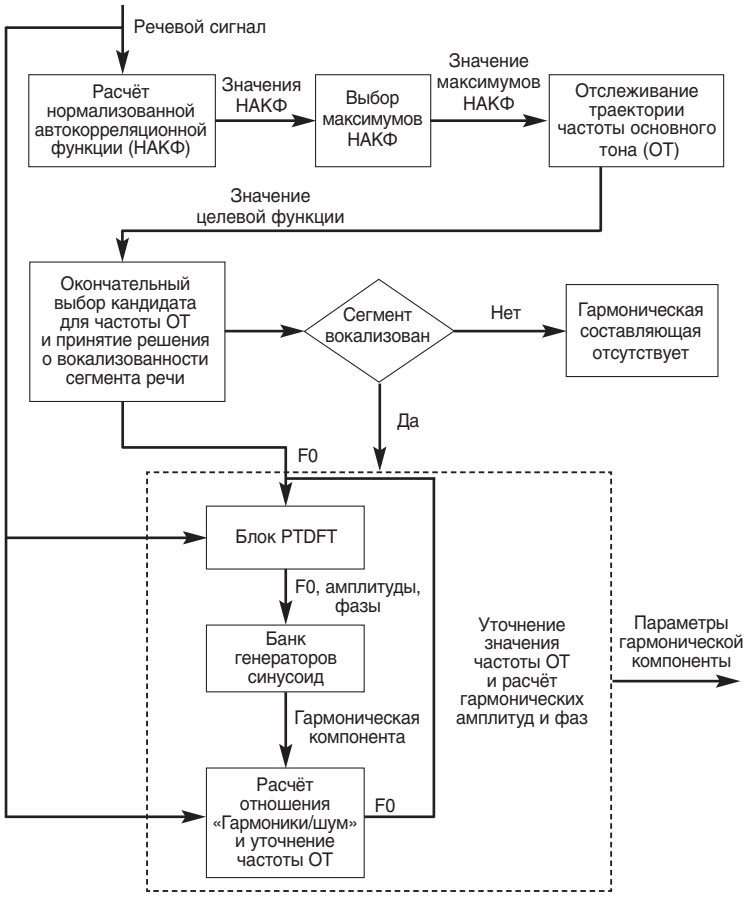

Рис. 3. Определение параметров гармоник и уточнение значения частоты основного тона в цикле с обратной связью

Целью данной процедуры является определение максимально гладкого контура частоты основного тона. После процедуры ДП в качестве предварительной оценки частоты основного тона на анализируемом сегменте выбирается кандидат j с минимальной стоимостью $D_{i,j}$. На ***рис. 3*** показана блок-схема алгоритма определения параметров гармоник и уточнения значения частоты основного тона в цикле с обратной связью.

Показателем точности определения параметров в данном случае может служить отношение «гармоники / шум»: $HNR = 10\lg(E_h/E^n)$, где E_H — энергия гармонической составляющей, синтезированной в соответствии с (2), E_H — энергия шумовой составляющей. Последняя определяется как разность между оригинальной речью и синтезированной гармонической компонентой: $r(i) = s(i) - h(i)$. Цикл с обратной связью для уточнения значения фундаментальной частоты выполняется после этапа грубой оценки. Целью данной процедуры является нахождение такого оптимального значения фундаментальной частоты, которое будет максимизировать значение HNR:

$$F_0^{opt} = arg\,max\,(\mathrm{HNR}(F_0)),\ F_{0min} \le F_0 \le F_{0max}.$$

Ядром данного процесса является PTDFT, а поиск оптимума осуществляется с помощью метода золотого сечения.

Для анализа переходных сегментов был использован подход ACELP в соответствии с рекомендациями ITU-T G.729 [15]. Пример декомпозиции речевого сигнала приведен на ***рис. 4***.

3. Конверсия голоса

3.1. Этап обучения

В работах [2, 16] была сделана попытка решения проблемы конверсии голоса с помощью таких методов преобразования спектра, как сопоставление кодовых книг амплитуд гармоник [16] и преобразование линейных спектральных частот (Line Spectral Frequencies — LSF) с помощью модели гауссовых смесей [2]. Принимая во внимание возможность использования модели (2), естественным было бы применить различные функции конверсии [17] для огибающих спектра каждой компоненты. При этом этап обучения осуществляется раздельно для составляющих модели (1).

На ***рис. 5*** показаны фазы обучения кодовых книг спектральных векторов.

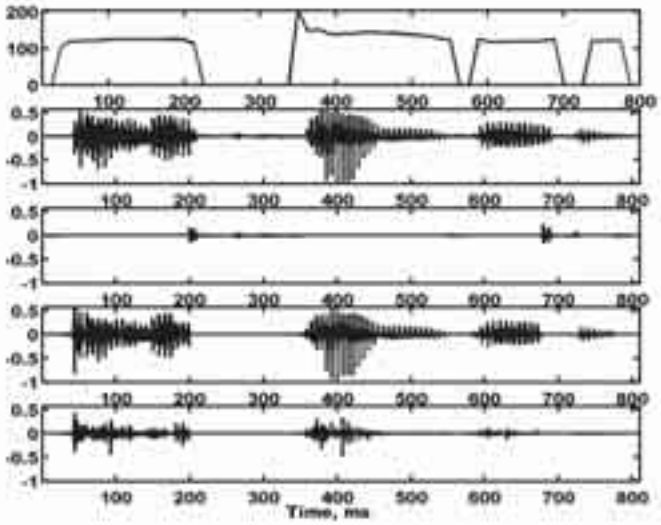

Рис. 4. Пример декомпозиции речевого сигнала. Сверху вниз: контур частоты основного тона, оригинальный речевой сигнал, переходные фреймы, гармоническая компонента, шумовая компонента

Из гармонического анализа для процесса конверсии берутся такие параметры, как спектральная огибающая, представленная LSF-коэффициентами, и фундаментальная частота F_0. Анализ переходных фреймов, осуществляемый с помощью метода ACELP [15], предоставляет для модификации LSF-коэффициенты фильтра, моделирующего вокальный тракт, период основного тона T_0, коэффициенты усиления последовательностей адаптивного и фиксированного возбуждения G_a и G_f соответственно.

Для осуществления преобразования таких параметров, как фундаментальная частота F_0, период основного тона T_0, коэффициенты усиления последовательностей адаптивного и фиксированного возбуждения G_a и G_f, использовался метод линейного преобразования математического ожидания и дисперсии. При этом предполагается, что математические ожидания этих параметров содержат существенную часть информации, специфической

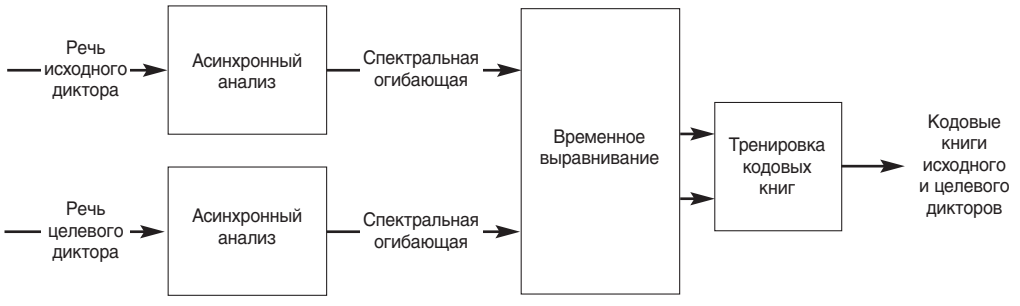

Рис. 5. Фазы этапа обучения кодовых книг

для каждого диктора. Предполагается также, что значения параметров каждого диктора подчиняются распределению Гаусса и имеют характерные средние значения и отклонения.

Обозначив модифицируемый параметр как $p_t^{S->T}$, можно определить линейное преобразование следующим образом [20]:

$$p_t^{S->T} = \frac{p_t^S - \mu^S}{\sigma^S}\sigma^T + \mu^T,$$

где p_t^T, p_t^S — один из параметров целевого и исходного дикторов соответственно, σ^S, μ^S, σ^T, μ^T — среднеквадратическое отклонение и математическое ожидание соответствующего параметра исходного и целевого дикторов соответственно.

В ходе обучения для установления более точного соответствия спектральных векторов исходного и целевого дикторов используется алгоритм динамической временной трансформации (DTW — Dynamic Time Warping) [20].

3.2. Процесс динамической временной трансформации

Предположим, имеется две последовательности наборов LSF-параметров: для целевого диктора — LSF^{tag} и для исходного — LSF^{source}. Необходимо выполнить выравнивание данных двух последовательностей по времени, т.е. таким образом сопоставить их по длине путём вставки или удаления элементов в LSF^{source}, чтобы общее среднеквадратическое отклонение между элементами этих последовательностей стремилось к минимальному значению.

Общее отклонение (расстояние) между последовательностями LSF^{tag} и LSF^{source} можно определить как $D\left(LSF^{tag}, LSF^{source}\right) = \sum_{s=1}^{Np} d\left(p_s\right)$, где $d(p_s)$ — расстояние между LSF_n^{source} и LSF_m^{tag}, N_p — количество элементов в пути. Тогда процесс нахождения оптимального пути сводится к минимизации общего отклонения: $P = arg\min_P \left(D\left(LSF^{tag}, LSF^{source}\right)\right)$.

Таким образом, задача сводится к нахождению функции выравнивания (пути):

$$P = p_1, \ldots, p_s, \ldots, p_{Np}, \quad p_s = (n_s, m_s),$$

каждое значение которой показывает, какой элемент последовательности LSF^{source} следует удалить, а какой вставить (***рис. 6***), чтобы достигнуть минимального значения общего отклонения.

Алгоритм для нахождения функции выравнивания (оптимального пути) может быть описан следующим образом:

Шаг 1. Вычислить матрицу локальных среднеквадратических отклонений d, каждый элемент которой является евклидовым расстоянием между двумя

соответствующими элементами последовательностей LSF^{source} и LSF^{tag} и определяется по следующему выражению:

$$d(n,m)=\sqrt{\sum_{k}^{10}\left(LSF_n^{source}(k)-LSF_m^{tag}(k)\right)^2},\ n=\overline{1,N},\ m=\overline{1,M}.$$

Шаг 2. Вычислить матрицу весов D, каждый элемент которой характеризует вклад соответствующего элемента матрицы d в общее среднеквадратическое отклонение.

Шаг 2.1. Положим начальное условие $D(1,1) = d(1,1)$. Вычислить первую строку матрицы D:

$$D(n,1)=D(n-1,1)+d(n,1),\ n=\overline{1,N}.$$

Шаг 2.2. Вычислить первый столбец матрицы D:

$$D(1,m)=D(1,m-1)+d(1,m),\ m=\overline{1,M}.$$

Шаг 2.3. Далее, двигаясь по матрице d слева направо снизу вверх, вычисляются следующие элементы матрицы D:

$$D(n,m)=\min\left[D(n,m-1),D(n-1,m-1),D(n-1,m)\right]+d(n,m),\ n=\overline{1,N},\ m=\overline{1,M}.$$

В процессе вычисления для каждой ячейки матрицы запоминается индекс соседней ячейки, которая вносит минимальный вклад в общую ошибку.

Шаг 3. Анализируя матрицу D в направлении от $D(N,M)$ до $D(1,1)$ и учитывая определённые на предыдущих этапах индексы ячеек, которые вносят меньший вклад в общее отклонение по сравнению с соседними, определяется наилучший путь $P = p_1, \ldots, p_k, \ldots, p_M$ с точки зрения минимизации величины общего отклонения.

Полученный в результате работы алгоритма путь $P = p_1, \ldots, p_k, \ldots, p_M$ является функцией сопоставления для обрабатываемых последовательностей, которая показывает, какие элементы необходимо удалить в исходной последовательности, а какие добавить.

Например, на ***рис. 7*** отображена функция сопоставления для двух наборов LSF-параметров исходного и целевого мужских дикторов с длиной $N=57$ и $M=68$ соответственно. Данным наборам соответствует фраза «Испорченный контакт».

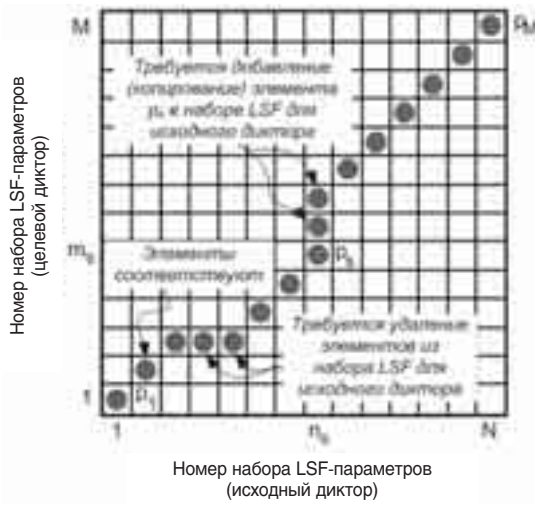

Рис. 6. Иллюстрация алгоритма динамической временной трансформации

3.3. Этап конверсии голоса

Функция конверсии векторов LSF для составляющих модели (1) имеет следующий вид:

$$F(x_t)=\sum_{i=1}^{N}p_i c_i,$$

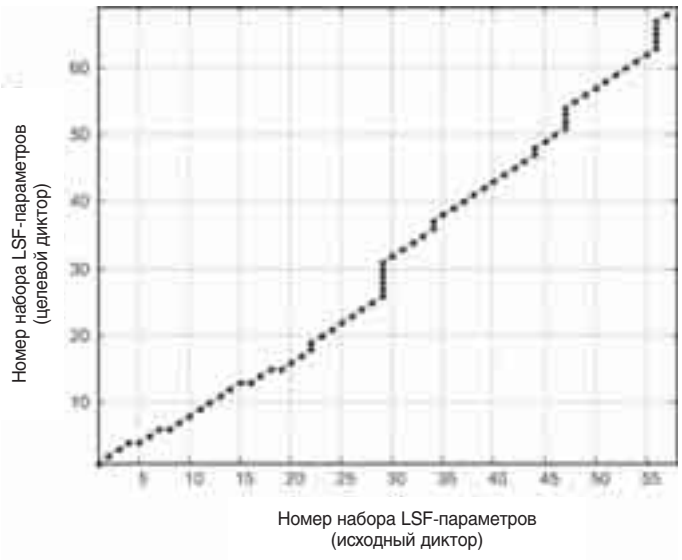

Рис. 7. Функция сопоставления для двух наборов LSF-параметров исходного и целевого мужских дикторов с длиной N=57 и M=68 соответственно

где p_i — вес, характеризующий вероятность принадлежности вектора x_t к i-му акустическому классу, представленному в кодовой книге размерности N центроидой c_i.

$$p_i = \frac{e^{-d_i}}{\sum_{j=1}^{N} e^{-d_j}},$$

где d_i — мера искажения: $d_i = \sum_{k=1}^{m} v_k |c_i - x_t|$.

Здесь величина m представляет собой размерность вектора, v_k — вес, рассчитанный по формуле обратного гармонического среднего, с помощью которого учитывается перцептуальный фактор близости смежных LSF:

$$v_k = \frac{1}{\omega_k - \omega_{k-1}} + \frac{1}{\omega_{k+1} - \omega_k},$$

где ω_k — k-й коэффициент LSF, ω_0=0, $\omega_{m+1} = \pi$.

На ***рис. 8*** показано, как выполняется процесс конверсии.

На вход системы поступает речевой сигнал, оцифрованный с частотой дискретизации 8 кГц. Детектор голосовой активности (VAD), реализованный в соответствии с [18], проверяет сегмент входного сигнала на наличие речи. Если

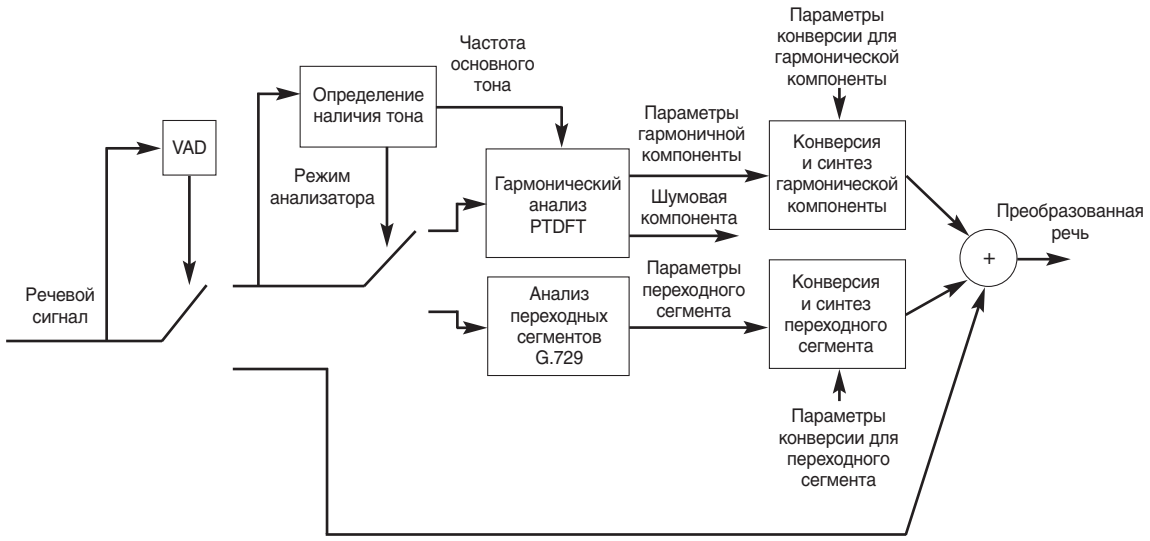

Рис. 8. Структурная схема процесса конверсии голоса

сегмент содержит тишину и/или фоновый шум, он передаётся напрямую на выход. Детектор тона определяет, будет ли речевой сегмент передан для обработки в модуль гармонического анализа либо в модуль анализа переходных сегментов. Параметры, выделенные с помощью одного из этих модулей анализа, подвергаются преобразованию, и далее осуществляется синтез фрагмента речи целевого диктора.

4. Экспериментальные результаты

Для сравнения использовалась система конверсии, полностью основанная на модели ACELP. Тесты показали, что голос, производимый предлагаемой системой, достаточно естественен и по разборчивости выше, чем в системе, основанной на подходе ACELP. Для примера была выбрана фраза на польском языке: «Lubić czardaszowy pląs» (***рис. 9***). 15 фраз из [19] были использованы для обучения.

Рис. 10 содержит спектрограммы примеров конверсии мужского голоса в женский. Очевидно, что результат конверсии предлагаемой системой лучше соответствует гармонической структуре речи целевого диктора и содержит меньше шума. Качество работы предлагаемой системы конверсии голоса оценивалось с помощью неформальных тестов прослушивания, которые показали, что узнаваемость диктора соответствует приблизительно 70%, реконструированная речь характеризуется достаточно высокой разборчивостью, хотя иногда характерны такие артефакты, как приглушённость и бормотание.

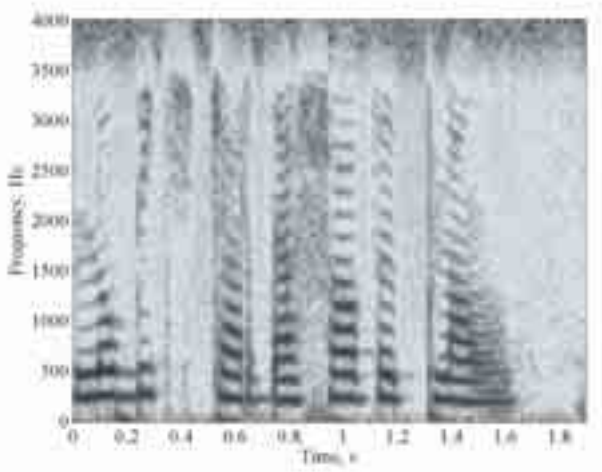

Рис. 9. Спектрограмма фразы целевого диктора

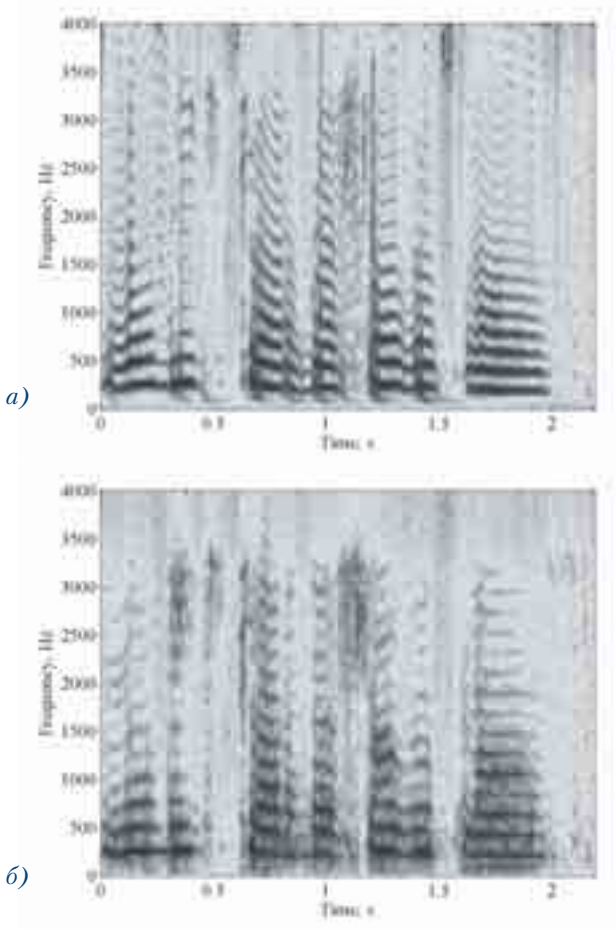

Рис. 10. Примеры конверсии голоса: а) системой, основанной на модели сепарации речевого сигнала; б) системой на базе ACELP-подхода

5. Заключение

В статье представлена система конверсии голоса, основанная на модели сепарации речевого сигнала на «гармоники+шум» и переходные фреймы с раздельной конверсией для каждой компоненты модели. Неформальные тесты прослушивания показали, что конвертированный речевой сигнал содержит небольшое количество артефактов, которые не мешают разборчивости и узнаваемости диктора.

Преимущество системы конверсии голоса в данном случае складывается из достоинств анализа-синтеза гармонической модели с достоинствами анализа и конверсии переходных сегментов во временной области. Благодаря данному подходу, значительно снижается доля исходного диктора в конвертированной речи, по сравнению, например, с системой конверсии на базе ACELP подхода.

Литература

1. *Moulines E. and Sagisaka Y., Eds.* «Voice conversion: state of the art and perspectives». Speech Communication, vol. 16, Feb. 1995.

2. *Pavlovets A., Kien T., Zubricki P. and Petrovsky A.* «Speech analysis — synthesis based on the PTDFT for voice conversion», in Proc. of the 2007 Int. Workshop on Spectral Methods and Multirate Sig. Proc., SMMSP, Moscow, Russia, Sep. 2007, pp. 203–210.

3. *Stylianou Y., Laroche J. and Moulines E.* «High-quality speech modification based on a harmonic + noise model», in Proc. of the European Conf. on Speech Communication and Technology EUROSPEECH, Madrid, Spain, Sep. 1995, pp. 451–454.

4. *Petrowsky A., Zubricki P. and Sawicki A.* «Tonal and noise components separation based on a pitch synchronous DFT analyzer as a speech coding method,» in Proc. European Conf. Circuit Theory and Design, Cracow, Poland, Sep. 2003, vol. 3, pp.169–172.

5. *Zubricki P., Pavlovec A. and Petrovsky A.* «Analysis-by-synthesis parameters estimation in the harmonic coding framework by pitch tracking modified DFT» in «New trends in audio and video», Dobrucki A., Petrovsky A. and Skarbek W. Eds. Byalystok 2006, pp. 233–246.

6. *Tremain T.* «The government standard linear predictive coding algorithm: LPC-10», Speech Technology Magazine, vol. 1, №. 2, 1982, pp. 40–49.

7. *Griffin D. and Lim J.* «Multiband excitation vocoder», IEEE Trans. Acoust., Speech and Sig. Proc., vol. 36, №8, pp. 1223 — 1235, Aug. 1988.

8. *Yegnanarayana B., d'Alessandro C. and Darsinos V.* «An iterative algorithm for decomposition of speech signals into periodic and aperiodic components», IEEE Trans. on Speech and Audio Proc., vol.6, № 1, pp. 1–11, Jan. 1998.

9. *Jackson P.J.B. and Shadle C.H.* «Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech», IEEE Trans. on Speech and Audio Proc., vol.9, №7, pp.713–726, Oct.2001.

10. *Abe M., Nakamura S., Shikano K. and Kuwabara H.* «Voice conversion through vector quantization», in Proc. of the Int. Conf. on Acoust., Speech and Sig. Proc. ICASSP, New York, USA, Apr.1988, vol.1, pp.655–658.

11. *Shlomot E., Cuperman V. and Gersho A.* «Hybrid coding: combined harmonic and waveform coding of speech at 4 kb/s», IEEE Trans. on Speech and Audio Proc., vol.9, № 6, pp. 632–646, Sep. 2001.

12. *Levine S. and Smith J.O.* «A sines+transients+noise audio representation for data compression and time/pitch scale modifications» in Proc. 105th Conv. Audio Eng. Soc., preprint #4781, Sep.1998.

13. *Sercov V. and Petrovsky A.* «An improved speech model with allowance for time-varying pitch harmonic amplitudes and frequencies in low bit-rate MBE coders», in Proc. of the European Conf. on Speech Communication and Technology EUROSPEECH, Budapest, Hungary, Sep.1999, pp.1479 — 1482.
14. *Talkin D.* «Robust algorithm for pitch tracking» in «Speech Coding and Synthesis», Kleijn W.B. and Palival K.K. Eds. Elsevier, Amsterdam, Netherlands, 1995.
15. ITU-T Rec. G.729, «Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear — prediction (CS-ACELP)», Mar.1996.
16. *Pavlovets A. and Petrovsky A.* «Voice conversion as a part of the voice analysis/synthesis system based on the periodic-aperiodic decomposition of speech», in Proc. of the 9th Int. Conf. on Pattern Recognition and Information Proc., PRIP, Minsk, Belarus, May2007, vol.2, pp. 71–76.
17. *Stylianou Y., Cappe O., Moulines E.* «Continuous probabilistic transform for voice conversion», IEEE Trans. on Speech and Audio Processing, vol. 6, № 2, pp. 131–142, March 1998.
18. ITU-T Rec. G.729, annex B, «A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70», Nov.1996.
19. *Grocholevski S.* «First database for spoken Polish», in Proc. Int. Conf. On Language Resources and Evaluation, Grenada, 1998, pp. 1059–1062.
20. *Huang X., Acero A., Hon H-W.* «Spoken Language Processing: a guide to theory, algorithms, and system development», Prentice Hall, NJ, 2001. — 980 p.

Павловец Александр Николаевич —

аспирант-заочник в Учреждении образования «Белорусский государственный университет информатики и радиоэлектроники». Закончил Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники» по специальности «Проектирование и технология электронных вычислительных средств». Работает на Заводе вычислительной техники им. С. Орджоникидзе. Область интересов — цифровая обработка речевых сигналов, кодирование речевых сигналов, проектирование проблемно-ориентированных средств вычислительной техники реального времени для мультимедиа-систем.

Лившиц Михаил Зеннадьевич —

кандидат технических наук, доцент кафедры Электронных вычислительных средств Учреждения образования «Белорусский государственный университет информатики и радиоэлектроники». Закончил Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники» по специальности «Электронные вычислительные средства». Область интересов — разработка реконфигурируемых аппаратных платформ для мультимедиа-систем, цифровая обработка сигналов, кодирование широкополосных речевых и аудиосигналов, повышение качества речевых сигналов, конверсия голоса.

Лихачёв Денис Сергеевич —

кандидат технических наук, доцент кафедры Электронных вычислительных средств Учреждения образования «Белорусский государственный университет информатики и радиоэлектроники». Закончил Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники» по специальности «Проектирование и технология электронных вычислительных средств». Область интересов — цифровая обработка речевых сигналов, системы компрессии речи, антропоморфическая обработка речи, конверсия голоса.

Использование закономерностей психоакустики в процедуре квантования параметров гармонической модели речевого сигнала

А.Н. Павловец,
аспирант
А.А. Петровский,
доктор технических наук, профессор

В данной работе рассматривается метод векторного квантования с переменной размерностью векторов для параметров гармонической модели речевого сигнала. Особенностью метода является использование закономерностей психоакустики в процедуре квантования амплитуд, что позволяет повысить качество реконструированного речевого сигнала и снизить вычислительную сложность алгоритмов квантования.

Abctract
The method of variable dimension vector quantization of harmonic model parameters is considered in this paper. The essence of the method is the incorporation of psychoacoustic principles into quantization procedure.

Введение

Большинство сигналов в природе, включая речь и музыку, могут быть описаны при помощи гармонической модели, которая определяется следующим набором параметров: фундаментальной частотой, амплитудой и фазой каждой частотной компоненты. Гармонический сигнал генерируется серией синусоид или гармонических компонент, частоты которых являются целочисленным кратным некоторой фундаментальной частоты. Данная модель является весьма эффективным решением для большого количества приложений кодирования сигнала, так как позволяет представить сигнал с помощью достаточно компактного набора параметров. Первые попытки представления речевого сигнала с помощью гармонической модели датируются началом 80-х годов [1].

Одним из фундаментальных вопросов в приложениях кодирования на базе гармонических моделей является квантование амплитуд гармоник, так как качество реконструированной речи в параметрических вокодерах в большой степени зависит от качества квантования параметров гармонической компоненты, несущей основную информацию о кодируемом речевом сигнале.

В настоящее время известно достаточно большое количество подходов к кодированию последовательности амплитуд гармоник. Скалярное квантование, например, квантует каждый элемент индивидуально; тем не менее, векторное квантование [2] является более предпочтительным подходом для современных алгоритмов низкоскоростных кодеров речи, что обусловлено лучшим соотношением качество/скорость передачи. Традиционные векторные квантователи строятся с учётом фиксированной длины векторов. В последних работах удалось добиться достаточно высокого качества квантования гармонических амплитуд благодаря применению схемы расщеплённого векторного квантования линейных спектральных пар, при этом прозрачное кодирование достигалось при скорости 23 бит/вектор [3]. Тем не менее, построение векторного квантователя с переменной длиной кодируемого вектора амплитуд гармоник выглядит более естественным решением ввиду того, что при этом не требуется дополнительных преобразований над входным вектором.

Использование особенностей слуховой системы человека при низкоскоростном кодировании речи было рассмотрено в работе [4], где закономерности психоакустики учитывались при построении огибающих спектров и при расчёте весовых коэффициентов для взвешивания ошибки квантования параметров в контексте MBE-вокодера. Таким образом, целью данной работы является разработка метода квантования векторов амплитуд гармоник с учётом особенностей восприятия речи человеком.

Квантование гармонических амплитуд

В контексте гармонической модели проблема квантования в большей степени связана с передачей вектора амплитуд гармоник. Если рассмотреть изменение спектра речевого сигнала во времени для разных дикторов (***рис. 1 а и 1 б***), можно сделать вывод, что векторы амплитуд гармоник, даже определяющие голос одного и того же диктора, имеют различную размерность в разные моменты времени.

К сожалению, математический аппарат векторного квантования был разработан для квантования векторов фиксированной длины и практически не используется с векторами переменной длины, такими, как векторы амплитуд гармоник. К решению данной проблемы возможны различные подходы. Одним из вариантов является использование собственной кодовой книги для каждой размерности [5]. Естественно, такой подход является малоприемлемым для использования в системах реального времени из-за серьёзных требований к объёму памяти. В наиболее широко применяемых решениях осуществляются различные преобразования над векторами переменной размерности, с тем чтобы привести их размерность к некоторому заданному фиксированному значению (с сохранением формы речевого спектра) с последующим применением техник векторного квантования. Примерами таких решений могут служить [3, 6–8]. Очевидным недостатком здесь является необходимость дополнительных преобразований и, следовательно, возможность внесения дополнительных искажений.

Одно из возможных решений — квантование фиксированного количества гармонических амплитуд: например, в кодере на базе линейного предсказания со смешанным возбуждением

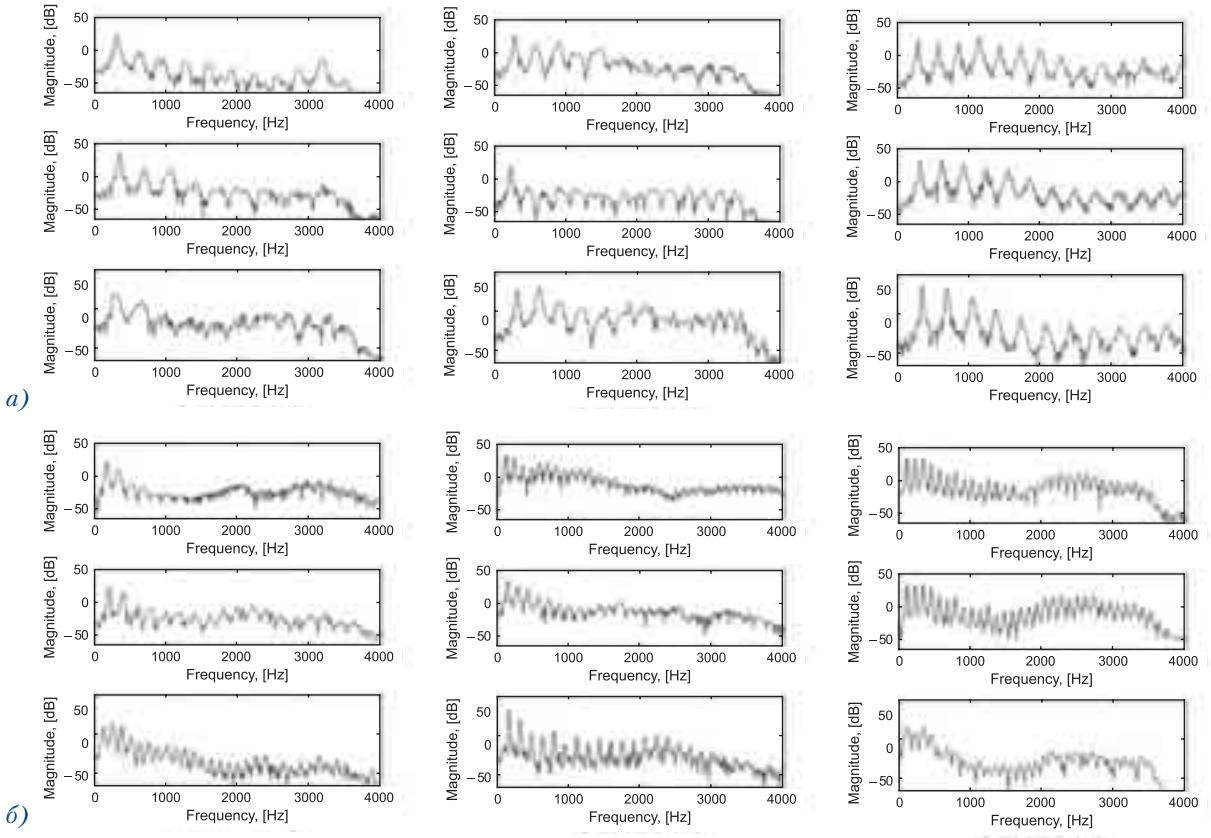

Рис. 1. Изменение спектра речи во времени: а) женский голос; б) мужской голос

(MELP — Mixed Excitation Linear Prediction) [9] векторное квантование используется для квантования первых 10-ти гармонических амплитуд, а амплитуды остальных гармоник считаются равными амплитуде последней (10-й) гармоники. Легко заметить, что 10 гармоник покрывают весь или почти весь речевой спектр для женских голосов с высокой частотой основного тона, в то время как для мужских голосов они могут покрыть только одну четвертую всего частотного диапазона (***рис. 1а и 1б***), что означает существенную потерю качества для мужских голосов по сравнению с женскими.

Наконец, в [10] была разработана схема векторного квантования с переменной размерностью векторов (от англ. Variable Dimension Vector Quantization — VDVQ). Тем не менее, поскольку в этом подходе не учитываются закономерности психоакустики, его трудно считать оптимальным. Далее будет рассмотрен математический аппарат VDVQ и его модификация с точки зрения восприятия речи человеком.

Векторное квантование с переменной размерностью векторов

В схеме VDVQ, предложенной в [10], кодовая книга квантователя содержит N_c кодовых векторов:

$$y_i,\ i = 0,...,N_c - 1$$

при

$$y_i^T = \left[y_{i,0} \quad y_{i,1} \quad \cdots \quad y_{i,N_v-1} \right],$$

где N_v — размерность кодового вектора.

Пусть поиск вектора гармонических амплитуд x с размерностью $N(\omega_0)$ и нормализованной частотой основного тона ω_0 осуществляется путём полного перебора в кодовой книге, тогда требуется рассчитать следующие расстояния:

$$d_i(x, \hat{y}_i),\ i = 0,...,N_c - 1,$$

где

$$\hat{y}_i^T = \left[\hat{y}_{i,1} \quad \hat{y}_{i,2} \quad \cdots \quad \hat{y}_{i,N(\omega_0)} \right],$$
$$\hat{y}_{i,j} = y_{i,k_j},\ j = 1,...,N(\omega_0)$$

при

$$k_j = \left[\frac{N_V \omega_j}{\pi} \right],\ \omega_j = j\omega_0,\ j = 1,...,N(\omega_0),$$

где [] означает округление к ближайшему целому.

Схема работает следующим образом: для каждого кодового вектора y_i путём расчёта набора индексов k_j извлекается вектор $\hat{y}_i$, имеющий ту же размерность, что и x. Эти индексы рассчитываются в соответствии с периодом основного тона и указывают на элементы y_i, ближайшие к позиции j-й гармоники в кодовой книге. После расчёта всех расстояний d_i для квантования x выбирается индекс кодового вектора с наименьшим расстоянием. В качестве расстояния (меры искажения) используется спектральное отклонение:

$$SD = \sqrt{\frac{1}{N(\omega_0)} \sum_{j=1}^{N(\omega_0)} \left(x_j - \hat{y}_j \right)^2}.$$

Модифицированная конфигурация схемы VDVQ, называемая IVDVQ, предложена в [11]. Данное изменение заключается в интерполяции элементов кодовых векторов y_i для получения действительных кодовых векторов $\hat{y}_i$. Индексы k_j в IVDVQ рассчитываются без операции округления:

$$k_j = \frac{N_V \omega_j}{\pi},\ \omega_j = j\omega_0,\ j = 1,...,N(\omega_0). \qquad (1)$$

Элемент $\hat{y}_{i,j}$ получается путём линейной интерполяции между двумя элементами вектора y_i, определяемыми индексами $\lfloor k_j \rfloor$ и $\lceil k_j \rceil$:

$$\hat{y}_{i,j} = y_{i,\lfloor k_j \rfloor} + \{k_j\}(y_{i,\lceil k_j \rceil} - y_{i,\lfloor k_j \rfloor}),$$

где $\{k_j\}$ обозначает дробную часть выражения (1). Обучение кодовых книг по методам VDVQ и IVDVQ представляет собой вариацию на тему алгоритма «k-средних» [12] и подробно описано в [11]. Результат применения метода к квантованию гармонических амплитуд отражён на ***рис. 2***; использовалась 10-разрядная кодовая книга.

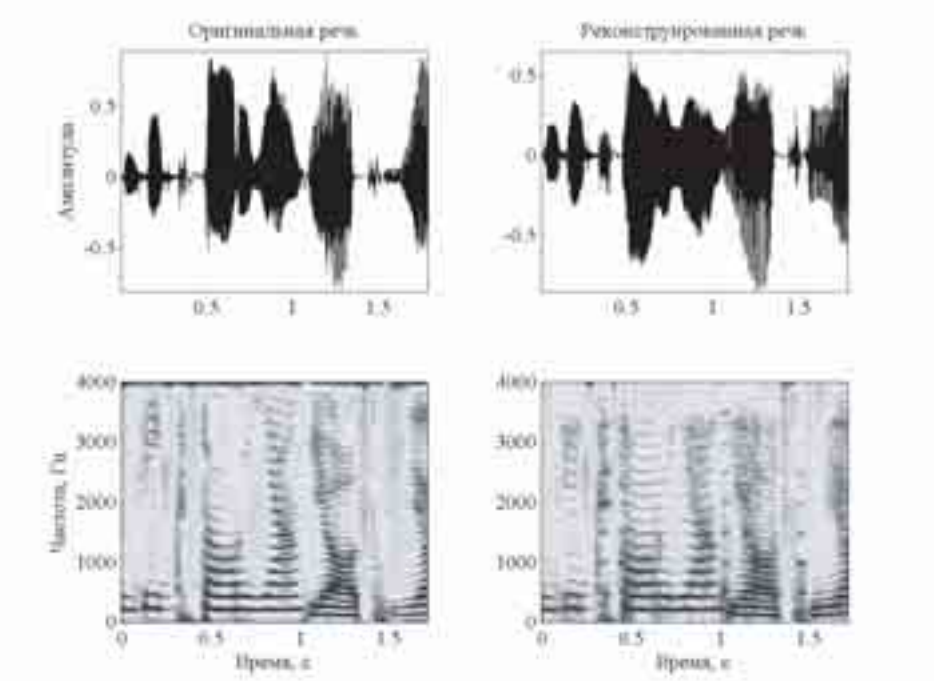

Рис. 2. Пример восстановления речи, кодированной с использованием метода VDVQ

VDVQ с психоакустически обоснованным ограничением длины вектора

Кодовые книги для VDVQ-метода обычно имеют большую длину кодовых слов (от 41 до 109 — в экспериментах [11]), что приводит к высоким требованиям к объёму памяти для их хранения. В то же время можно видеть, что последние амплитуды гармоник спектра имеют незначительную величину, особенно в случае мужской речи (рис. 1б). Следовательно, имеет смысл ограничить размерность квантуемого вектора таким образом, чтобы не учитывать достаточно малые амплитуды.

Схожая проблема существует в рамках модели речевого сигнала «гармоники плюс шум» [13, 14], где необходимо найти максимальную частоту вокализованности (ограничить спектр гармонической компоненты). Алгоритм, предложенный в [13], осуществляет проверку спектра на гармоничность в окрестности амплитуд гармоник, и в случае, если спектр в области двух смежных проверяемых гармоник оказался негармоническим, проверка прекращается. В качестве максимальной частоты вокализованности принимается последняя гармоника частоты основного тона, предшествующая негармонической области спектра. Всё же данный алгоритм является в большой степени эвристическим и использует при оценке некоторые заранее определённые опытным путём пороговые значения.

Модель анализа речевого сигнала, рассмотренная в [15], предполагает разделение речи на гармоническую и шумовую компоненту по всему спектру. Используя закономерности психоакустики, можно определить, в какой степени шумовая компонента влияет на восприятие человеком гармонической компоненты, т.е. определить гармоники, не влияющие на восприятие речи в целом.

Для решения данной проблемы использовалась психоакустическая модель Джонстона [16]. Данная модель позволяет рассчитать порог маскирования «шум маскирует тон» в частотной области с использованием следующей последовательности действий:

1) Сегмент шумовой компоненты взвешивается временным окном и подвергается ДПФ;

2) Спектр мощности шумовой компоненты суммируется в критических частотных полосах, измеряемых в барках [17]:

$$B_i = \sum_{n=bl_i}^{bh_i} P(n),$$

где $P(n)$ — n-й частотный компонент спектра мощности; bl_i и bh_i — номера начального и конечного спектрального отсчёта, попадающих в i-ю критическую частотную полосу.

Шкала барков получается с помощью следующего преобразования [17]:

$$z(f) = 1 + 13arctg(0{,}76f) + 3.5arctg((f/7{,}5)^2),$$

где f — частота в Гц. Для ДПФ размерности 256 и частоты дискретизации F_s=8000 Гц параметры критических частотных полос приведены в таблице 1.

3) Рассчитывается функция распространения возбуждения по базилярной мембране для оценки эффектов маскирования в нескольких критических частотных полосах [18]:

$$S_{i,j} = 10^{(15{,}81+7{,}5(k+0{,}474)-17{,}5\sqrt{1+(k+0{,}474)^2})/10},$$

где $k=i-j$; i — номер барка маскируемого сигнала; j — номер барка маскирующего сигнала.

4) Вычисляется распространение спектральной энергии барка в каждой критической частотной полосе как свёртка спектра мощности B_i с функцией распространения возбуждения $S_{i,j}$:

$$\begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ \dots \\ C_{18} \end{bmatrix} = \begin{bmatrix} S_{1,1} & S_{1,2} & S_{1,3} & \dots & S_{1,18} \\ S_{2,1} & S_{2,2} & S_{2,3} & \dots & S_{2,18} \\ S_{3,1} & S_{3,2} & S_{3,3} & \dots & S_{3,18} \\ \dots & \dots & \dots & \dots & \dots \\ S_{18,1} & S_{18,2} & S_{18,3} & \dots & S_{18,18} \end{bmatrix} \times \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ \dots \\ B_{18} \end{bmatrix}$$

5) Рассчитываются коэффициенты тональности для каждой критической частотной полосы:

$$\alpha_i = \min\left(\frac{SFM_{dB}(i)}{SFM_{dB\max}}, 1\right),$$

где $SFMdB(i)$ — мера спектральной пологости в i-ой критической частотной полосе:

$$SFM_{dB} = 10[\log_{10}(GM) - \log_{10}(AM)],$$

где AM и GM — среднее арифметическое и среднее геометрическое значение спектра мощности в i-ой критической частотной полосе; SFM_{dBmax} — максимальное значение меры спектральной пологости, равное 60 дБ.

Таблица 1

Параметры критических частотных полос приведены для ДПФ размерности 256 и частоты дискретизации Fs=8000 Гц

Номер критической полосы	Номера элементов ДПФ	Количество элементов ДПФ	Частоты, Гц
1	1…3	3	0…94
2	4…6	3	94…187
3	7…10	4	187…312
4	11…13	3	312…406
5	14…16	3	406…500
6	17…20	4	500…625
7	21…25	5	625…781
8	26…29	4	781…906
9	30…35	6	906…1094
10	36…41	6	1094…1281
11	42…47	6	1281…1469
12	48…55	8	1469…1719
13	56…64	9	1719…2000
14	65…74	10	2000…2312
15	75…86	12	2312…2687
16	87…100	14	2687…3125
17	101…118	18	3125…3687
18	119…128	9	3687…4000

6) Определяются смещения порогов маскирования:

$$O_i = 5{,}5(1-\alpha_i),\ i=1\ldots18.$$

7) Производится расчёт порогов маскирования в критических полосах и их ренормализация:

$$T_i = 10^{\log_{10}(C_i)-O_i/10},\ i=1\ldots18.$$

Для ренормализации требуется определить ошибку распространения спектральной энергии барка. Для этого предполагается, что на слуховую систему воздействует гипотетический раздражитель, спектральная энергия которого в критической частотной полосе равна единице:

$$\begin{bmatrix} C_{E1} \\ C_{E2} \\ C_{E3} \\ \dots \\ C_{E18} \end{bmatrix} = \begin{bmatrix} S_{1,1} & S_{1,2} & S_{1,3} & \dots & S_{1,18} \\ S_{2,1} & S_{2,2} & S_{2,3} & \dots & S_{2,18} \\ S_{3,1} & S_{3,2} & S_{3,3} & \dots & S_{3,18} \\ \dots & \dots & \dots & \dots & \dots \\ S_{18,1} & S_{18,2} & S_{18,3} & \dots & S_{18,18} \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}$$

Ренормализованные пороги маскирования определяются так:

$$T_i^{'} = T_i - 10\log_{10}(C_{Ei}),\ i=1\dots18.$$

8) Окончательные значения порогов маскирования определяются так:

$$T_i^f = \max\left(T_i^{'}, ATH(f)\right),\ i=1\dots18,$$

где $ATH(f)$ — функция, аппроксимирующая значение абсолютного порога слышимости [17] и рассчитываемая с помощью следующего выражения для частот, равных значениям гармоник частоты основного тона:

$$ATH(f) = 3.64 f^{-0.8} - 6.5e^{-0.6(f-3.3)^2} + 10^{-3} f^4,$$

где f — частота в кГц.

Максимальной частотой вокализованности считается последняя гармоника частоты основного тона, превышающая порог маскирования.

На ***рис. 3*** показан результат расчёта порога маскирования и определения максимальной частоты вокализованности для вектора амплитуд гармоник. Очевидно, что вычислительная сложность поиска в кодовой книге в данном случае будет снижена более чем в 2 раза.

Таким образом удаётся ограничить размерность вектора амплитуд гармоник на основании закономерностей психоакустики и тем самым снизить вычислительную сложность процесса его квантования. Результат применения метода отражён на ***рис. 4***; использовалась 10-разрядная кодовая книга.

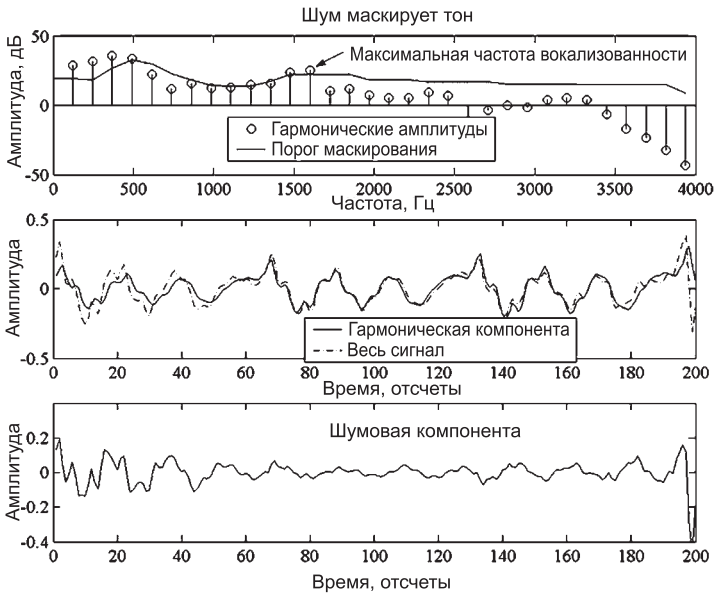

Рис. 3. Маскирование амплитуд гармоник шумовой компонентой

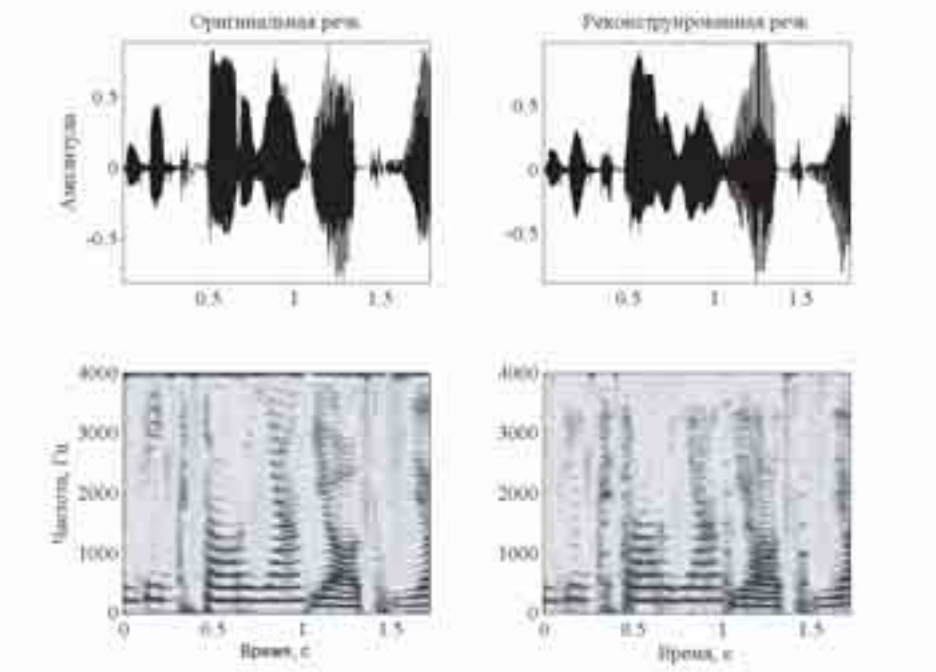

Рис. 4. Результат применения метода VDVQ с психоакустически мотивированным ограничением длины вектора

Сравнение качества методов квантования векторов амплитуд гармоник переменной длины

Поскольку предлагаемые методы квантования основаны на использовании особенностей слуха человека, классические параметры, по которым можно их сравнить (отношение «сигнал/шум», спектральное отклонение и т.д.), не смогут обеспечить корректную оценку качества. В то же время оценка качества по шкале MOS (Mean Opinion Score) требует наличия специально оборудованного помещения и определённого количества подготовленных слушателей. Таким образом, целесообразным будет произвести оценку качества реконструированной речи с помощью такого параметра, при расчёте которого использовалась бы модель слуха человека. Таким параметром является модифицированная величина искажений спектра барков (MBSD — Modified Bark Spectral Distortion) [19], искажения в данном случае определяются как средняя разность субъективных оценок громкости.

Для сравнительной оценки качества квантователи, построенные на базе предложенных методов, были использованы в составе вокодера, основанного на декомпозиции речевого сигнала на периодическую и апериодическую компоненты [15]. В ходе эксперимента квантованию подвергались только амплитуды гармоник (использовались десятиразрядные кодовые книги), прочие параметры не квантовались. Результаты тестирования качества реконструированной речи для различных вариантов квантования приведены в ***таблице 2***.

Таблица 2

Качество реконструированной речи при использовании различных подходов для квантования векторов амплитуд гармоник

	VDVQ	VDVQ+психоакустически обоснованное ограничение длины вектора
MBSD	5,5973	5,3348

Таким образом, психоакустически модифицированный вариант квантования векторов амплитуд гармоник показал по результатам измерений лучшее качество с точки зрения восстановления речи.

Заключение

В данной статье были рассмотрены методы квантования векторов амплитуд гармоник речевого сигнала. Метод квантования векторов переменной длины является весьма удобным для использования с такими параметрами гармонической модели речи, как амплитуды, поскольку отпадает надобность в дополнительных преобразованиях. Предложенные методы, в основе которых лежат преобразования, использующие закономерности психоакустики, позволяют повысить качество реконструированной речи и снизить вычислительную сложность алгоритмов квантования.

Литература

1. *Almeida L., Tribolet J.* «Nonstationary spectral modeling of voiced speech», IEEE Trans. Acoustics, Speech, Signal Processing., vol.ASSP-31, №3, pp. 664–678, June 1983.
2. *Gersho A., Gray R.M.* Vector Quantization and Signal Compression. Kluwer Academic, Norwell, USA, 1992.
3. *Павловец А., Петровский А.* «Квантование огибающей спектра в вокодере, основанное на декомпозиции речевого сигнала на периодическую и апериодическую составляющие», Цифровая обработка сигналов, №3, Москва, 2005 г., с. 13–21.
4. *Серков В., Петровский А.* «Использование закономерностей психоакустики при низкоскоростном кодировании речи», Доклады 3-й междунар. конф. «Цифровая обработка сигналов и её применения», DSPA'2000, Москва, 2000, с. 241–244.
5. *Adoul J.-P., Delprat M.* «Design algorithm for variable-length vector quantizers» in Proc. Allerton Conf. on Circuits, Syst., Comput, 1986, pp. 1004–1011.
6. *McAulay R.J., Quatiery T.F.* «Sinusoidal Coding» in «Speech Coding and Synthesis» (W.Klein and K. Palival, eds.), Amsterdam: Elsevier Science Publishers, 1995, pp. 121–176.
7. *Nishiguchi M., Inoue A., Maeda Y., and Matsumoto J.* «Parametric speech coding-HVXC at 2.0–4.0 kbps» in Proc. IEEE Speech Coding Workshop, pp. 84–86, Porvoo, Finland, June 1999.
8. *Li C., Lupini P., Shlomot E., and Cuperman V.* «Coding of variable dimension speech spectral vectors using weighted nonsquare transform vector quantization», IEEE Trans. Speech, and Audio Processing, vol. 9, no. 6, pp. 622–631, 2001.
9. *Supplee L., Cohn R., Collura J., McCree A.* «MELP: the new federal standard at 2400 bps», in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'97, vol.2, pp. 1591–1594, Munich, Germany, April 1997.
10. *Das A., Rao A., Gersho A.* «Variable-dimension vector quantization», IEEE Signal Processing Letters, vol.3, no.7, pp. 200–202, 1996.

11. *Chu W.C.* «A novel approach to variable dimension vector quantization of harmonic magnitudes», in Proc. 3rd IEEE International Symposium on Image and Signal Processing and Analysis, vol.1, pp.537–542, Rome, Italy, September 2003.
12. *MacQueen, J.B.* «Some Methods for Classification and Analysis of Multivariate Observations», in Proc. Fifth Berkeley Symp. Math. Statistics and Probability, vol.1, pp. 281–296, 1967.
13. *Stylianou Y.* «Applying the harmonic plus noise model in concatenative speech synthesis», IEEE Transactions on Speech and Audio Processing, vol.9, №1, pp. 21–29, Jan. 2001.
14. *Bao C., Lukasiak J., Ritz C.* «A novel voicing cut-off determination for low bit-rate harmonic speech coding», in INTERSPEECH-2005, pp. 2709–2712.
15. *Petrovsky A., Zubricki P., Sawicki A.* «Tonal and noise components separation based on a pitch synchronous DFT analyzer as a speech coding method», in Proc. European Conf. Circuit Theory and Design, Cracow, Poland, Sep. 2003, vol.3, pp.169–172.
16. *Johnston J.* «Estimation of perceptual entropy using noise masking criteria», in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'88, vol.5, pp. 2524–2527, New York, NY, USA, April 1988.
17. *Zwicker E., Fastl H.* «Psychoacoustics: facts and models», Springer-Verlag, Berlin, 1990.
18. *Schroeder M.R., Atal B.S., Hall J.L.* «Optimizing digital speech coders by exploiting masking properties of the human ear», Journal of the Acoustical Society of America, vol.66, pp.1647–1652, 1979.
19. *Петровский А.А.* Объективная оценка качества восстановленного аудио сигнала перцептуальным ПДВП-кодером на базе периферийной модели уха человека // Сборник докладов 5 Международной научной конференции «Цифровая обработка сигналов и её применение» (DSPA'2003), т. 2, Москва, Россия, 2002. С. 123–126.

Павловец Александр Николаевич —

аспирант-заочник в Учреждении образования «Белорусский государственный университет информатики и радиоэлектроники». Закончил Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники» по специальности «Проектирование и технология электронных вычислительных средств». Работает на Заводе вычислительной техники им. С. Орджоникидзе. Область интересов — цифровая обработка речевых сигналов, кодирование речевых сигналов, проектирование проблемно-ориентированных средств вычислительной техники реального времени для мультимедиа-систем.

Петровский Александр Александрович —

доктор технических наук, профессор. Работает в Учреждении образования «Белорусский государственный университет информатики и радиоэлектроники», кафедра «Электронные вычислительные средства». Закончил Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники» по специальности «Электронные вычислительные машины». Главные научные интересы лежат в области цифровой обработки сигналов речи и звука для целей компрессии, распознавания, редактирования шума, а также в области проектирования проблемно-ориентированных средств вычислительной техники реального времени для систем мультимедиа.
Член НТО РЭС им. А.С.Попова, IEEE, EURASIP, AES.

Построение психоакустической модели в области вейвлет-коэффициентов для перцептуальной обработки звуковых и речевых сигналов

А.А. Петровский,
кандидат технических наук, доцент

В статье предлагается методология построения дерева пакета дискретного вейвлет-преобразования для перцептуальной обработки сигналов речи и звука при минимальном множестве узлов, определённой глубине декомпозиции структуры дерева, заданной частоте дискретизации сигнала, когда ошибка аппроксимации критических частотных полос минимальна в области барков. Также приведена процедура расчёта маскирующих порогов в пространстве вейвлет-коэффициентов.

Abstract

The methodology of a wavelet packet (WP) tree construction for time-frequency processing of speech and audio signals is proposed to obtain a WP tree for the minimal number of nodes, certain WP tree decomposition depth, the required sampling frequency, the minimal approximation error of the critical frequency bands. The procedure of masking thresholds calculation in the wavelet coefficients domain is also described in the given article.

Введение

Пакет дискретного вейвлет-преобразования (ПДВП), или, другими словами, быстрое вейвлет-преобразование [1], является важным инструментом частотно-временной декомпозиции звуковых и речевых сигналов для различных приложений мультимедиа-систем [1, 2]. ПДВП часто используется в субполосной перцептуальной обработке сигналов звука и речи [2-4] в критических частотных полосах согласно психоакустической модели Zwicker [5]:

— расстояние между центральными частотами соседних критических частотных полос определяется формулой

$$z = F(f) = 13 \cdot \arctan(0.00076 \cdot f) + 3.6 \cdot \arctan\left(\left(\frac{f}{7500}\right)^2\right), [Барк] \quad (1)$$

где f — частота в герцах (единица измерения в данном масштабе — 1 барку);

— ширина критических частотных полос определяется формулой

$$CBW(f) = 25 + 75 \cdot \left(1 + 1.4 \cdot \left(\frac{f}{1000}\right)^2\right)^{0.69}. \ [Гц] \quad (2)$$

При этом также вычисляются характеристики психоакустической модели восприятия человеком акустической информации, такие как пороги маскирования, а именно:

— абсолютный порог слышимости АНТ (absolute threshold of hearing), частотная зависимость которого аппроксимируется выражением:

$$ATH(f) = 3.64 \cdot \left(\frac{f}{1000}\right)^{-0.8} - 6.5 \cdot e^{-0.6\left(\frac{f}{1000} - 33\right)^2} + 10^{-3} \cdot \left(\frac{f}{1000}\right)^4, \ [Гц] \quad (3)$$

где f — частота в герцах;

— частотное маскирование (simultaneous masking), проявляющееся при воздействии маскера в течение некоторого времени одновременно на разных частотах сигнала;

— маскировка во временной области (temporal masking): если громкий звук маскирует следующий за ним слабый звук, то явление называется «маскировкой вперёд» (post-masking) и может продолжаться от 5 мс до 300 мс в зависимости от силы и длительности маскера; при «маскировке назад» (pre-masking) громкий звук маскирует звук, воспроизводимый до него, длительность которого составляет примерно 20 мс.

Итак, пусть $\{\varphi_n(t) : z \in Z\}$ определяет множество структур деревьев ПДВП и пусть $\{E \subset \{(l,n) : 1 \le l \le L,\ 0 \le n \le 2^l\}\}$ представляет собой узлы дерева ПДВП, тогда отрезок $[0,1)$ разделяется на диадические интервалы $I_{l,n} = [n2^l,\ (n+1)\ 2^l]$, которые соответствуют специфическому множеству узлов E. В частности, $\{\varphi_{l,n,k}(t) : (l,n) \in E,\ k \in Z\}$, где $\varphi_{l,n,k}(t) \triangleq 2^{-\frac{l}{2}} \varphi_{l,n}(2^{-l}t - k)$ является базовой формой в пространстве сигнала $\overline{span}\{\varphi_{0,0}(t-k) : k \in Z\}$. Узел $(l,n) \in E$ дерева ПДВП ассоциируется с частотной полосой, у которой центральная частота и полоса пропускания приблизительно задаются следующими соотношениями:

$$f_{l,n} = 2^{-l}\left(GC^{-1}(n) + 0.5\right) \cdot \frac{f_s}{2}, \quad (4)$$

$$\Delta f_{l,n} = 2^{-l} \cdot \frac{f_s}{2}, \quad (5)$$

где $GC^{-1}(n)$ — обратный код перестановок Грея, f_s — частота дискретизации сигнала.

ПДВП реализуется на выбранной структуре дерева, поиск которой основывается на известном утверждении [6]: любая комбинация целых индексов $(l,n,k) \in Z$, для которых вейвлеты сконцентрированы на двоичных интервалах $[n2^{-l}, (n+1)\,2^{-l}]$ из диапазона $[0,\infty)$, соответствует ортогональным базисам $\psi_{l,n,k}(t)$, $\varphi_{l,n,k}(t)$ из пространства $L^2(k)$. Утверждение доказывает существование множества структур ПДВП, причём ПДВП ассоциируется с алгоритмом выбора лучшей структуры преобразования $\{E \subset \{(l,n): 1 \le l \le L,\ 0 \le n \le 2^l\}\}$ из множества структур путём изменения и минимизации определённой меры качества.

Цель данной статьи — показать методологию построения дерева ПДВП для минимального множества узлов (l,n), определённой глубины декомпозиции l, заданной частоты дискретизации сигнала f_s, когда ошибка аппроксимации критических частотных полос минимальна в области барков, а также получить процедуру расчёта маскирующих порогов в пространстве вейвлет-коэффициентов.

1. ПДВП, согласованный со шкалой критических частотных полос

Для того чтобы получить аппроксимацию шкалы критических частотных полос с помощью ПДВП, необходимо осуществить декомпозицию дерева ПДВП таким образом, чтобы расстояние между центрами субполос составляло 1 барк. Следует отметить, что ширина критических частотных полос $CBW(f)$ — монотонно увеличивающаяся функция частоты (2). Для формирования низкочастотных полос требуется интенсивная декомпозиция ПДВП в сравнении с характером изменения дерева ПДВП для аппроксимации высокочастотных полос.

Дано дерево ПДВП $(l,n) \in E_m$ и его вейвлет-коэффициенты $X_{l,n,k}$. Интегральная перцептуально взвешенная ошибка аппроксимации шкалы критических частотных полос деревом $(l,n) \in E_m$ ПДВП в области барков может быть определена следующим образом:

$$Q_E = \frac{1}{L} \sum_{\substack{\text{ДЛЯ} \\ \forall (l,n) \in E_m}} \left[\widehat{CBW}_{Zw}(z) - \widehat{CBW}_{E_m}(z_{(l,n)})\right]^2 \cdot \widehat{W}(z). \qquad (6)$$

Здесь ширина критических частотных полос $\widehat{CBW}_{Zw}(z)$ в Гц — функция центральных частот соседних критических частотных полос, заданных в барках, то есть:

$$\widehat{CBW}_{Zw}(z) = CBW(F^{-1}(z)), [\text{Гц}] \qquad (7)$$

$\widehat{CBW}_{Zw}(z)$ определяет шкалу критических частотных полос в модели Zwicker [5];

$\widehat{CBW}_{E_m}(z_{(l,n)})$ — аппроксимация критических частотных полос деревом ПДВП $(l,n) \in E_m$;

$z_{(l,n)}$ — центр полосы (l,n) в барках дерева ПДВП E_m вычисляется для центральной частоты $f_{(l,n)}$, заданной в Гц, как $z_{(l,n)\,\in E_m} = F(f_{(l,n)})$, где F — преобразование (1). Перцептуальная взвешивающая функция $\widehat{W}(z)$, учитывающая определённые частотные свойства наружного и среднего уха, задаёт меньшее распределение ошибки аппроксимации шкалы критических частотных полос в области средних частот по сравнению с низкочастотным и высокочастотным диапазонами и определяется в шкале дБ как функция частоты [7]:

$$W_{\text{дБ}}(f) = -0.6 \times 3.64(10^{-3} f)^{-0.8} + 6.5 \times \exp(-0.6 \times (10^{-3} f - 3.3)^2) - 10^{-3}(10^{-3} f)^4, \qquad (8)$$

Также $\widehat{W}(z)$ может быть переопределена для барков как

$$\widehat{W}(z) = \widehat{W}(F(f)) = W(F^{-1}(z)) = W(f), \qquad (9)$$

где $W(f) = 10^{(W_{\text{дБ}}(f)/20)}$.

Минимизация ошибки Q_E (6) может позволить автоматизировать процесс построения оптимального дерева ПДВП $(l,n) \in E_{CB}$ для шкалы критических частотных полос.

На ***рис. 1*** показано дерево ПДВП (Critical Band Wavelet Packet Decomposition ($CB - WPD$)) [3, 8], полученное эмпирически, которое осуществляет разделение частотного интервала сигнала на полосы согласно критической шкале частот: $CB - WPD{:}(l,n) \in E_{CB}$, $l=\overline{0,8}$, где E_{CB} обозначает множество узлов дерева ПДВП, соответствующего $CB - WPD$. Дерево $CB - WPD$ делит частотный диапазон [0–22,05 кГц] на 25 неравномерных полос $CBW(f)$, то есть на 25 барков. Корневой узел $(l,n) = (0,0)$ данного дерева соответствует всему частотному диапазону сигнала. Каждый внутренний узел дерева $(l,n) \in E$, названный узлом предка, делится на два потомка: 1-й потомок и 2-й потомок, ассоциируемые соответственно с высокочастотной и низкочастотной фильтрацией, выходные сигналы (вейвлет-коэффициенты) которых децимируются в соотношении два к одному:

$$X_{l,n,k}(t) = \left\langle x(t), \varphi_{l,n,k}(t) \right\rangle, (l,n) \in E_{CB}, k \in Z. \qquad (10)$$

Для систем перцептуальной обработки широкополосных речевых сигналов предлагается следующее критическое дерево ПДВП $CB - WPD$ [8]:$(l,n) \in E_{CB}$, $l=\overline{0,7}$ (***см. рис. 2***), где ширина полосы анализируемого сигнала равна 16 кГц и обработка ведётся в 24 барках.

Разрешающая способность человеческого уха ограничивает длину анализируемого фрейма в пределах 5–10 мс для области верхних частот и 100 мс для нижних частот. Выбор фильтра прототипа преобразования, длины его вейвлет-

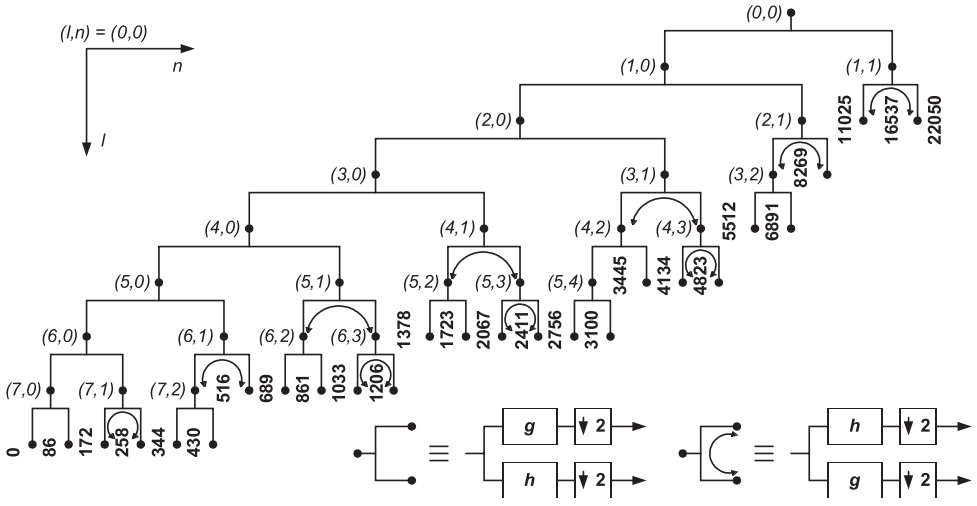

Рис. 1. Структура критического дерева ПДВП $CB - WPD{:}(l,n) \in E_{CB}$, $l=\overline{0,8}$

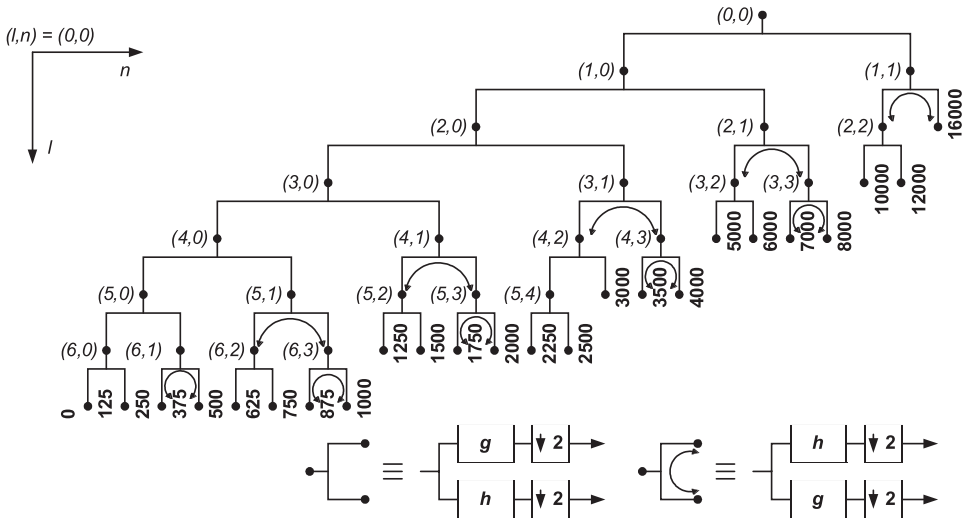

Рис. 2. Структура критического дерева ПДВП CB – WPD:(l,n) ∈ E_{CB}, l=0,7

функции в общем случае зависит от максимального размера окна обработки (временного разрешения) и нулевых моментов [1]. Использование вейвлет-функций семейства Добеши позволяет обеспечить хорошую частотную избирательность, которая увеличивается с числом масштабных уровней в дереве преобразования.

На ***рис. 3*** и ***рис. 4*** показаны аппроксимации центральной частоты и ширины каждой частотной полосы критической шкалы частот (соответственно) деревом ПДВП, структура которого приведена на рис. 1. Аналогичные результаты для дерева ПДВП (см. рис. 2) иллюстрируются на ***рис. 5*** и ***рис. 6*** соответственно. Здесь непрерывная линия соответствует критической шкале частотных полос согласно модели Zwicker, а кружки — декомпозиции дерева ПДВП. Сопоставление данного результата с другими структурами дерева *CB – WPD* [9, 10] даёт определённый выигрыш в величине ошибки Q_E (от 0 до 4 дБ), а неформальные тесты прослушивания показывают лучшее восприятие восстановленных сигналов в перцептуальных системах компрессии и редактирования шумов.

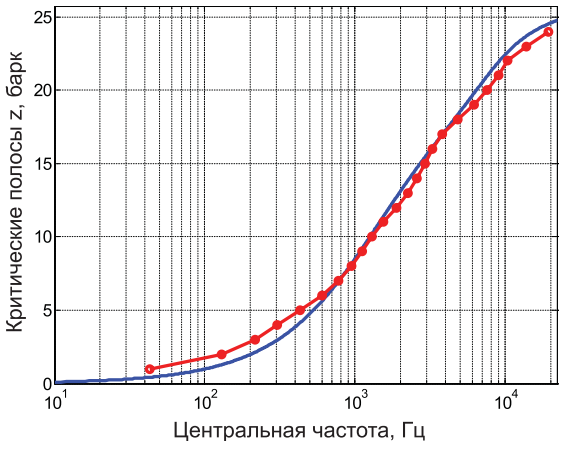

Рис. 3. Аппроксимация центральных частот CB – WPD:(l,n) ∈ E_{CB}, l=0,8

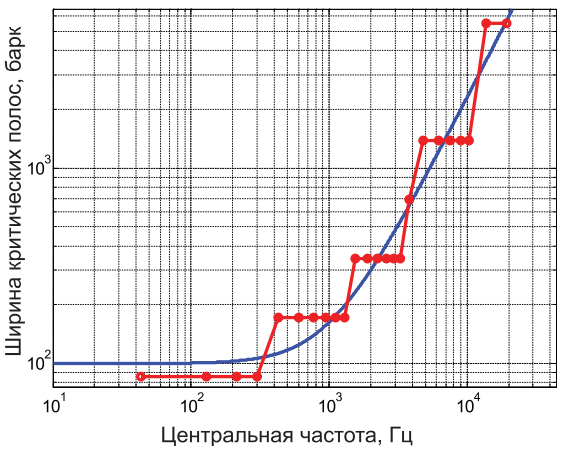

Рис. 4. Аппроксимация ширины критических частотных полос CB – WPD:(l,n) ∈ E_{CB}, l=0,8

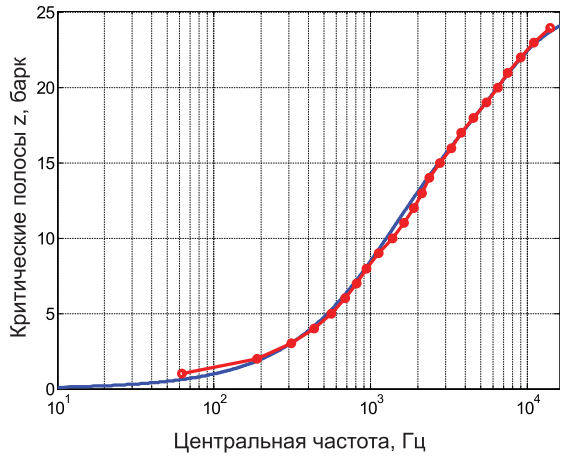

Рис. 5. Аппроксимация центральных частот CB – WPD:(l,n) ∈ E_{CB}, $l=\overline{0,7}$

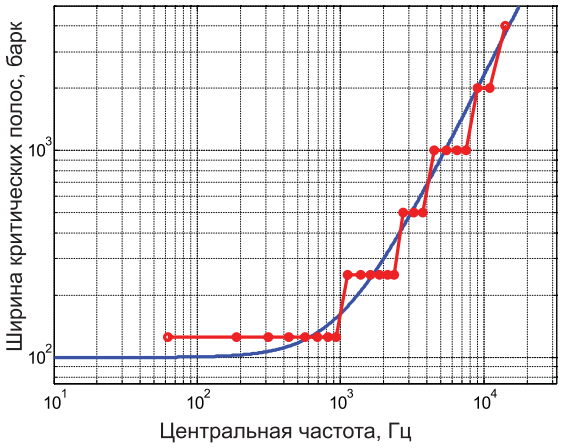

Рис. 6. Аппроксимация ширины критических частотных полос CB – WPD:(l,n) ∈ E_{CB}, $l=\overline{0,7}$

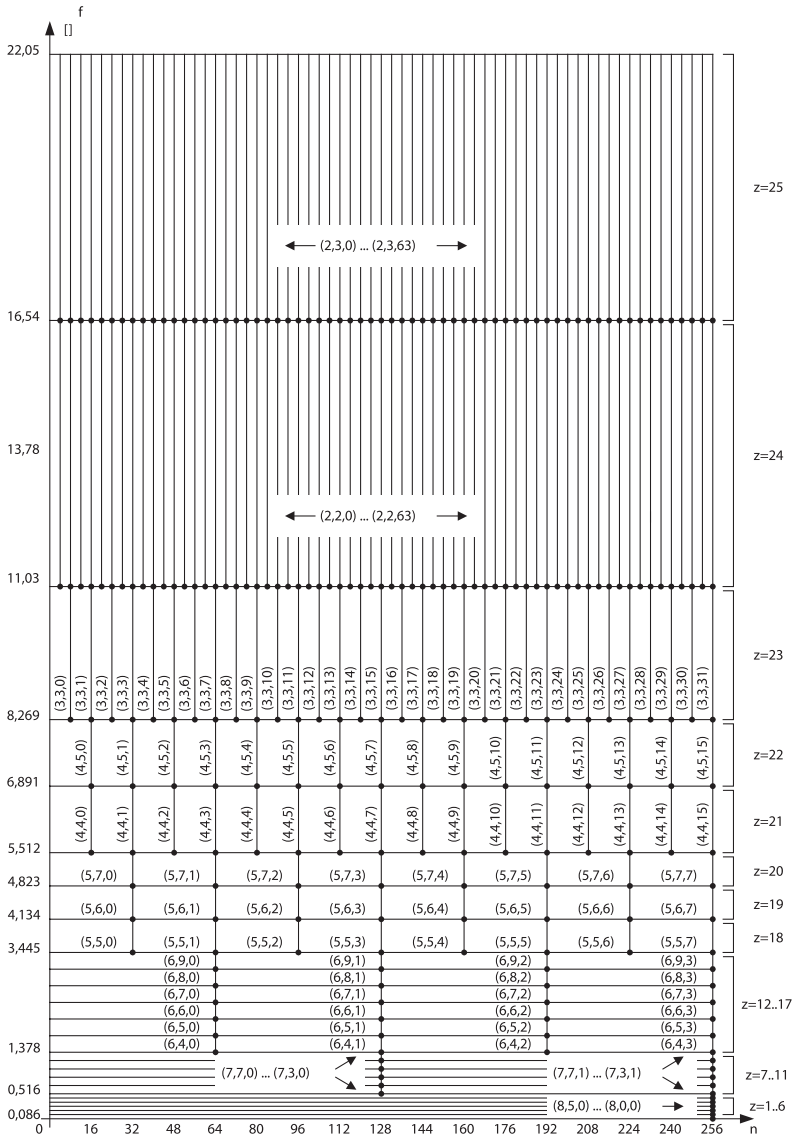

Рис. 7. Частотно-временной план структуры дерева ПДВП CB – WPD:(l,n) ∈ E_{CB}, $l=\overline{0,8}$, f_s = 44.1 кГц

Частотно-временной план для структуры дерева ПДВП $CB-WPD:(l,n)\in E_{CB}$, $l=\overline{0,8}$, $f_s=$ 44.1 кГц (рис. 1) [3,8] представлен на ***рис. 7***. Ширина каждой клеточки есть длина фрейма и определяется как $F_l=2^l$ ($F_{min}=2$ отсчётам и $F_{max}=256$ отсчётам). Следовательно, длина анализируемого окна равна $W=(P-1)\,(F_{l-1})+1$ отсчётов. Для первого уровня $l=1$ преобразования определяющей является область верхних частот и длина окна W равна 40 отсчётам при длине фильтра прототипа $P=40$. Для уровня $l=8$ преобразования наибольшая частотная разрешающая способность находится в области нижних частот, а длина окна W равна 9946 отсчётам.

В ***табл. 1*** и ***табл. 2*** приведены показатели аппроксимации критических частотных полос структурами деревьев ПДВП $CB-WPD:(l,n)\in E_{CB}$, $l=\overline{0,8}$ и $CB-WPD:(l,n)\in E_{CB}$, $l=\overline{0,7}$ соответственно.

Таблица 1

Аппроксимации критических частотных полос деревом ПДВП $(l,n)\in E_{CB}$, $l=\overline{0,8}$, $f_s=44.1$ кГц

№ барка	Узел	Кол-во вейвлет-коэфф.		Параметры полосы (Гц)		
z	l	n	K	Нижняя	Центр	Верхняя
1	8	0	1	0.00	43.07	86.13
2	8	1	1	86.13	129.20	172.27
3	8	2	1	172.27	215.33	258.40
4	8	3	1	258.40	301.46	344.53
5	8	4	1	344.53	387.59	430.66
6	8	5	1	430.66	473.73	516.80
7	7	3	2	516.80	602.93	689.06
8	7	4	2	689.06	775.19	861.33
9	7	5	2	861.33	947.46	1033.59
10	7	6	2	1033.59	1119.72	1205.86
11	7	7	2	1205.86	1291.99	1378.13
12	6	4	4	1378.13	1550.39	1722.66
13	6	5	4	1722.66	1895.42	2068.19
14	6	6	4	2068.19	2239.95	2411.72
15	6	7	4	2411.72	2583.98	2756.25
16	6	8	4	2756.25	2928.51	3100.78
17	6	9	4	3100.78	3273.04	3445.31
18	5	5	8	3445.31	3789.84	4134.37
19	5	6	8	4134.37	4478.89	4823.41
20	5	7	8	4823.41	5167.95	5512.50
21	4	4	16	5512.50	6201.56	6890.63
22	4	5	16	6890.63	7579.69	8268.75
23	3	3	32	8268.75	9646.87	11025.00
24	2	2	64	11025.00	13781.16	16537.50
25	2	3	64	16537.50	19293.75	22050.00

Таблица 2

Аппроксимации критических частотных полос деревом ПДВП $(l,n) \in E_{CB}$, $l = \overline{0,7}$, $f_s = 32$ кГц

№ барка	Узел	Кол-во вейвлет-коэфф.		Параметры полосы (Гц)		
z	l	n	K	Нижняя	Центр	Верхняя
1	7	0	1	0.00	62.5	125.0
2	7	1	1	125.0	187.5	250.0
3	7	2	1	250.0	312.5	375.0
4	7	3	1	375.0	437.5	500.0
5	7	4	1	500.0	562.5	625.0
6	7	5	1	625.0	687.5	750.0
7	7	6	1	750.0	812.5	875.0
8	7	7	1	875.0	937.5	1000.0
9	6	4	2	1000.0	1125.0	1250.0
10	6	5	2	1250.0	1375.0	1500.0
11	6	6	2	1500.0	1675.0	1750.0
12	6	7	2	1750.0	1875.0	2000.0
13	6	8	2	2000.0	2125.0	2250.0
14	6	9	2	2250.0	2375.0	2500.0
15	5	5	4	2500.0	2750.0	3000.0
16	5	6	4	3000.0	3250.0	3500.0
17	5	7	4	3500.0	3750.0	4000.0
18	4	4	8	4000.0	4500.0	5000.0
19	4	5	8	5000.0	5500.0	6000.0
20	4	6	8	6000.0	6500.0	7000.0
21	4	7	8	7000.0	7500.0	8000.0
22	3	4	16	8000.0	9000.0	10000.0
23	3	5	16	10000.0	11000.0	12000.0
24	2	3	32	12000.0	14000.0	16000.0

2. Процедура расчёта порогов маскирования в вейвлет-области

Пусть даны $CB - WPD{:}(l,n) \in E_{CB}$, частотно-временной план дерева ПДВП E_{CB} и коэффициенты $X_{l,n,k}$. Процедура расчёта порогов маскирования выглядит следующим образом [8]:

- Вычислить спектральную энергию барка:

$$A_{CB}(z) = \sum_{k=0}^{K-1} X_{z,k}^2 , \qquad (11)$$

- Оценить тональность сигнала в каждой критической частотной полосе и значения индексов $a_{tmn}(z)$ и $a_{nmn}(z)$ уменьшения спектральной энергии барка соответственно для тоновых и шумовых маскеров:

— индекс $a_{tmn}(z)$, который оценивает отношение маскирования тоном шума, задается так:

$$a_{tmn}(z) = -0.275 \cdot z - 15.025 [\partial Б]; \quad (12)$$

— индекс маскирования шумом шума оценивается как константа

$$a_{nmn}(z) = -25 [\partial Б], \quad (13)$$

так как ПДВП уже внесло некоторое спектральное перекрытие;

— среднее значение тональности маскеров в каждой критической полосе определяется маскирующим индексом

$$a_{CB}(z) = \eta \cdot a_{tmn}(z) + (1-\eta) \cdot a_{nmn}(z) [\partial Б], \; z = \overline{1.25}, \quad (14)$$

где η — тональный коэффициент:

$$\eta = \min\left(\frac{SFM(z)_{\partial Б}}{SFM_{\partial Б \max}}, 1\right), \quad (15)$$

где $SFM_{\partial Б}$ — мера спектральной пологости [11], рассчитываемая как

$$SFM(z) = \frac{\left(\prod_{k=0}^{K-1} X_{z,k}^2\right)^{\frac{1}{K}}}{\frac{1}{K}\sum_{k=0}^{K-1} X_{z,k}^2}, \quad (16)$$

$SFM_{\partial Б \max}$ — максимальное значение меры пологости спектра. Для заданного фильтра прототипа $SFM_{\partial Б \max}$ = –25 дБ.

- Найти спектральную энергию барка z с учётом тональности сигнала:

$$D_{CB}(z) = 10 \cdot \log\left(A_{CB}(z) \cdot 10^{\frac{a_{CB}(z)}{10}}\right) [\partial Б], \quad (17)$$

- Вычислить разброс энергии барка $C_{CB}(z)$ как свёртку $D_{CB}(z)$ с функцией разброса $B(z)$ в каждой критической частотной полосе z:

$$C_{CB}(z) = 10 \cdot \log\left(\frac{1}{K}\sum_{k=1}^{25} 10^{\frac{D_{CB}(k)}{10}} \cdot 10^{\frac{B(z-k)}{10}}\right) [\partial Б], \; z = \overline{1.25}, \quad (18)$$

где функция разброса $B(z)$ вычисляется как

$$B(z) = a + \frac{v+u}{2} \cdot (z+c) - \frac{v-u}{2} \cdot \sqrt{d + (z+c)^2}, \quad (19)$$

а параметры функции a, v, u, d, c приведены в первой строке ***табл. 3***.

Таблица 3

Параметры функции разброса

Функция разброса	v	u	d	c	a
Барк-шкала	30 дБ/барк	–25 дБ/барк	0.3	0.05	15
Временная шкала	0.0825 дБ/F^*_{min}	–0.0412 дБ/F^*_{min}	0.3	0.157	0.032/F^*_{min}

- Вычислить пороги маскирования во временной области. Аналогично, частотному маскированию, во временном маскировании уже присутствуют некоторые элементы перекрытия в АЧХ, обусловленные деревом ПДВП. Предполагается, что временное маскирование аддитивно к сигналу и определяется через коэффициенты ПДВП в каждой критической частотной полосе (см. ***рис. 8***). Максимальное временное разрешение для ПДВП имеет место в критических частотных полосах верхних частот, которые имеют минимальную протяжённость по времени $F^*_{min} = 2$ отсчёта или 0.0454 *мс*. Временная функция разброса $B(k)$ в вейвлет-области задаётся так:

$$B(k) = a + \frac{v+u}{2}\cdot(k+c) - \frac{v-u}{2}\cdot\sqrt{d+(k+c)^2}\ [\partial Б]. \qquad (20)$$

Её параметры вдоль оси времени определяются как v = 20 *дБ/мс* = 0.0825 *дБ*/F^*_{min} и u = 20 *дБ/мс* = –0.0412 *дБ*/F^*_{min} (см. табл. 3, строка 2. F^*_{min} — минимальная длина анализируемого фрейма). На рис. 8 схематически показано проявление временного маскирования.

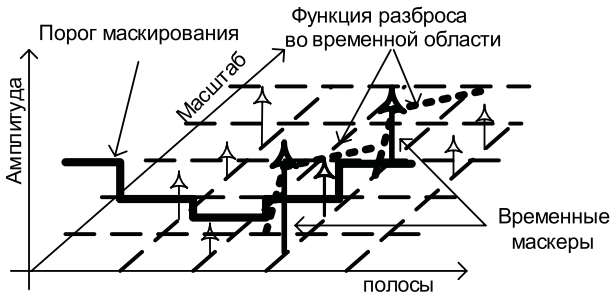

Рис. 8. Схема временного маскирования в соответствующей критической частотной полосе

- Вычислить энергию вейвлет-коэффициентов в каждой критической частотной полосе z:

$$E_z(k) = X^2_{z,k},\ k = \overline{1, K-1}, z = \overline{1.25}. \qquad (21)$$

- Определить временную функцию разброса энергии в каждой критической частотной полосе z как свёртку $E_z(k)$ и функции разброса $B(k)$:

$$F_z(m) = \frac{1}{K}\sum_{k=0}^{K-1} E_z(k)\cdot 10^{\frac{B(K-k)}{10}},\ m = \overline{1, K-1}. \qquad (22)$$

- Найти временной фактор маскирования в полосе z как результат сравнения величин:

$$F_z(m) \geq E_z(m), k = \overline{1, K-1}. \qquad (23)$$

Если данное соотношение выполняется, то в соответствующей критической частотной полосе имеет место временное маскирование, в противном случае его нет.

- Оценить частотно-временной маскирующий порог $M_{CB}(z)$ в каждой критической частотной полосе:

$$M_{CB}(z) = C_{CB}(z) \cdot \max\left(\frac{F_z(k)}{E_z(k)}, 1\right) [\partial Б], k = \overline{1, K-1}. \quad (24)$$

- Вычислить значение маскирующего порога $T_{CB}(z)$ в соответствующей критической частотной полосе из сравнения временно-частотного маскирующего порога $M_{CB}(z)$ с минимальным значением абсолютного порога слышимости $ATH(z)$ (3):

$$T_{CB}(z) = \max\left(ATH(z), M_{CB}(z)\right)[\partial Б]. \quad (25)$$

Заключение

Предложенная процедура расчёта параметров психоакустической модели в области вейвлет-коэффициентов нашла применение при реализации систем кодирования звуковых и широкополосных речевых сигналов [3], а также используется в моделировании переходных компонент сигнала на основе согласованной подгонки с адаптивным словарём, формируемым структурой дерева ПДВП в параметрических кодерах [12]. Ведение обработки сигнала и реализация психоакустической модели в одной области исключает необходимость перерасчёта параметров модели, что позволяет предотвратить дополнительную погрешность.

Литература

1. *M. Vetterli, J. Kovacevic.* Wavelets and Subband coding. — Englewood Cliffs. NJ: Prentice-Hall, 1995. 488 p.

2. *A. Spanias, T. Painter, V. Atti.* Audio signal processing and coding. — Hoboken, NJ: John Wiley & Sons, Inc., 2007. 464 p.

3. *Al. Petrovsky, D. Krahe, A.A. Petrovsky.* Real-time wavelet packet-based low bit rate audio coding on a dynamic reconfiguration system. // AES, Convention paper 5778, 114th Convention, 2003 March 22-25, Amsterdam, The Netherlands. 22 p.

4. *Петровский А.* Субполосное перцептуальное кодирование звуковых и речевых сигналов. Мн: Бестпринт, 2008 г. 220 с.

5. *E. Zwicker, H. Fastl.* Psychoacoustics: Facts and Models. — Berlin, Germany: Springer-Verlag, 1990. 380p.

6. *M.V. Wickerhauser.* Adaptive Wavelet Analysis from Theory to Software. — A.K. Peters Ltd., Massachusetts, 1994. 486 p.

7. ITU-R Rec. BS.1387, Method for Objective Measurements of Perceived Audio Quality, 998.

8. *Al. Petrovsky, M. Parfieniuk, A. Borowicz, A. Petrovsky.* Auditory modeling via frequency warped transforms. // AES, Convention paper 7446, 124th Convention, May 2008, Amsterdam, The Netherlands. 15 p.

9. *D. Sinha, A.H. Tewfik.* Low bit rate transparent audio compression using adapted wavelets. // IEEE Trans. Signal Processing, vol.41, no.12, pp. 3463–3479, Dec. 1993.

10. B.Carnero, A.Drygajlo. Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms. // IEEE Trans. Signal Processing, vol.47, no.6, pp. 1622–1635, June 1999.

11. *Ковалгин Ю.А., Вологдин Э.И.* Цифровое кодирование звуковых сигналов. Санкт-Петербург: Корона-Принт, 2004 г. 240 с.

12. *Al. Petrovsky, E. Azarov, A. Petrovsky.* Harmonic representation and auditory model-based parametric matching and its application in speech/audio analysis. // AES, Convention paper 7705, 126th Convention, May 2009, Munich, Germany. 13 p.

Выбор языковых средств пользователем для формулирования ответа в ходе диалога

Е.А. Снюгина,
научный сотрудник

Одной из основных задач при создании IVR-системы является моделирование потенциальных ответов пользователя в ходе диалога человек-компьютер. Чем точнее будет смоделирована грамматика для входных фраз, тем выше будет уровень распознавания. В работе мы рассматриваем, от чего зависит набор языковых средств, формирующих ответ пользователя. По результатам проведённого нами эксперимента указаны обнаруженные зависимости.

Abctract

One of the main tasks connected with the creation of IVR systems is modelling the potential answers of user in the course of human-computer dialogue. The more precisely the grammar for the input phrases is modelled, the higher the recognition level will be. In the article we examine the linguistic means which form the user's answer. We summarize the dependences that we have discovered based on the results of the experiment conducted as part of our research.

Введение

В последнее время в связи с необходимостью автоматической обработки большого количества запросов системы IVR приобретают всё большее распространение. Они выполняют функции оператора и могут применяться в call-центрах различных компаний, в авиа- и железнодорожных кассах. Например, их используют для обработки запросов, связанных с заказом билетов, с пополнением банковского счета и т.д. Система IVR представляет собой диалоговую систему компьютер-человек. Запрос пользователя является одновременно ответом системе на предлагаемые возможности (которые могут быть представлены как в виде вопроса, так и в виде некоторой инструкции или перечисления вариантов ответа).

Для систем распознавания одной из важнейших задач является моделирование языковых грамматик. Чем точнее смоделирована грамматика, тем выше уровень распознавания. С одной стороны, необходимо постараться описать все потенциальные варианты ответов; с другой стороны, теоретически ответов может быть достаточно большое количество. Однако на практике они могут реализовываться не всегда, что может привести к избыточному моделированию и, таким образом, негативно сказаться на качестве распознавания. В нашем случае необходимо понять, от чего зависит набор языковых средств, формирующих высказывания ответа.

Эксперимент

Область разговорной речи исследована ещё не в полной мере. Для того чтобы рассмотреть, как люди спонтанно отвечают на вопросы, мы провели опрос по телефону. Между тестирующим и тестируемым отсутствовал зрительный контакт, и информация воспринималась так же, как и при диалоге с автоматическими системами.

Количество испытуемых: 56. Возрастная категория: 17–70 лет.

Участникам эксперимента задавались вопросы в виде теста. Всего было составлено 7 тестов по 2 диалога в каждом, по 2 вопросительные реплики[1] в каждом диалоге. Суть каждого диалога была одна и та же. В первом диалоге пользователю предлагалось выбрать вариант из сфер досуга: кинопоказы, театральные постановки или концерт, а затем, в зависимости от ответа, предлагалось выбрать из определённых кинотеатров, театров или концертных залов. В следующем диалоге обыгрывалась примерно такая же ситуация, только сначала предлагалось выбрать одну из интересующих стран, а затем понравившийся аспект: география, история, экономика или политическая ситуация.

Одну и ту же мысль можно сформулировать различными способами, поэтому мы поставили задачу рассмотреть влияние различных формулировок вопросительных высказываний на ответ пользователя. Таким образом составлено по 7 вариантов каждой вопросительной реплики (в каждом тесте использовалась своя). Например, реплики, касающиеся выбора интересующей страны: «Выберите интересующую вас страну: Англия, Франция, Германия, Италия»; «Выберите, какая страна вас интересует: Англия, Франция, Германия, Италия»; «Я могу предоставить вам информацию по Англии, Франции, Германии, Италии. О чём бы вы хотели узнать?»; «Я могу предоставить вам информацию по Англии, Франции, Германии и Италии. Выберите, что вас интересует»; «Я могу предоставить вам информацию по Англии, Франции, Германии и Италии. О чём вы хотите узнать?»; «Выберите, что вам нужно: Англия, Франция, Германия, Италия»; «Что вас интересует: Англия, Франция, Германия, Италия?». Каждый вариант отличался от другого в лексико-синтаксическом плане. Цель подобных различий в формулировках вопросительных реплик — возможность формирования различных потенциальных вариантов ответов (не по содержанию, а по форме) на вопросы с одним и тем же смыслом.

[1] Под вопросительными репликами мы подразумеваем высказывание, где пользователю предоставляются варианты ответа; в этой же реплике возможен вопрос типа «О чём бы вы хотели узнать?», «Что вас интересует?» и т.д.

Один из вариантов теста:
I диалог:

1) Приветствие. Я могу предоставить вам информацию из области кинопоказов, театральных постановок и концертов. Выберите, что вас интересует.

2) Выберите, какой кинотеатр вас интересует[2]: «Аврора», Дом кино, «Кристалл-Палас», «Художественный» (Выберите, какой театр вас интересует: Михайловский, БДТ, «Приют Комедианта», Балтийский дом. Выберите, какой концертный зал вас интересует: БКЗ «Октябрьский», ДК Ленсовета, ДК «Выборгский»).

II диалог:

1) Приветствие. Я могу предоставить вам информацию по Англии, Франции, Германии, Италии. О чём бы вы хотели узнать?

2) Выберите интересующий вас аспект: история, экономика, география, политическая ситуация.

Сразу оговоримся, что пользователи в данном случае были ограничены достаточно жёсткими рамками, так как в вопросе практически озвучивался ответ. Однако в любом случае выбор лексических средств, порядок слов, развёрнутость ответа зависели от пользователя. Отметим, что вопросы относились к существительным, что также ограничивает область нашего исследования, так как в конечном итоге основное внимание уделялось рассмотрению изменений падежных форм существительного (а вместе с тем и наличию того или иного предлога, относящегося к существительному) в зависимости от формулировки задаваемого вопроса.

Мы условно разбиваем вопросительные реплики на три группы — всего 199 реплик (***таблица 1***):

Таблица 1

Группы вопросительных реплик

I группа	II группа	III группа
Перечисление вариантов в именительном падеже (вопрос отсутствовал), например: «Выберите, что вас интересует: кинопоказы, театральные постановки, концерты»	Перечисление вариантов не в именительном падеже + вопрос, подразумевающий ответ не в именительном падеже, например, «Я могу предоставить вам информацию по Англии, Франции, Германии, Италии. О чём бы вы хотели узнать?»	Перечисление вариантов не в именительном падеже + вопрос, подразумевающий ответ в именительном падеже, например: «Я могу предоставить вам информацию по нескольким аспектам: по истории, экономике, географии и политической ситуации. Что вас интересует?»
134 реплики	39 реплик	26 реплик

[2] В данной реплике происходит уточнение по первому запросу — выбор конкретного кинотеатра, театра или концертного зала.

I группа.

134 реплики первой группы подразумевали, что выбранный в качестве ответа один из предложенных вариантов будет выражен в именительном падеже. В итоге в именительном падеже были даны 125 ответов. Из них 98 были даны именно в той форме, которую бы подразумевал краткий ответ на поставленный вопрос, т.е. на высказывание «Выберите, что вас интересует: кинопоказы, театральные постановки, концерты» ответ был «кинопоказы» и т.д., в зависимости от выбранного пользователем варианта. 21 ответ, помимо варианта ответа в именительном падеже, содержал ещё слова типа «наверное», «тогда», «давайте», «пускай», «скорее всего». Так как порядок слов в русском языке свободный, то место подобных слов точно определить невозможно. Однако и здесь форма ответа была краткой, например: «допустим, Англия» или «Дом кино, наверное». Подобные слова появляются в ответах и других групп и характерны для разговорной речи. Некоторые из них с семантикой неуверенности («наверное», «тогда», «пусть будет», «допустим») можно объяснить тем, что для пользователя данная ситуация — игровая и ему в принципе всё равно, что выбирать, а также и тем, что количество предложенных вариантов для данной области интереса было всего 3–4, то есть, по сути, выбирать было «не из чего», и пользователь как бы «смирялся» с тем, что ему предложили.

В целом можно сказать, что, вне зависимости от лексико-синтаксического наполнения вопросительной реплики, практически все ответы даны в краткой форме. Только один ответ из этих 125 ответов был: «Меня всё интересует... Именно из всего... Ну, для начала Франция» (на вопрос: «Что вас интересует: Англия, Франция, Германия, Италия?»).

II группа.

Отметим, что краткие ответы были характерны для всех групп вопросительных реплик. Единственное отличие — в том, что во второй и третьей группах ответ предусматривал ещё и наличие соответствующего предлога.

Во второй группе вопросительных реплик был заложен потенциальный вариант ответа либо в дательном, либо в предложном падеже. В 23 ответах падеж существительного согласуется с тем, который предполагается в вопросительной реплике. А именно, либо с самим вопросом[3]: **«О чём** бы вы хотели узнать?» — **«О политической ситуации»** (16 ответов), либо с падежом существительного, использованным при перечислении вариантов ответа[4]: «Я могу предоставить вам информацию по нескольким аспектам: **по истории, экономике, географии или политической ситуации.** О чём бы вы хотели узнать?» — **«По истории»** (5 ответов). 2 ответа были такие: «Итал**ии...** хотелось бы узнать»; здесь «Итал**ии**» согласуются с вопросительной репликой, но из-за отсутствия предлога в ответе остаётся неясным, с чем именно происходит согласование.

Однако из 39 вопросительных реплик 12 ответов были даны в именительном падеже. Необходимо сказать, что ответы давались в именительном падеже чаще, если вопросительная реплика из данной группы была не первая. В случае если данная реплика была третьей — четвёртой, то 9 из 23 потенциальных ответов были даны в именительном падеже. А в случае если первая — 3 из 16 возможных.

[3] Согласование с предложным падежом.

[4] Согласование с дательным падежом.

Таким образом, проанализировав предыдущие реплики соответствующих тестов, мы предполагаем, что если по ходу диалога звучат однотипные (требующие в ответе именительного падежа) реплики, то, даже несмотря на последующие модификации реплик (в том смысле, что они уже не требуют именительного падежа в ответе), ответы всё равно достаточно часто будут звучать в именительном падеже, так как, по-видимому, формируется некий шаблон, по которому строятся дальнейшие ответы на подобные вопросы.

Из 39 потенциальных ответов 21 ответ, как и в первой группе, представлен в краткой форме (в именительном падеже — 7 ответов, в дательном падеже — 4 ответа, в предложном падеже — 10 ответов). В этой группе ответов, как и в предыдущей, слова типа «допустим», «наверное» употреблены аналогично — с краткими ответами. Например, «Допустим, о театральных постановках» (3 на 39 ответов).

III группа.

Согласование ответов происходит по тому же принципу, что и в предыдущей группе. Но заметим, что, в отличие от предыдущей группы, где в ответе не подразумевался именительный падеж, в данной группе вопрос («Что вас интересует?» / «Выберите, что вас интересует») подразумевал ответ в именительном падеже. И в данном случае процентное соотношение ответов в именительном падеже выше, чем в предыдущей группе (12 из 39 во II группе и 19 из 26 в III группе). Выбор именительного падежа является приоритетным для этой группы. Наличие слов типа «наверное», «допустим» аналогично предыдущим группам (3 на 26 ответов).

Рассмотрим те ответы, которые мы ещё не описали. Подобные ответы мы не рассматриваем в контексте их отнесённости к тому или иному вопросу, так как они могут появиться вне зависимости от того, какой потенциальный ответ мы формируем.

А) Ответы, имеющие более развёрнутые формулировки. Можно сказать, что они дублируют в своём содержании часть вопроса:

1) «Я могу предоставить вам информацию по Англии, Франции, Германии, Италии. О чём бы вы хотели узнать?» — «По Италии хотел бы узнать».

2) «Что вас интересует: Англия, Франция, Германия, Италия?» — «Меня всё интересует. Именно из всего... Ну, для начала Франция».

3) «Я могу предоставить вам информацию по кинопоказам, театральным постановкам и концертам. О чём бы вы хотели узнать?» — «Персонально я хотел бы узнать о концертах».

4) «Я могу предоставить вам информацию по Англии, Франции, Германии и Италии. О чём вы хотите узнать?» — «Италии... хотелось бы узнать».

5) «Я могу предоставить вам информацию из области кинопоказов, театральных постановок и концертов. О чём вы хотите узнать?» — «Я хочу узнать, что в ближайшее время произойдёт в областях...»

На этих примерах достаточно ясно видно, что, несмотря на относительное согласование (то есть если в вопросе употребляется тот или иной глагол, то при ответе используется тот же глагол; падежные формы вопросительных реплик и ответов также согласуются), окончательный выбор языковых средств, как и порядок слов, остаются индивидуальными. В то же время, вероятно,

при большем количестве подобных ответов можно было бы проследить подробнее характерные для них тенденции.

Б) Предложение своей линии ведения диалога: «Выберите, какой кинотеатр вас интересует: «Аврора», Дом кино, «Кристалл-Палас», «Художественный». — «В зависимости от фильма и удобного сеанса»; «Я могу предоставить вам информацию по Англии, Франции, Германии и Италии. Выберите, что вас интересует.» — «А Испании там нет?». Закладывать грамматики подобных ответов в систему, ограниченную по словарю, нет необходимости из-за невозможности дальнейшего сотрудничества между человеком и системой таким образом. В данном случае можно предложить переформулировать вопрос или дать прослушать пример того, каким образом следует отвечать.

В) В отличие от предыдущего примера (где пользователем предлагался ответ, не соответствующий содержанию вопросительной реплики), в некоторых случаях пользователи выбирали два варианта ответа:
1) «Выберите, что вам нужно: «Аврора», Дом кино, «Кристалл-Палас», «Художественный». — «*А два можно?* Дом кино и «Кристалл».
2) «Выберите, какая страна вас интересует: Англия, Франция, Германия, Италия?» — «*А две можно?* Тогда Англия и Германия».

В случаях систем именно с такой подачей информации необходимо учитывать возможность подобных запросов и либо изначально оговаривать возможность выбора только одного варианта, либо иметь возможность удовлетворять несколько направлений запроса. Отметим, что для разговорной речи характерно отсутствие выражения тех или иных элементов, этим и объясняется в вопросе «А два можно?» выпадение слова «кинотеатр», а в вопросе «А две можно?» — слова «страна».

Г) В одном из вариантов ответа прозвучало: «Ещё раз, пожалуйста. География». Заметим, что глагол в данном запросе опущен, что является одной из особенностей разговорной речи. Это говорит о необходимости возможности повтора запроса, а также включения в грамматику разговорного варианта для подобной просьбы.

Д) В 4 случаях происходит сокращение словосочетания ответа: вместо «театральные постановки», «о театральных постановках», «по театральным постановкам» пользователь говорит «театр», «о театре», «по театрам» соответственно, в 3 случаях вместо сложного существительного «кинопоказы» — «кино». Это говорит о тенденции к упрощению используемых средств выражения в разговорной речи.

Ж) В 4 случаях пользователь отказывается от предложенных вариантов, при этом используя разговорный стиль: «ни то, ни другое», «нет таких», «в данной ситуации ни о чём», — что надо иметь в виду при составлении грамматик.

З) Возможен длинный монолог пользователя (в основном, в случае, если принцип работы ему непонятен). Заложить грамматики подобных монологов, конечно, невозможно, поэтому в качестве выхода из подобной ситуации предлагается определение времени, в течение которого пользователь может осуществить свой ответ, и если он не укладывается в это ограничение, то автоматическое переключение на секретаря (так как лишние вопросы и уточнения будут вызывать только раздражение).

И) Одним из вариантов ответа прозвучало слово «последнее», что также свидетельствует о сильном влиянии тенденций разговорной речи при выборе языковых средств. Однако если подобный вариант малоупотребителен, то будет логичнее предлагать переформулировать ответ.

Выводы

1) При составлении грамматик необходимо учитывать особенности разговорной речи.

2) Общая база[5] позволяет коммуникантам не дублировать в диалоге то, что было выражено в предыдущем высказывании, и, затрачивая меньшее количество языковых средств, передать весь смысл задуманного. Таким образом, прослеживается тенденция к кратким формулировкам ответов, что наглядно представлено практически полным отсутствием развёрнутых предложений.

3) В случае, когда есть возможность сформулировать ответ в именительном падеже, большинство ответов будет даваться именно в такой форме, так как именительный падеж наиболее употребителен из всех падежных форм в разговорной речи [1].

4) Есть смысл на основе данных по наиболее употребительным формам глаголов, существительных и пр. в разговорном языке задавать вопросы так, чтобы в гипотетическом ответе прозвучала именно наиболее употребительная форма. Отказавшись от менее употребительных форм слов в вопросе, мы таким образом зададим пользователю более простой путь для формулирования ответа и тем самым, по всей вероятности, сократим возможное количество ответов.

Литература

1. *Земская Е.А.* «Русская разговорная речь. Лингвистический анализ и проблемы обучения». М., 2006.
2. *Meisel William, Ph.D.,* «VUI visions. Expert views on effective Voice User Interface Design», 2006.
3. http://www.genling.nw.ru/Staff/Chernigo/Minerva/content.html

Снюгина Е.А. —

научный сотрудник фирмы «Центр Речевых Технологий». Филолог.
Область научных интересов — речевые коммуникации, психолингвистика.

[5] Общие знания, которые позволяют не реализовывать словесно всю фразу, т.е., например, в разговоре на вопрос: «Как ты себя чувствуешь?» естественно будет звучать формулировка: «Нормально». Нет необходимости произносить: «Я чувствую себя нормально» — более того, в разговорной речи подобная конструкция, на наш взгляд, была бы избыточной.

Текстозависимая верификация диктора по голосу на основе коллектива решающих правил

Т.В. Левковская,
кандидат технических наук

Рассматриваются основные этапы и методы обработки речевых сигналов при решении задач автоматического распознавания личности по голосу. Описывается экспериментальная система текстозависимой верификации диктора на основе коллектива решающих правил, в которой принятие решения выполняется путём анализа и объединения оценок трёх классификаторов: на основе методов динамического программирования, векторного квантования и сравнения интегральных характеристик голоса. Приводятся результаты экспериментальных исследований надёжности верификации дикторов на речевом материале изолированно произнесённых названий цифр, показывающие эффективность применения принципов коллективного принятия решения в задачах автоматического распознавания диктора по голосу.

Abctract

The basic stages and methods of speech signals processing for automatic person recognition by his/her voice are considered. The experimental text-dependent speaker verification system based on multi-stream approach is described. The decision-making is carried out by the analysis and fusion of the scores of three classifiers based on dynamic time warping (DTW) methods, vector quantization (VQ) and comparison of integrated voice characteristics. Results of experimental researches of speaker verification reliability on the speech material of isolated digits are given. The obtained results show effectiveness of multi-stream approach in automatic speaker recognition tasks.

Введение

В настоящее время большое внимание уделяется развитию биометрических технологий, которые предназначены для получения и использования индивидуальных биологических данных человека, называемых биометриками, в целях его идентификации.

Наряду с такими биометриками, как отпечаток пальца, рисунок радужной оболочки глаза, структура ДНК и др., использование индивидуальных характеристик голоса предоставляет бесконтактный, этически корректный способ получения биометрической информации, позволяет осуществить скрытое наблюдение за человеком и его идентификацию, обеспечивает возможность удалённого доступа к конфиденциальной информации, в том числе по телефону.

Задача распознавания диктора по голосу может быть разделена на две подзадачи: идентификация и верификация. Идентификация диктора — это процесс определения говорящего из заданного набора дикторов. Распознаваемый голос сравнивается с эталонными голосами, и из набора выбирается тот диктор, голос которого в наибольшей степени соответствует данному. В случае верификации говорящий вначале предъявляет свой идентификатор (объявляет, кто он такой), а затем система определяет, принадлежит ли распознаваемый голос диктору с указанным идентификатором или нет. В задаче верификации при росте числа пользователей время принятия решения не увеличивается и является постоянным для различного числа пользователей. Это определяет возможность более широкого применения систем верификации, чем систем идентификации.

В последнее время разрабатывается множество экспериментальных и коммерческих приложений распознавания личности по голосу [1], которые применяются в банковских системах, системах безопасности и массового обслуживания в целях контроля доступа. Как правило, биометрические системы контроля доступа (пропускного контроля, доступа к информации, физической защиты закрытых объектов от несанкционированного проникновения посторонних лиц и т.п.) строятся на принципах автоматической верификации. Большинство верификаторов личности по голосу являются зависимыми от текста. Верификация диктора осуществляется по фиксированной парольной фразе, которая может быть изменена, или по конкретным словам, порядок произношения которых определяется самой системой случайным образом. В последнем случае снижается возможность фальсификации голосового пароля.

1. Принципы построения верификатора диктора по голосу

Основные компоненты и принципы функционирования систем автоматического распознавания личности по голосу детально описаны в ряде фундаментальных работ [2, 3]. Верификатор диктора по голосу функционирует в двух режимах: обучения и распознавания. Режим обучения предназначен для создания эталонных моделей голосов пользователей системы. Каждой модели ставится в соответствие идентификатор диктора, на основании речи которого была построена модель. Созданные эталоны сохраняются в базе данных с указанием персонального идентификатора (кода). В режиме распознавания пользователь вводит свой код и произносит пароль. Основные этапы обработки речевых сигналов при распознавании диктора по голосу таковы:

- предобработка и выделение информативных признаков, характеризующих индивидуальные особенности голоса человека;
- классификация, т.е. сравнение информативных признаков с эталонами и вычисление оценки соответствия;

- принятие решения об индивидуальности говорящего путём сравнения полученной оценки соответствия с заранее установленным пороговым значением.

Информативными признаками, наиболее часто используемыми в современных системах автоматического распознавания речи, являются кепстральные коэффициенты, рассчитанные по параметрам линейного предсказания речи, и мел-кепстральные коэффициенты, полученные с помощью дискретно-косинусного преобразования. Для сохранения информации о динамике речевых характеристик параметрическое описание обычно дополняется дельта-параметрами, которые представляют собой производные по времени от полученных признаков.

Последовательность векторов параметров является динамической моделью. Распространённым способом нелинейного во времени сравнения эталонной и анализируемой последовательностей векторов параметров является метод динамического программирования. Моделирование речи при помощи методов векторного квантования является более компактным описанием признакового пространства при формировании эталонных моделей голосов и эффективным способом распознавания диктора по голосу. Модель векторного квантования представляет собой конечное число типичных для конкретного диктора векторов параметров, совокупность которых образует кодовую книгу. Сравнение интегральных характеристик, вычисленных на продолжительных интервалах речи, таких как средний основной тон, средний спектр, кепстр сигнала и др., также достаточно эффективно используется для выявления индивидуальных особенностей голоса диктора.

В задачах распознавания диктора широко используются вероятностные модели, к которым относятся скрытые Марковские модели, модели Гауссовых смесей.

В последнее время при реализации систем автоматической верификации диктора применяются комбинации как разных наборов информативных признаков, так и разных способов классификации [4, 5]. Принятие решения осуществляется путём анализа и объединения полученных оценок используемых классификаторов.

2. Способы оценки эффективности систем верификации

Основным критерием качества биометрических систем являются вероятности ошибок первого и второго рода. Ошибка первого рода — это вероятность ложного отказа в доступе клиенту, имеющему право доступа (FRR — False Rejected Rate). Ошибка второго рода — вероятность ложного доступа, когда система ошибочно опознает чужого как своего (FAR — False Accepted Rate). Коэффициент равной вероятности ошибок (EER — Equal Error Rate) представляет точку совпадения вероятностей ошибок первого и второго рода. Изменение соотношения ошибок 1-го и 2-го рода достигается за счёт изменения порога принятия решения.

Система с двумя типами ошибок имеет много возможных уровней порога принятия решения. Для оценки качества таких систем традиционно используется кривая относительной характеристики функционирования (ROC — Receiver Operating Characteristics) [3]. В общем случае вероятность ложного срабатывания (FAR) откладывается по горизонтальной оси, а вероятность верного распознавания (1 — FRR) — по вертикальной оси.

Для задач верификации диктора также применяется оценка, отражающая компромисс ошибок детектирования (DET — Detection Error Tradeoff) [2]. В случае DET-графика значения

ошибки откладываются по обеим осям (FAR — по горизонтальной оси, FRR — по вертикальной оси), что позволяет чётко отличать эффективность распознавания одной системы от другой.

3. Структура верификатора диктора по голосу на основе коллектива решающих правил

Для повышения надёжности верификации предлагается использовать подход, основанный на принципах коллективного принятия решения. В разработанной экспериментальной системе автоматической верификации диктора по голосу (***рис. 1***) предлагается использовать три типа классификаторов: на основе методов динамического программирования (ДП), векторного квантования (ВК) и сравнения интегральных характеристик (интегральный классификатор — ИК).

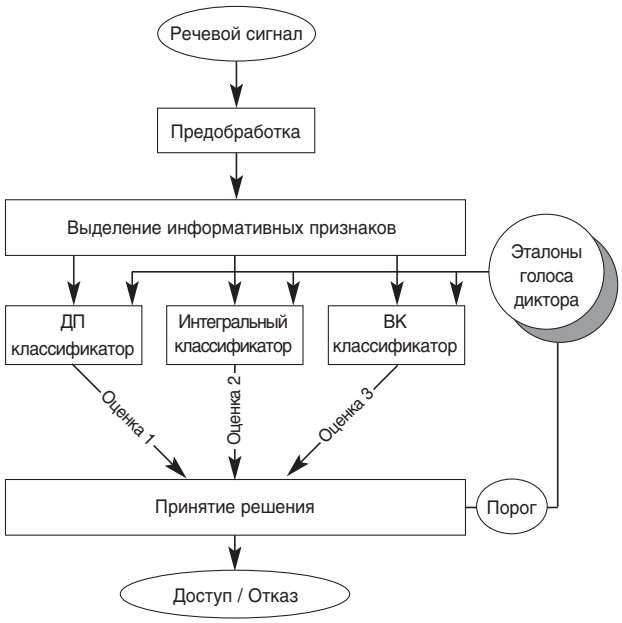

Рис.1. Общая схема верификатора на основе коллектива решающих правил

В качестве параметрического описания речевого сигнала использовались кепстральные коэффициенты, рассчитанные с помощью дискретно-косинусного преобразования, и их дельта-параметры. Нелинейное во времени сравнение параметрических описаний анализируемой речевой реализации и подготовленного в процессе обучения динамического эталона выполняется модифицированным ДП-методом [6, 7]. Главным достоинством этого метода является то, что он позволяет определить вероятность присутствия распознаваемых элементов речи в непрерывном речевом потоке и оценить их временное местоположение в условиях разного рода акустических помех.

На основе динамических параметров создаётся интегральная модель, представляющая собой средний кепстр по всей речевой реализации парольной фразы, а также формируется кодовая книга векторов параметров с помощью алгоритма векторного квантования. В режиме распознавания эталоны интегральной модели сравниваются с интегральными признаками входной речевой реализации, а кодовая книга сравнивается не с входными векторами, а с кодовой книгой входной реализации сигнала, подвергнутого такой же процедуре векторного квантования, как и при обучении.

Результатом сравнения каждого из классификаторов является расстояние между двумя моделями. Если расстояние меньше порогового значения, то диктор опознаётся как тот, за кого он себя выдаёт. Окончательное принятие решения может осуществляться как путём сравнения взвешенной суммы полученных оценок используемых классификаторов с порогом (рис. 1), так и путём объединения частных решений каждого классификатора в форме голосования.

4. Результаты экспериментальных исследований надёжности верификации дикторов

Исследования по надёжности верификации были проведены на одном и том же речевом материале для всех дикторов (названия цифр). Для тестирования была создана экспериментальная речевая БД, включающая образцы голосов 35 взрослых дикторов (13 женщин, 22 мужчин). Запись производилась с периодичностью 14 дней в офисных условиях. Каждый диктор произносил цифры от 0 до 9 по 2 раза. Речевые сигналы записывались в монорежиме с частотой дискретизации 11025 Гц, 16 бит на отсчёт. Всего было выполнено 5 серий записи.

БД условно была разделена на две части. Первая часть включала 15 дикторов (5 женщин и 10 мужчин). Эталоны каждого диктора создавались на речевом материале первых двух серий записи, после чего определялись пороговые значения для принятия идентификационного решения. Остальные серии записи использовались для тестирования. Во вторую часть входили оставшиеся дикторы, которые не участвовали в процессе обучения. Экспериментальная оценка эффективности системы проводилась на речевом материале обеих групп дикторов.

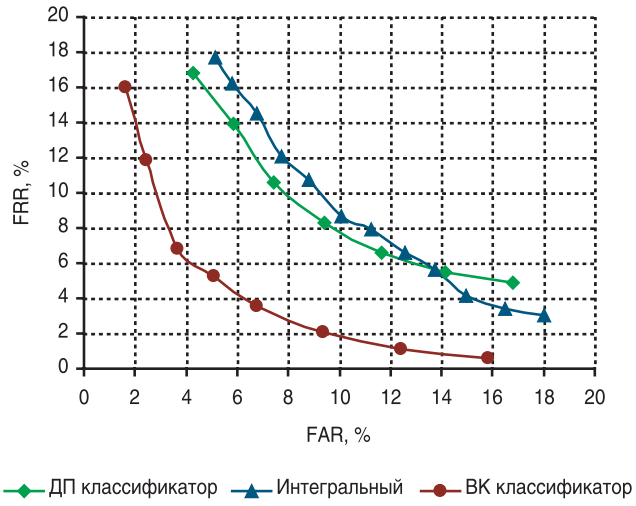

Рис. 2. DET зависимости ошибок верификации при использовании трёх типов классификаторов

На ***рис. 2*** показаны DET зависимости процента ошибок первого (FRR) и второго (FAR) рода при независимом использовании классификаторов ДП, ИК и ВК. Экспериментальные данные получены на речевом материале изолированно произнесённых названий цифр первой группы дикторов. Представленные результаты наглядно демонстрируют более низкий уровень ошибок верификации (примерно в 2 раза) при использовании ВК классификатора по сравнению с двумя остальными.

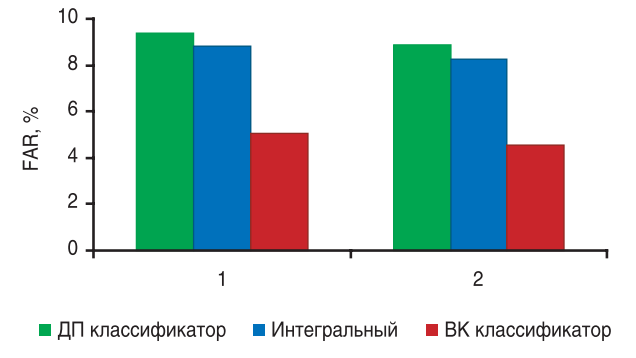

Рис. 3. Ошибки ложного доступа для своих (1) и чужих (2) дикторовпри использовании трёх типов классификаторов

Результаты ошибок ложного доступа (FAR) для двух групп дикторов представлены на ***рис. 3***. В первую группу входили свои дикторы, эталонные модели голосов которых были созданы в процессе обучения; во вторую — чужие дикторы, для которых эталоны отсутствовали. Из рис. 3 следует, что средние значения FAR практически не отличаются для своих и чужих дикторов при использовании каждого из классификаторов.

Во второй серии экспериментов были получены сравнительные результаты верификации на речевом материале разной длительности. Принятие решение выполнялось на основе анализа частных решений, полученных для каждой из произнесённых цифр, в форме голосования. Зависимость коэффициента равной вероятности ошибок (EER) от длительности анализируемой последовательности цифр, состоящей из одной, трёх, пяти, семи и десяти цифр соответственно, изображена на ***рис. 4***. Значение EER уменьшается примерно по экспоненте для каждого классификатора. Лучший результат получен для ВК классификатора.

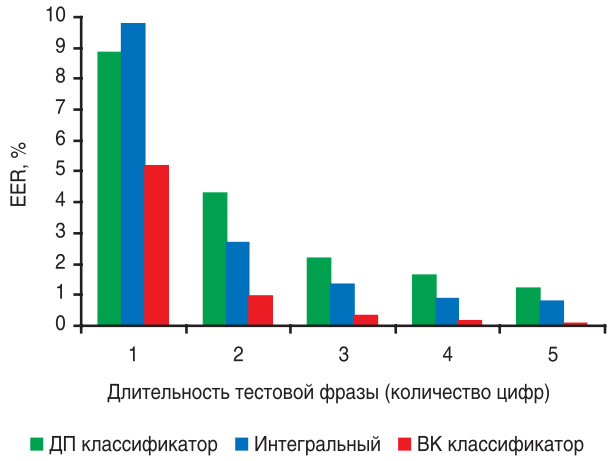

Рис. 4. Зависимость коэффициента равной вероятности ошибок (EER) от длительности анализируемой последовательности цифр (1, 3, 5, 7 и 10 цифр соответственно)

Последняя серия экспериментов связана с исследованием надёжности верификации при совместном использовании всех трёх классификаторов, а также их комбинаций. Средние значения ERR первой группы дикторов (***таб. 1***) и FAR двух групп дикторов (***таб. 2***) получены при анализе речевых сообщений разной длительности. Как и в предыдущей серии экспериментов, длительность тестовой фразы определялась количеством слов (1, 3, 5, 7 и 10 цифр).

Таблица 1

Коэффициенты равной вероятности ошибок (%) текстозависимой верификации дикторов на основе коллектива решающих правил

Классификаторы	Длительность тестовой фразы (кол-во цифр)				
	1	3	5	7	10
ДП, ВК	5,249	2,271	1,686	0,271	0
ИК, ВК	6,88	1,849	0,832	0,33	0,067
ДП, ИК	6,597	1,822	0,599	0,253	0,202
ДП, ИК, ВК	5,377	1,313	0,387	0,119	0

Таблица 2

Ошибки ложного доступа (%) текстозависимой верификации своих (1) и чужих (2) дикторов на основе коллектива решающих правил

Классификаторы	Группа дикторов	Длительность тестовой фразы (кол-во цифр)				
		1	3	5	7	10
ДП, ВК	1	5,782	0,502	0,204	0,086	0
	2	5,221	0,803	0,55	0,501	0,404
ИК, ВК	1	6,779	0,963	0,474	0,283	0,067
	2	6,368	1,406	1,021	0,923	0,779
ДП, ИК	1	6,59	1,002	0,506	0,348	0,202
	2	6,274	1,522	1,138	1,031	0,966
ДП, ИК, ВК	1	5,849	0,615	0,265	0,16	0
	2	5,246	1,066	0,817	0,765	0,685

Сравнительный анализ полученных результатов показывает значительное снижение ошибок верификации при увеличении длительности анализируемой тестовой фразы. Результаты экспериментальных исследований позволяют сделать вывод, что использование нескольких классификаторов обеспечивает уменьшение коэффициента равной вероятности ошибок первого (ложный отказ в доступе) и второго (ложный доступ) рода и увеличивает надёжность верификации.

Заключение

Результаты экспериментальных исследований наглядно демонстрируют эффективность применения принципов коллективного принятия решения в задачах автоматического распознавания диктора по голосу. Коллектив решающих правил в разработанной системе может быть дополнен. Система открыта для использования дополнительных критериев принятия решения с целью дальнейшего повышения надёжности верификации и устойчивости системы в реальных условиях её функционирования.

Направления дальнейших исследований связаны с разработкой формантного анализатора речевых сигналов, анализом просодических характеристик речи, использованием статистических методов распознавания голосов дикторов, применением аппарата нечёткой кластеризации для решения задачи отбора наиболее информативных признаков и классификации в условиях присутствия шумов и помех, проведением экспериментальных исследований с использованием доступных речевых баз данных.

Основные результаты получены в ходе выполнения научно-исследовательской работы «Разработка экспериментальной бимодальной биометрической системы контроля доступа

на основе характеристик лица и голоса человека» государственной комплексной программы научных исследований «Научные основы информационных технологий и систем».

Литература

1. *Хитров М.В.* Речевые технологии на CEBIT 2008 // Речевые технологии. — С.-Петербург: Издательский дом «Народное образование», 2008, № 2. С. 79–80.

2. *Rosenberg A.E., Soong F.K.* Recent research in automatic speaker recognition, in Advances in Speech Signal Processing, Furui, S. and Sondhi, M.M., Eds.,Marcel Dekker, New York, 1991. — P.701–738.

3. *Furui S.* «An overview of speaker recognition technology», ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 1994. — P.1–9.

4. *Rylov A.S., Chyzhdzenka V.A., Leukouskaya T.V.* The discriminant-stochastic approach of the speaker verification for entry control by the biometrical technologies // Proc. of the 9-th Intern. Conf. «Speech and Computer — SPECOM'2004, St.-Petersburg, 2004. — P. 377–381.

5. *Gupta H., Hautamaki, V., Kinnunen, T., Franti, P.* «Field Evaluation of Text-Dependent Speaker Recognition in an Access Control Application», Proc. of the 10th International Conference «Speech and Computer» — SPECOM'2005, Patras, Greece, 17–9 October, 2005. — P. 551–554.

6. *Lobanov B., Levkovskaia T.* Recognition of words and words-sequences in running speech // Proc Digital image processing. — Minsk: Institute Eng. Cybernetics, Academy of Science of Belarus, Minsk, 1997. — P. 154–161.

7. *Левковская Т.В.* Идентификация спектральных изображений речи // Анализ цифровых изображений. Мн.: ОИПИ НАН Беларуси, 2003. С. 186–193.

Левковская Т.В. —

кандидат технических наук, старший научный сотрудник Объединённого института проблем информатики НАН Беларуси. В 1997 г. защитила кандидатскую диссертацию «Исследование и разработка методов фонемного распознавания речи» (научный руководитель — доктор технических наук Б.М. Лобанов).
Область научных интересов — анализ сигналов, распознавание образов.

Алгоритмы преобразования сложноструктурированных объектовдля синтеза речи по тексту

Л.И. Цирульник,
кандидат технических наук
Ю.С. Гецевич,
аспирант

Одним из путей расширения использования систем синтеза речи является обработка и озвучивание не только текстовой информации, но и сложноструктурированных объектов, таких как таблицы, рисунки, формулы и т.д. Преобразование подобных объектов в орфографический текст является частным случаем задачи анализа сцен, но требует создания специальных алгоритмов, учитывающих структуру обрабатываемых объектов. При этом критерием корректности созданных алгоритмов должна являться, по мнению авторов, не только достоверность полученной информации, но и адекватность смыслового восприятия сформированного орфографического текста.
В работе рассматриваются сложноструктурированные объекты MS Word, предлагается шкала оценок сложноструктурированных объектов по критерию их сложности для смыслового восприятия, приводятся алгоритмы преобразования таких объектов в орфографический текст, описываются особенности программной реализации разработанных алгоритмов.

Abctract
One of the ways to extend the application of text-to-speech synthesis systems is to process not only textual information, but also complex-structured objects such as tables, figures, formulas, etc. The conversion of complex-structured objects into orthographic text is a particular case of scene analysis problem. Such conversion requires the development of special algorithms which take into account the structure of processed objects. The criteria of correctness of created algorithms are, on the author's opinion, not only accuracy of obtained information, but also adequacy of sense understanding of generated orthographic text.

The complex-structured objects of MS Word are considered in the paper, the rating scale of complex-structured objects by criterion of sense understanding complexity is suggested, the algorithms of transformation of complex-structured objects into orthographic text are given, the specificities of software implementation of developed algorithms are described.

Введение

При разработке систем синтеза речи по тексту предполагается, как правило, что на вход системы подаётся текстовый файл, содержащий, впрочем, не только орфографический текст, но и аббревиатуры, числа, сокращения и т.д. [1].Однако практика показывает, что такое ограничение на формат входного файла затрудняет широкое использование системы синтеза речи по тексту. Входная информация во многих случаях может содержаться в файлах в pdf, doc и других широко распространённых форматах.

Наиболее удачное решение озвучивания информации в произвольном формате — это предварительное её преобразование к одному определённому формату и последующая обработка стандартными модулями системы синтеза речи. При этом, очевидно, блок преобразования формата должен входить в состав системы синтеза речи.

Преобразование форматов входных данных требует решения дополнительных задач. Многие широко распространённые форматы допускают наличие в тексте рисунков, формул, таблиц и других объектов, требующих особой обработки. Оценкой корректности производимой обработки должна быть, по мнению авторов, адекватность смыслового восприятия сформированного текста.

В данной работе описываются исследования сложноструктурированных объектов MS Word и алгоритмы их преобразования в орфографический текст, который подаётся на вход системы синтеза речи по тексту. Под *сложноструктурированным объектом* понимается любой фрагмент содержимого файла MS Word, который: а) не является текстом и требует преобразования к текстовому формату либо б) является фрагментом текста, требующим особого интонационного выделения.

Примером первого типа сложноструктурированных объектов являются формулы, примером второго типа — заголовки.

1. Типы сложноструктурированных объектов MS Word и их классификация

Для выявления различных типов сложноструктурированных объектов были проанализированы текстовые документы MS Word, относящиеся к научному и художественному стилям. Соответствующие doc-файлы содержат в сумме 700 страниц и включают 43 таблицы, 67 формул и 211 рисунков.

В результате анализа выделены следующие основные типы сложноструктурированных объектов: оглавление, заголовок, список, перекрёстная ссылка, таблица, рисунок, формула.

Необходимо отметить, что в общем случае сложноструктурированные объекты могут быть вложенными: например, формула или рисунок могут являться содержимым ячейки таблицы.

Выявленные сложноструктурированные объекты, а также их составляющие были классифицированы по степени их значимости для смыслового восприятия текста. Для такой классификации разработана шкала оценок, приведённая в ***таблице 1***.

Таблица 1

Шкала оценок значимости объектов для смыслового восприятия текста

Оценка	Значение
0	Объект является незначимым и не должен быть преобразован в орфографический текст
1	Объект является значимым, но сложным для понимания смысла при восприятии соответствующего текста на слух
2	Объект является значимым; без его преобразования к орфографическому тексту и последующего озвучивания будет утерян смысл фрагмента текста

Детальный анализ сложноструктурированных объектов позволил выявить их структуру и оценить значимость самих объектов и их составляющих для смыслового восприятия текста.

1.1. Структура и степень значимости объекта «Оглавление»

Объект «Оглавление» представляет собой перечень строк. Каждая строка содержит текст заголовка, заполнитель (последовательность одинаковых символов) и номер страницы. Заполнитель и номер страницы не являются обязательными составляющими строки оглавления.

Степень значимости составляющих объекта «Оглавление», а также соответствующие пояснения приведены в таблице 2.

Таблица 2

Степень значимости составляющих объекта «Оглавление»

Наименование составляющей объекта «Оглавление»	Степень значимости	Комментарии
Текст заголовка	2	При озвучивании позволяет слушателю получить общее представление о содержимом текста
Заполнитель	0	Не является информативным
Номер страницы	1	При озвучивании позволяет слушателю получить представление об объёме соответствующего раздела

1.2. Структура и степень значимости объекта «Заголовок»

Объект «Заголовок» представляет собой абзац, отличительной особенностью которого является наличие в соответствующей строке особого стиля: например, «Заголовок 1», «Заголовок 2» и т.д.

Степень значимости объекта «Заголовок» для смыслового восприятия текста равна 2.

1.3. Структура и степень значимости объекта «Список»

Объекты «Список» в MS Word делятся на две основные категории: маркированные (характеризующиеся наличием маркера перед каждым элементом списка) и нумерованные (характеризующиеся наличием порядкового номера перед каждым элементом списка).

Каждый элемент и маркированного, и нумерованного списков может состоять, в общем случае, из части предложения (при перечислении), целого предложения или нескольких предложений. Элементы маркированного списка отделяются, как правило, точкой с запятой («;»), а элементы нумерованного списка заканчиваются, как правило, точкой.

Степень значимости для смыслового восприятия составляющих объекта «Список» приведена в таблице 3.

Таблица 3

Степень значимости составляющих объекта «Список»

Наименование составляющей объекта «Список»	Степень значимости	Комментарии
Текст элемента списка	2	Отражает смысловое содержание
Маркер	0	Не является информативным
Порядковый номер	2	Является элементом, способствующим пониманию смысла текста

1.4. Структура и степень значимости объекта «Перекрёстная ссылка»

В документе MS Word могут содержаться перекрёстные ссылки следующих типов: абзац, заголовок, закладка, сноска, концевая сноска, рисунок, таблица, формула.

В качестве перекрёстной ссылки в документе может находиться:

- текст объекта (заголовка, закладки, абзаца) либо название объекта (рисунка, таблицы, формулы);
- порядковый номер соответствующего объекта (абзаца, закладки, (концевой) сноски, заголовка; номер абзаца, в котором находится закладка);
- номер страницы, на которой находится соответствующий объект;
- слово «выше» или «ниже».

Степень значимости каждой из категорий приведена в таблице 4.

Таблица 4

Степень значимости различных категорий объекта «Перекрёстная ссылка»

Тип перекрёстной ссылки	Степень значимости	Комментарии
Текст либо название объекта	2	Является пояснением сути перекрёстной ссылки
Порядковый номер объекта	0; 2	При всех типах перекрёстных ссылок, кроме (концевой) сноски, не является информативной и имеет степень значимости, равную 0; в случае перекрёстной ссылки на (концевую) сноску степень значимости равна 2
Номер страницы, на которой находится соответствующий объект	0	Не является информативным
Слово «выше» или «ниже»	2	Содержит пояснительную информацию

1.5. Структура и степень значимости объекта «Таблица»

Таблица состоит из двух основных составляющих: названия и непосредственно таблицы.

Название таблицы находится, как правило, перед таблицей и состоит из слова «таблица», за которым может следовать порядковый номер, и наименования таблицы.

Степень значимости названия таблицы равна 2.

Непосредственно таблица имеет следующие характеристики, важные для её преобразования в текстовый вид: наличие или отсутствие заголовков (подзаголовков) строк и столбцов; количество заголовков (подзаголовков) строк и столбцов; количество столбцов; количество строк.

Степень значимости непосредственно таблицы зависит от её сложности, которая может быть вычислена в соответствии с формулой:

$$S = (k_r + k_{pr}) n_r + (k_c + k_{pc}) n_c + k_m \sum_{i=1}^{l} S_i , \qquad (1)$$

где k_r, k_{pr}, k_c, k_{pc}, k_m — коэффициенты сложности, соответственно, строк таблицы; заголовков строк таблицы; столбцов таблицы; заголовков столбцов таблицы; сложноструктурированных объектов, являющихся содержимым таблицы; n_r, n_c, l — количество, соответственно, строк, столбцов и сложноструктурированных объектов в таблице; S_i — сложность i-го сложноструктурированного объекта, входящего в таблицу.

Используемые в формуле (1) коэффициенты удовлетворяют уравнению:

$$k_r + k_{pr} + k_c + k_{pc} + k_m = 1$$

— и вычисляются экспериментальным путём.

Соответствие сложности таблицы и её степени значимости также определяется экспериментально; увеличение сложности таблицы влечёт уменьшение степени её значимости для смыслового восприятия текста.

1.6. Структура и степень значимости объекта «Рисунок»

Рисунок состоит из двух основных составляющих: подрисуночной подписи и непосредственно рисунка.

Подрисуночная подпись — это отдельный абзац, состоящий, как правило, из слова «Рисунок» (или «Рис.»), за которым может следовать порядковый номер рисунка и его название.

Степень значимости подрисуночной подписи для смыслового восприятия текста равна 2.

Непосредственно рисунки делятся на следующие основные категории:

- растровые изображения;
- объекты внешних приложений, позволяющие создавать графики, диаграммы и т.д., например, объекты MS Visio или MS Excel;
- рисунки, выполненные с использованием средств рисования MS Word (Word-рисунки);
- «смешанные» рисунки, содержащие две или более составляющих, принадлежащих к различным категориям.

Степень значимости непосредственно рисунка зависит от его типа и сложности. Обзор степеней значимости рисунков для смыслового восприятия приведён в таблице 5.

Таблица 5

Степень значимости составляющих объекта «Рисунок»

Категория рисунка	Степень значимости	Комментарии
Растровое изображение	0	В общем случае не может быть преобразовано в текст, описывающий его содержимое
Объект внешнего приложения	0	Требует обработки фрагментов, не являющихся объектами MS Word
Word-рисунок	0–2	В зависимости от сложности объекта
«Смешанный» рисунок	0	Содержит фрагменты, которые в общем случае не могут быть преобразованы в текст

Сложность рисунка определяется в зависимости от количества и типа составляющих его фигур и может быть вычислена по формуле:

$$S = k_l n_l + k_b n_b + k_a n_a + k_{bs} n_{bs} + k_m n_m + k_s n_s \qquad (2)$$

где k_l, k_b, k_a, k_{bs}, k_m, k_s — коэффициенты сложности, соответственно, линии, основной фигуры, фигурной стрелки, элемента блок-схемы, выноски, звезды или ленты; n_l, n_b, n_a, n_{bs}, n_m, n_s — количество, соответственно, линий, основных фигур, фигурных стрелок, элементов блок-схемы, выносок и звёзд или лент на рисунке.

Используемые в формуле (2) коэффициенты удовлетворяют уравнению:

$$k_l + k_b + k_a + k_{bs} + k_m = 1$$

— и вычисляются экспериментальным путём.

Соответствие сложности рисунка и его степени значимости также должно быть определено экспериментальным путём; при этом при увеличении сложности рисунка его степень значимости будет уменьшаться.

1.7. Структура и степень значимости объекта «Формула»

Объект «Формула» может рекурсивно включать в себя следующие структуры: дроби; верхние и нижние индексы; радикалы; крупные операторы (например, сумма, произведение и т.п.); круглые, квадратные, фигурные, угловые скобки и их разновидности; скобки с разделителями; тригонометрические, обратные тригонометрические, гиперболические, обратные гиперболические функции; диакритические знаки; матрицы.

Кроме того, формула может содержать обычный (не являющийся математическим) текст, а также следующие типы символов: основные математические символы; греческие буквы; буквоподобные символы; операторы; стрелки; отношения с отрицанием; геометрические символы.

Значимость объекта «Формула» для смыслового восприятия текста, как и объектов «Рисунок» и «Таблица», зависит от его сложности. Очевидно, что если формула обладает большой сложностью, то преобразование её в текстовый вид и последующее озвучивание вызовет у слушателя затруднения при восприятии сути формулы. Следовательно, при увеличении сложности формулы степень её значимости для смыслового восприятия текста уменьшается. Сложность формулы может быть вычислена в соответствии с рекурсивным выражением:

$$S = k_{fr}S_{fr} + k_iS_i + k_rS_r + k_pS_p + k_bS_b + k_{fn}S_{fn} + k_dS_d + k_mS_m \qquad (3)$$

где k_{fr}, k_i, k_r, k_p, k_b, k_{fn}, k_d, k_m — коэффициенты сложности, соответственно, дроби, индекса, радикала, крупного оператора, скобки, функции, диакритического знака, матрицы; S_{fr}, S_i, S_r, S_p, S_b, S_{fn}, S_d, S_m — среднее значение сложности, соответственно, дробей, индексов, радикалов, крупных операторов, скобок, функций, диакритических знаков, матриц.

Коэффициенты формулы (3) удовлетворяют уравнению:

$$k_{fr} + k_i + k_r + k_p + k_b + k_{fn} + k_d + k_m = 1 \qquad (4)$$

Для составляющих объекта «Формула», не содержащих внутри себя дробей, индексов, радикалов, крупных операторов, скобок, функций, диакритических знаков и матриц, сложность может быть вычислена в соответствии с выражением:

$$S = k_sn_s + k_{gr}n_{gr} + k_ln_l + k_on_o + k_an_a + k_rn_r + k_{gm}n_{gm} \qquad (5)$$

где k_s, k_{gr}, k_l, k_o, k_a, k_r, k_{gm} — коэффициенты сложности, соответственно, основных математических символов, греческих букв, буквоподобных символов, операторов, стрелок, отношений с отрицанием, геометрических символов; n_s, n_{gr}, n_l, n_o, n_a, n_r, n_{gm} — количество, соответственно, основных математических символов, греческих букв, буквоподобных символов, операторов, стрелок, отношений с отрицанием и геометрических символов в формуле.

Коэффициенты формулы (5) удовлетворяют уравнению:

$$k_s + k_{gr} + k_l + k_o + k_a + k_r + k_{gm} = 1 \qquad (6)$$

Коэффициенты, входящие в состав формул (4) и (6), вычисляются экспериментальным путём.

2. Алгоритмы преобразования сложноструктурированных объектов MS Word к орфографическому тексту

Как уже указывалось ранее, под сложноструктурированным объектом понимается не только фрагмент текста, который требует преобразования к текстовому формату, но и фрагмент, требующий особого интонационного выделения. Для того, чтобы распознанный в Word-документе объект второго типа мог быть идентифицирован на последующих этапах преобразования текста и соответствующим образом обработан, были введены специальные теги, добавляемые в выходной документ, которые являются индикаторами наличия того или иного объекта: например, строки оглавления, заголовка и т.д.

2.1. Алгоритмы преобразования объектов, требующих особого интонационного выделения

К таким объектам относятся: оглавление и его составляющие; заголовок; элементы списка; перекрёстная ссылка.

При обработке входного документа распознаётся наличие сложноструктурированного объекта, он разбивается на составляющие, и затем каждая из составляющих определённым образом обрабатывается. Составляющие, степень значимости которых для смыслового восприятия текста равна 0, удаляются из документа.

Правила преобразования этих объектов в процессе обработки документа представлены в таблице 6.

Таблица 6

Правила преобразования некоторых сложноструктурированных объектов в текст

Входной объект или его составляющая	Выходной текст
Текст заголовка оглавления	<content_a> Текст заголовка оглавления </content_a>
Номер страницы в заголовке оглавления	<content_p> Номер страницы в заголовке оглавления </content_p>
Заголовок	<paragraph_a > Заголовок </ paragraph_a >
Номер элемента списка	<list_n> Номер элемента списка</list_n>
Текст элемента списка	<list_t> Текст элемента списка </list_t>
Слова «Выше» («Ниже») как содержимое перекрёстной ссылки	<cross_reference_ab> «Выше» («Ниже») </cross_reference_ab>
Текст как фрагмент перекрёстной ссылки	<cross_reference_t> Текст </cross_reference_t>
Номер как фрагмент перекрёстной ссылки	<cross_reference_n> Номер </cross_reference_n>

2.2. Алгоритмы обработки объектов, требующих преобразования к текстовому формату

К объектам, требующим преобразования к текстовому формату, относятся таблицы, рисунки и формулы. В общем случае можно определить три варианта преобразования к текстовому формату таких объектов:

- **краткий,** когда для таблиц преобразуются к текстовому виду только их названия, для рисунков — только подрисуночные подписи, формулы заменяются словом «формула»;

- **средний,** при котором преобразуются к текстовому виду таблицы, рисунки и формулы, сложность которых не превышает некоторого заданного порога (причём значение порога будет различным для разных типов объектов); для объектов, сложность которых превышает порог, алгоритм преобразования будет таким же, как и в первом случае;

- **подробный,** при котором преобразуются к текстовому виду все объекты, вне зависимости от их сложности.

В последнем случае смысловое восприятие текста будет, вероятно, затруднено из-за большой сложности объектов. Тем не менее, целесообразно предоставить пользователю системы возможность выбирать наиболее подходящий для него режим.

Очевидно, что в любом случае не будут преобразованы к текстовому виду рисунки, которые являются растровыми изображениями или объектами внешних приложений.

2.2.1. Алгоритм преобразования таблиц в орфографический текст

На вход алгоритма подаётся таблица MS Word. В результате преобразования формируется орфографический текст, включающий заголовки и подзаголовки строк таблицы, столбцов таблицы, содержимое ячеек таблицы, а также специальные теги для заголовков и подзаголовков строк, столбцов и содержимого ячеек.

Алгоритм состоит из следующих шагов.

Шаг 1. Вычисляется общее количество строк таблицы n, количество заголовков и подзаголовков строк nc, общее количество столбцов таблицы m, количество заголовков и подзаголовков столбцов mc.

Шаг 2. Текущий номер строки i принимает значение $(nc+1)$.

Шаг 3. Текущий номер столбца j принимает значение (mc+1).

Шаг 4. Для всех k от 1 до nc формируется текст: <table_header_r> $T[i,k]$ </table_header_r>, где i — номер строки таблицы, k — номер столбца таблицы, $T[i, k]$ — содержимое ячейки i, k таблицы T.

Шаг 5. Для всех l от 1 до mc формируется текст: <table_header_c> $T[l,j]$ </table_header_c>.

Шаг 6. Формируется текст: <table_cell> $T[l,j]$ </table_cell>.

Шаг 7. Значение j увеличивается на 1.

Шаг 8. Если $j \le m$, переход к шагу 4. Иначе — переход к шагу 9.

Шаг 9. Значение i увеличивается на 1.

Шаг 10. Если $i \le n$, переход к шагу 3. Иначе — конец алгоритма.

Отличительными особенностями данного алгоритма являются следующие:

а) перечисление заголовков и подзаголовков (при их наличии) строки и столбца таблицы перед содержимым ячейки, которая находится на пересечении данных строки и столбца (благодаря такому перечислению из сформированного текста понятно, что именно отражает содержимое каждой ячейки; это особенно важно при «озвучивании» таблиц, содержащих большое количество строк и столбцов или несколько подзаголовков строк и столбцов);

б) вставка в текст специальных тегов, указывающих начало/конец заголовков (подзаголовков) строк и столбцов, а также начало/конец содержимого ячейки таблицы (благодаря вставке специальных тегов появляется возможность особого интонационного выделения заголовков и подзаголовков строк и столбцов, а также содержимого ячеек таблицы на последующих этапах обработки текста).

2.2.2. Алгоритм преобразования рисунков к орфографическому тексту

На вход алгоритма подаётся рисунок, включающий только фигуры MS Word. В результате преобразования формируется орфографический текст, содержащий названия фигур, тексты фигур, а также специальные теги для названий и текстов фигур.

Алгоритм включает следующие шаги.

Шаг 1. «Считывается» первая фигура от верхнего левого угла рисунка (определяется по координатам фигур).

Шаг 2. Если фигура не принадлежит к одному из классов «Основные фигуры», «Элементы блок-схемы», «Выноски», «Звёзды и ленты», то переход к шагу 5.

Шаг 3. Формируется текст: <figure_shape> <Название фигуры> </figure_shape>, где <Название фигуры> — текстовое название конкретной фигуры, например, «прямоугольник», «ромб», «куб» и т.д.

Шаг 4. Если внутри фигуры содержится текст, то формируется текстовая последовательность: <figure_text> <Текст фигуры> </figure_text>, где <Текст фигуры> — текстовое содержимое фигуры.

Шаг 5. Если «считаны» все фигуры, то переход к шагу 6, иначе — считывается очередная фигура по принципу «сверху вниз, слева направо» (определяется по координатам фигур) и осуществляется переход к шагу 2.

Шаг 6. Конец алгоритма.

Особенности данного алгоритма следующие:

а) к текстовому формату не преобразуются стрелки и соединительные линии (эта особенность связана с тем, что стрелка может быть направлена из объекта в тот же объект; стрелки могут циклически соединять несколько объектов, причём данные объекты не охватывают всего рисунка; из одного объекта может «выходить» несколько стрелок, направленных в различные объекты);

б) если одна и та же фигура входит в классы «Основные фигуры» и «Элементы блок-схемы», то при её преобразовании к тексту формируется название, взятое из класса «Основные фигуры» (к таким фигурам относится, например, ромб, название которого в классе «Основные фигуры» — «ромб», а в классе «Элементы блок-схемы» — «Блок-схема: решение»).

Как показали результаты экспертного тестирования разработанного алгоритма, выбор названия фигуры из класса «Основные фигуры» способствует лучшему смысловому восприятию сформированного текста.

2.2.3. Алгоритм преобразования формул к орфографическому тексту

На вход алгоритма подаётся формула MS Word. В результате преобразования формируется орфографический текст, содержащий названия элементов формулы и специальные теги для различных типов элементов.

Шаги алгоритма следующие.

Шаг 1. Переменной $Formula$ присваивается значение исходной формулы.

Шаг 2. Переменной F_i присваивается значение очередного элемента формулы.

Шаг 3. Если F_i не принадлежит ни одному из классов «дробь», «индекс», «радикал», «крупный оператор», «скобка», «функция», «диакритический знак», «матрица», «интеграл», «логарифм», то переход к шагу 7.

Шаг 4. В зависимости от класса F_i формируется определённый открывающий тег: например, для класса «дробь» — <formula_fr>, для класса «индекс» — <formula_i>, для класса «радикал» — <formula_r> и т.д.

Шаг 5. Переменной $Formula$ присваивается значение F_i; осуществляется рекурсивный переход к шагу 2.

Шаг 6. В зависимости от класса F_i формируется соответствующий закрывающий тег: например, для класса «дробь» — </formula_fr>, для класса «индекс» — </formula_i>, для класса «радикал» — </formula_r> и т.д. Затем осуществляется переход к шагу 8.

Шаг 7. Формируется текст, соответствующий названию элемента F_i.

Шаг 8. Если достигнут конец формулы, переход к шагу 9, иначе — переход к шагу 2.

Шаг 9. Конец алгоритма.

Особенностью алгоритма является то, что он учитывает наличие «сложных» элементов, таких как дробь, радикал, крупный оператор и т.д., которые, в свою очередь, обрабатываются так же, как исходная формула, благодаря рекурсивной структуре алгоритма. Кроме того, при реализации алгоритма была составлена таблица соответствия символов, используемых в формулах, и их текстовых эквивалентов, например, ε — «эпсилон», sin — «синус» и т.д.

3. Особенности программной реализации алгоритмов

Обработка файлов MS Word осуществляется в два этапа. На первом этапе входной файл в формате doc преобразовывается к валидному xml-документу. На втором этапе в xml-документе на основе тегов распознаются сложноструктурированные объекты, в результате преобразования которых формируется последовательный орфографический текст, включающий специальные теги, маркирующие сложноструктурированные объекты.

3.1. Особенности преобразования документа MS Word к xml-документу

Входными данными обработчика является произвольный документ MS Word. В процессе обработки формируется временный html-документ, который затем преобразовывается в корректный по оформлению и наполнению

xml-документ (***рис. 1***). Для обработки файлов в форматах html и xml используются стандартные библиотеки классов Microsoft (Microsoft.Office. Interop.Word.dll, SgmlReaderDll.dll).

Основным действием преобразования к xml-документу является фильтрация тегов и их атрибутов, после которой в результирующем документе остаются только теги и атрибуты, необходимые для идентификации и обработки сложноструктурированных объектов.

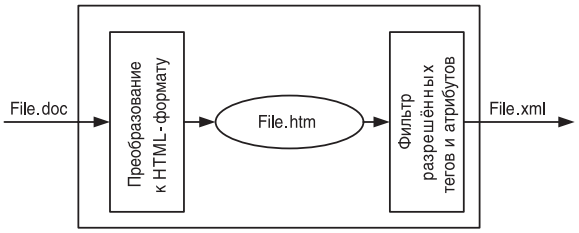

Рис. 1. Процесс преобразования документа MS Word в xml-документ

3.2. Особенности преобразования сложноструктурированных объектов в орфографический текст

Существует два подхода к обработке xml-документов, принципиальная разница между которыми заключается в способе загрузки документа и доступа к его содержимому. При первом подходе, реализованном в рамках DOM-технологии (Document Object Model — объектная модель документа) [2] xml-документ полностью загружается в оперативную память и после этого обрабатывается. При втором подходе, реализованном в рамках SAX-технологии (Simple API for XML) [3], обработка документа происходит последовательно, с использованием относительно небольшого буфера выделенной памяти, в котором хранится минимально необходимая для обработки документа информация.

Преимуществом DOM-технологии[1] является возможность произвольного доступа к содержимому документа. Специальные языки программирования, разработанные для DOM-технологии, позволяют очень удобно записывать правила обработки для целых групп тегов или конкретных атрибутов тегов. Основным недостатком DOM-технологии является использование большого объёма оперативной памяти, так как для обработки документа необходимо загрузить в память всё дерево тегов.

Преимуществом SAX-технологии является, наоборот, использование относительно небольшого объёма оперативной памяти. Платой за это являются определённые неудобства при программной реализации правил обработки xml-документов. Решение по обработке тега и его содержимого необходимо принимать почти вслепую, поскольку неизвестно, что было вложено в тег ранее и будет вложено дальше. Невозможно также обобщать правила обработки xml-документов; кроме того, такие правила программируются на языке, не приспособленном для обработки тегов с их атрибутами.

Реализация разработанных алгоритмов была осуществлена с использованием как DOM-обработчика, так и SAX- обработчика.

Для реализации в рамках DOM-технологии использовался язык преобразования xml-документов XSLT; правила преобразования записывались в файл стилей styles.xsl (***рис. 2***).

[1] Под DOM понимается технология, при которой для обработки документа требуется его полная загрузка в оперативную память.

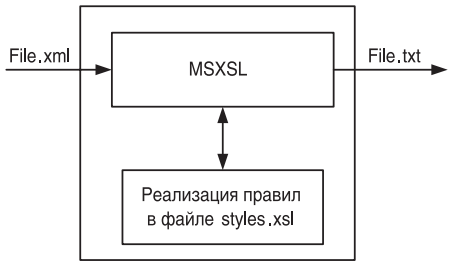

Рис. 2. Схема обработки xml-документа с использованием XSLT

Для выбора данных из исходного дерева тегов формировались запросы на языке запросов XPath. Применение языка XSLT осуществлялось с использованием процессора Microsoft XML Parser 4.0 [4]. Он принимает на вход исходное дерево тегов (xml-файл), выходными данными является файл как результат применения правил-стилей XSLT к xml-файлу. Обработка каждого узла данным процессором осуществляется применением к нему всех правил, шаблону условия которых удовлетворяет данный узел. Вычислительная сложность обработки составляет $O(n*m)$, где n — количество узлов, m — количество правил [5].

Реализация разработанных алгоритмов с использованием SAX-технологии заключается фактически в написании кода, составляющего методы StartTag, StopTag, TagContent. Такой код был реализован путём создания классов на C++, содержащих методы обработки различных сложноструктурированных объектов. Программа читает входной xml-документ порциями около 200 байт и посылает информацию на обработку (***рис. 3***). Обработчик идентифицирует в каждой поступившей порции данных открывающие и закрывающие теги. Для распознанных тегов в блоке идентификации объектов осуществляется поиск начала и завершения конкретного сложноструктурированного объекта. Каждый распознанный объект записывается в кеш-память, после чего вызывается обработчик данного объекта; результат преобразования объекта подаётся на выход.

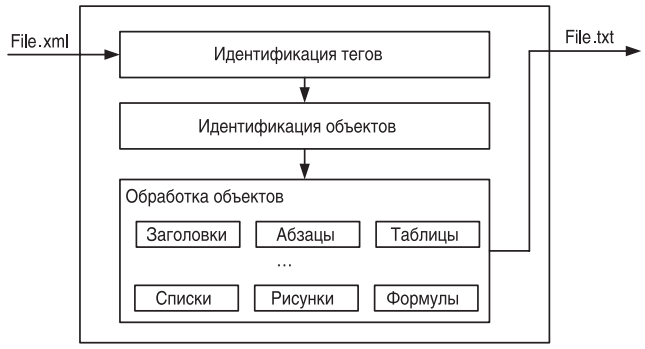

Рис. 3. Схема обработки xml-документа с использованием SAX

Оба обработчика тестировались на нескольких xml-документах. Результаты тестирования представлены в таблице 7, где отражено количество сложноструктурированных объектов, обрабатываемых в соответствии с описанными ранее алгоритмами, а также используемый объём памяти и скорость обработки входных документов на компьютере, имеющем процессор Intel® Core(TM)2 с тактовой частотой 2x1.80 гГц.

Таблица 7

Результаты тестирования программных модулей обработки xml-документов с использованием XSLT- и SAX-технологий

Наименование xml-документа	Объём документа, кБ	Кол-во сложностр. объектов	Объём памяти, мБ (XSLT/ SAX)	Скорость обработки, с (XSLT/ SAX)
Компьютерный синтез и клонирование.xml	1 394	27 131	8 / 1,5	0,6 / 1,33
Финансовые таблицы.xml	20 498	382 541	70 / 1,5	600 / 15

Как видно из таблицы, при применении SAX-технологии требуется значительно меньший объём памяти для обработки xml-документов, чем при применении XSLT-технологии. Этот факт объясняется необходимостью загрузки в память всего документа в программном модуле, использующем XSLT-технологию; при использовании SAX-технологии требуемый объём памяти определяется размером максимального сложноструктурированного объекта, входящего в обрабатываемый документ.

Скорость обработки документов при использовании XSLT-технологии резко уменьшается при увеличении объёма документа и количества сложноструктурированных объектов в нём. При использовании SAX-технологии обработка xml-документов, бо̀льших по объёму и количеству сложноструктурированных объектов, влечёт незначительное увеличение скорости обработки.

Заключение

Предлагаемые в работе алгоритмы программно реализованы и используются в составе системы синтеза речи MultiPhone.

Достоинством алгоритмов является то, что они направлены на преобразование сложноструктурированных объектов в орфографический текст с учётом адекватности смыслового восприятия полученного текста. Предложены несколько режимов обработки сложноструктурированных объектов, классифицированные по степени подробности преобразования.

Программная реализация разработанных алгоритмов осуществлена с использованием двух различных технологий, что позволит использовать программные модули как для мобильных устройств, характеризующихся малым объёмом памяти и низким быстродействием, так и для приложений, существенной характеристикой которых является расширяемость и масштабируемость.

Двухэтапная обработка сложноструктурированных объектов, включающая преобразование к xml-документу на первом этапе и обработку сложноструктурированных объектов в составе xml-документа на втором этапе, в котором реализованы разработанные алгоритмы, позволяет осуществить преобразование сложноструктурированных объектов в других форматах путём добавления модулей формирования xml-документа на основе, например, pdf-файлов.

За рамками данного исследования осталась разработка тестов для оценки адекватности смыслового восприятия сформированного текста и соответствующее тестирование предлагаемых алгоритмов. На решение данных задач будут направлены дальнейшие усилия авторов.

Литература

1. *Лобанов Б.М., Цирульник Л.И.* «Компьютерный синтез и клонирование речи», Минск: Белорусская наука, 2008. — 342 с.
2. http://ru.wikipedia.org/wiki/DOM
3. http://ru.wikipedia.org/wiki/SAX
4. http://www.w3.org/TR/xslt
5. http://www.zdnet.de/security/news/0,39029460,39198860,00.htm

Цирульник Лилия Исааковна —

Окончила факультет прикладной математики и информатики Белорусского государственного университета. Кандидат технических наук, старший научный сотрудник лаборатории распознавания и синтеза речи Объединённого института проблем информатики Национальной академии наук Беларуси, автор более 40 научных работ по проблемам компьютерного синтеза и клонирования речи. Область научных интересов — методы автоматического анализа и синтеза речевых сигналов, человеко-машинные системы речевого общения, речевые компьютерные технологии.

Гецевич Юрий Станиславович —

Окончил факультет прикладной математики и информатики Белорусского государственного университета, факультет математики и информатики университета в Мангейме (Германия). Аспирант, младший научный сотрудник лаборатории распознавания и синтеза речи Объединённого института проблем информатики Национальной академии наук Беларуси. Область научных интересов — методы синтеза белорусской и русской речи по тексту, человеко-машинные системы речевого общения, речевые компьютерные технологии.

Квазиречевой видеонавигатор для слепых

Б.М. Лобанов,
доктор технических наук
О.Г. Сизонов,
соискатель

Описывается структура электронного «видящего и говорящего» устройства-поводыря для слепых, основанного на принципах преобразования видеоизображения в речеподобный сигнал. Приводятся предварительные результаты экспериментального исследования эффективности предложенного метода.

Abctract

A structure of the electronic «seeing and speaking» guide-device for blind people, based on the principles of conversion of a videoimage into a speech-similar signal is described. Preliminary results of an experimental investigation of effectiveness of the proposed method are given.

Введение

Как известно, незрячий человек испытывает огромные трудности при самостоятельном передвижении в пространстве квартиры, улицы, приусадебного участка, города. В настоящее время он использует для навигации специальную трость или сопровождающего (человека или собаку). В данной работе рассматривается возможность создания в помощь слепому принципиально нового электронного «видящего и говорящего» устройства-поводыря на принципах преобразования видеоизображения в речеподобный сигнал.

До недавнего времени задача распознавания зрительных образов не рассматривалась в направлении помощи слепым. Начало было положено в 2001 году, когда в Стенфордском университете США был начат проект под названием «Blind Navigator» («Навигатор для слепых»), который ставит целью создать для слепого человека телеробот-шапку [1]. Голова, шея, туловище человека с такой шапкой направляют взор двух телекамер, закреплённых на голове, в нужную точку пространства впереди идущего. Миниатюрный компьютер, вмонтированный в «шапку», может распознавать около ста часто встречающихся в комнате предметов. После этого синтезатор речи сообщает идущему слепому информацию о том, что за объекты

встречаются на его пути. Точно так же, как курсор мышки компьютера управляется рукой и скользит по экрану, аналогично курсор взора двух телекамер будет управляться головой и скользить по реальному трёхмерному физическому пространству.

Существенные результаты по этому проекту, пригодные для внедрения, до сих пор отсутствуют. Это связано, прежде всего, с огромными трудностями в решении проблемы машинного распознавания трёхмерных зрительных образов и видеосцен.

1. Основная идея и конечная цель

Очевидно, что человеческий мозг справляется с задачей распознавания образов намного лучше любой существующей сегодня искусственной системы. Поэтому, если реализовать механизм адекватного преобразования визуальной информации в акустическую, а когнитивную работу оставить человеку, можно добиться ощутимых результатов и избежать решения проблемы машинного распознавания зрительных образов. Для этого предлагается выделить информативные параметры изображения и преобразовать их в адекватный звуковой сигнал. Для того чтобы способствовать лучшему восприятию и запоминанию звука человеком, принято решение приблизить его к звучанию человеческой речи, т.е. на основе поступающей видеоинформации синтезировать речеподобный сигнал. Здесь можно провести отдалённую аналогию с любым человеческим языком, в том смысле, что реальные объекты описываются набором звуков, который человек может запомнить и распознать.

Основная идея заключается в следующем. Система содержит в своём составе электронную фотокамеру, которая совершает одномоментные снимки пространства впереди незрячего человека. Снимок цветного формата в форме электронного файла поступает для обработки на вход процессора. Из пространственного видеокадра специальная программа формирует последовательный временной ряд сегментов, которые последовательно озвучиваются, образуя псевдослова. Разные сцены впереди незрячего будут давать различные видеокадры, а разные кадры будут порождать для слепого разные псевдослова. В процессе тренировки слепой человек, согласно нашей гипотезе, сможет на слух распознать набор из нескольких сотен типовых псевдослов. После нескольких сеансов обучения с поводырём, двигаясь в привычном пространстве, он сможет узнавать впереди себя знакомые предметы и сцены, ориентируясь на слуховое восприятие псевдослов.

Конечной целью является создание технической системы «Квазиречевой видеонавигатор» для инвалидов по зрению на базе мобильного телефона, оснащённого телекамерой и специальным программным обеспечением для навигации в пространстве. Реализованная на базе мобильного телефона система может оказаться, по нашему мнению, недорогой и простой для тиражирования.

К настоящему времени в лаборатории распознавания и синтеза речи ОИПИ НАНБ накоплен значительный опыт синтеза речевых, в том числе и речеподобных, сигналов [2]. В перспективе, при реализации достаточно

эффективных алгоритмов распознавания и анализа изображений, вместо речеподобного сигнала планируется использовать встроенную в мобильный телефон систему генерации навигационного текста и речи.

2. Алгоритмы реализации

Исходное изображение сцены (***рис. 1***) делится на заданное число «временных срезов». Для преобразования каждого среза в речеподобный сигнал определяется значение усреднённого цвета и контраста. Каждый срез делится на 32 части и определяется значением яркости на каждом участке. Речеподобный сигнал формируется методами спектрально-полосного синтеза путём последовательного во времени сканирования картинки. При этом оси X картинки соответствуют временные отсчёты сигнала, а оси Y — его частотный спектр.

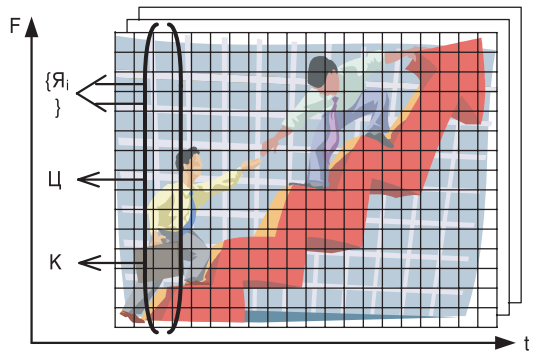

Рис. 1. Исходное изображение сцены

Выходной речеподобный сигнал формируется в соответствии со схемой, представленной на ***рис. 2***. Яркость среза определяет входные значения для используемых цифровых фильтров. Средний цвет — входные значения для генераторов тона, шума. Контраст — соотношение шума и тона.

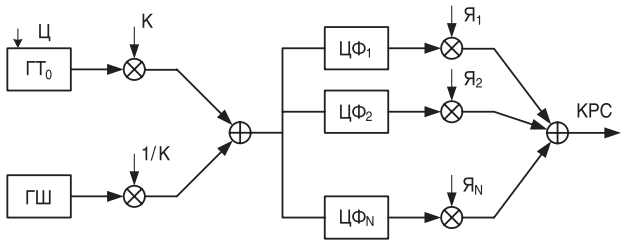

Рис. 2. Преобразователь изображения в речеподобный сигнал

Битовая матрица рисунка приводится к цветовой схеме HSV [3]. Цвет в цветовой схеме HSV изменяется в диапазоне 0÷360°. Яркость изменяется в диапазоне 0÷1. При минимальной яркости сигнал имеет нулевую амплитуду, при максимальной яркости амплитуда приравнивается к единице. Под контрастностью здесь принимается разница между максимальным и минимальным значениями яркости участка. Чем выше средняя контрастность среза — тем сильнее сигнал тона, чем контрастность ниже — тем больше зашумлённость.

На практике цветные снимки дают при синтезе достаточно сложно воспринимаемый звук, поэтому введена возможность приближения картинки к чёрно-белой палитре. Это может быть исполнено двумя способами. Первый из них предлагает вручную задавать значение порогов максимальной и минимальной яркости. При втором способе автоматически вычисляется дисперсия яркости, которая и определяет пороги белого и чёрного.

ГШ — генератор шума.
$ГТ_0$ — генератор основного тона.
ЦФ — цифровой фильтр.
КРС — квазиречевой сигнал.
К — контраст среза.
$\{Я_i\}$ — вектор яркости среза.
Ц — средний цвет среза.

В синтезаторе квазиречевого сигнала используется 32 полосовых фильтра Баттерворда второго порядка с полосами пропускания, определяемыми по шкале Мела в диапазоне 200–4850 Гц. Источником основного тона служит генератор пилообразного сигнала, источником шумового сигнала — генератор белого шума.

Частота основного тона звука задаётся значениями среднего цвета среза:

$$F_0 = \pi/[F_{min} + (F_{max} - F_{min})]^{*}\text{Ц}.$$

Согласно схеме (рис. 2) сигнал основного тона умножается на значение контрастности, а сигнал шума — на обратную величину. Тем самым задаётся соотношение тон/шум в синтезированном сигнале.

3. Программная модель навигатора

По описанным выше алгоритмам разработана программная модель навигатора. Программа позволяет загружать список файлов изображений. При выборе элемента этого списка выводится соответствующее изображение. При этом происходит обработка битовой матрицы, разбиение её по вертикали и горизонтали, синтез квазиречевого сигнала, который можно прослушать и сохранить в файл. Главный интерфейс программы представлен на ***рис. 3***, а интерфейс настроек — на ***рис. 4***.

Для обработки изображений предусмотрены два режима: «ручной» — с возможностью установки порогов чёрного и белого цветов; «автоматический» —

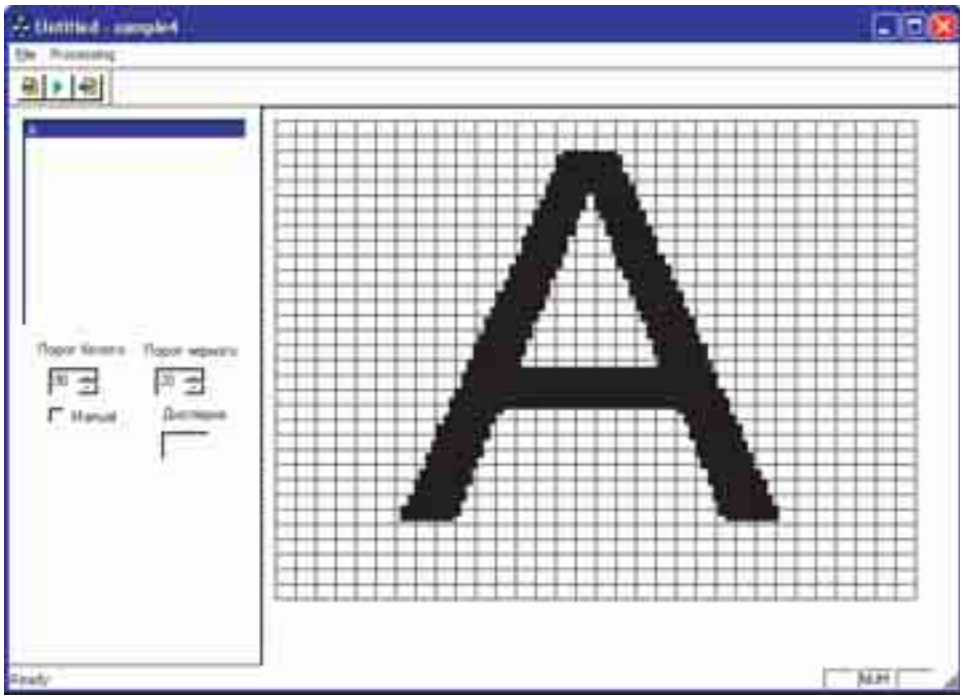

Рис. 3. Главный интерфейс программы

по автоматически вычисленной дисперсии. Для изменения характеристик синтезируемого звукового сигнала используются следующие переменные:

- количество делений изображения по горизонтали — 8, 16, 32, 64;
- Fmax — максимальная частота основного тона в Гц;
- Fmin — минимальная частота основного тона в Гц;
- относительная длительность импульсов сигнала основного тона.

Для проверки качества преобразования изображения в речеподобный сигнал на вход системы подавались различные картинки, а для выходного сигнала рассчитывались соответствующие им динамические спектрограммы (сонограммы). Примеры сонограмм, полученных для изображений цифр, приведены на ***рис. 5***.

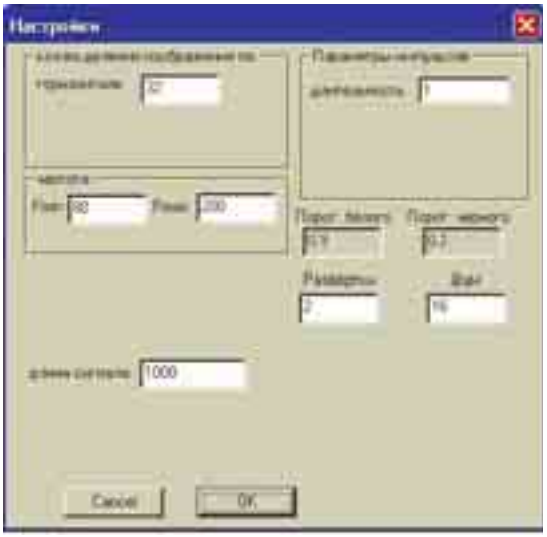

Рис. 4. Интерфейс настроек

Рис. 5. Примеры сонограмм для изображений цифр

Для увеличения информативности звукового сигнала, кроме синтеза по основному направлению (слева направо), применяется изменение направления сверху вниз и справа налево. При этом звуковой сигнал, синтезированный на дополнительных развёртках, последовательно добавляется к основному. Такой подход предоставляет дополнительные пространственные составляющие и обеспечивает большую однозначность звуковым образам. В системе предусмотрена возможность задания длительности развёртки кадра (в мсек.) и числа развёрток (от 1 до 3). На ***рис. 6*** представлены сонограммы сигналов с двойной развёрткой.

Рис. 6. Примеры сонограмм сигналов с двойной развёрткой

4. Экспериментальное исследование метода

Цель исследования — изучить и оценить трудоёмкость обучения звуковым образам изображений, эффективность их слухового распознавания незрячим человеком, определить направления развития описанной модели синтеза в задачах навигации.

Эксперимент 1. Оценка обучаемости и эффективности распознавания на изображениях цифр.

Методика: Испытуемому, тотально слепому человеку К.Л., предоставляется возможность самостоятельно изучить звуковые образы цифр от 0 до 9. Для этого автоматически озвучивается название каждой выбранной пользователем цифры и воспроизводится её синтезированный звуковой образ. Обучаемый сам определяет, когда он закончил обучение, т.е., по его мнению, готов узнавать предъявляемые цифры. Далее компьютерной программой ему предъявляются в случайном порядке звуковые образы, а испытуемый должен назвать, каким цифрам они соответствуют. Общее количество предъявлений — 100.

Результаты эксперимента (матрицы спутывания) представлены в ***таблицах 1 и 2*** для одинарной и двойной развёрток при длительности стимулов 0,5 и 1 сек. соответственно.

Таблица 1

Матрица спутывания для одинарной развёртки

Предъявлено \ Распознано	0	1	2	3	4	5	6	7	8	9
0	8						2			
1		9			1					
2			10							
3				10						
4					10					
5						8			2	
6	3						6		1	
7								10		
8						5			5	
9										10

Общая эффективность узнавания при одной развёртке — 86%, при двойной развёртке — 95%.

Эксперимент 2. Изучение восприятия сходства и различия изображений на фотографиях прямых и пересекающихся пешеходных дорожек.

Таблица 2

Матрица спутывания для двойной развёртки

		Распознано									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Предъявлено	**0**	9						1			
	1		10								
	2			10							
	3				9					1	
	4					10					
	5						9	1			
	6							9			1
	7								10		
	8							1		9	
	9										10

Методика: Испытуемому предъявляются пары синтезированных звуковых образов, полученные на основе фотографий. Задача испытуемого — определить на слух, на каких фотографиях отображены сходные объекты, а на каких — различные.

На ***рис. 7*** представлены примеры фотографий, воспринятых как схожие изображения прямых пешеходных дорожек, а на ***рис. 8*** — пересекающихся.

Рис. 7. Фотографии прямых пешеходных дорожек

Рис. 8. Фотографии пересекающихся пешеходных дорожек

В результате эксперимента испытуемый осуществил на слух правильную классификацию предъявленных фотографий прямых и пересекающихся пешеходных дорожек с надёжностью 93%.

Заключение

Проведённые эксперименты показали достаточно высокий уровень обучаемости слепого человека и способностей к распознаванию звуковых образов предъявляемых изображений. Остаются проблемы обработки изображений для избавления от таких артефактов, как блики, тени и пр.

В целом, однако, предложенный подход можно характеризовать как многообещающий для создания достаточно недорогого и эффективного нового средства навигации слепых в пространстве.

Цель настоящей статьи была бы достигнута в полной мере, если бы она послужила стимулом к привлечению необходимого финансирования для реализации действующей системы на базе мобильного телефона.

Литература

1. Материалы Стэндфордской конференции, 2003 г. // [Электронный ресурс] http://mediax.stanford.edu/news/conference_nov03/dave_grossman.pdf.

2. *Воробьев В.И., Давыдов А.Г., Лобанов Б.М.* Синтез речеподобных сигналов с использованием аллофонов // Сб. трудов XIII сессии Российского акустического общества, М.: «Геос», 2003, с. 110–114.

Лобанов Борис Мефодьевич —

Почётный радист СССР (1981), обладатель серебряной и бронзовой медалей ВДНХ СССР (1983), главного приза международного конкурса фирмы HEWLETT-PACKARD за работу «Распознавание голоса» (1992). С 1987 — член Международного акустического общества, с 1994 — координатор Белорусского отделения Европейской сети по компьютерной лингвистике и речи, с 1995 — член Европейской ассоциации речевых исследований, с 2001 — эксперт Европейской сети языковых технологий. Член докторских советов по защите диссертаций (ОИПИ НАН Беларуси, БГУИР, БГУ, МГЛУ). С 1998 — профессор БГУИР, с 2003 — профессор Университета в Белостоке.

О.Г. Сизонов —

Окончил факультет информационных технологий Белорусского Национального технического университета. Соискатель учёной степени кандидата технических наук. С 2007 года — младший научный сотрудник лаборатории распознавания и синтеза речи Объединённого института проблем информатики Национальной академии наук Беларуси. Область научных интересов — методы лингвистической обработки русского текста в синтезе речи по тексту, синтез и обработка речевых и квазиречевых сигналов, применение синтеза речи в системах реабилитации инвалидов по зрению и слуху.

Криминалистический учёт лиц по фонограммам их речи

С.Е. Лячканов,
соискатель

Описывается новый вид криминалистических учётов — учёт лиц по фонограммам их речи (фоноскопический учёт), созданный в органах внутренних дел Беларуси. Раскрывается актуальность, структурная схема и направления дальнейшего развития фоноучёта.

Abctract

The new kind of criminalistic record — the record of persons by phonograms of their speech (the phonoscopic record) — created in law-enforcement bodies of Belarus is described. The urgency, the block diagramme and directions of further development of the phonorecord are presented.

Введение

Голос и речь человека всегда имели большое криминалистическое значение. По крайней мере, в XVII веке в Англии уже использовалось слуховое распознавание по голосу в качестве доказательства в суде [1, 2]. Особая роль в доказывании отводится криминалистической идентификации — установлению наличия или отсутствия тождества того или иного материального объекта (в данном случае человека) по его отображениям [3], в частности, по его голосу и речи.

В СССР криминалистическая идентификационная экспертиза звукозаписей речи в ходе расследования уголовного дела впервые проведена в 1949 г. [4, 2], а с 1971 г. заключения комплексной криминалистической экспертизы звукозаписей использовались в качестве доказательства в суде [5, 2].

Потребность правоохранительных органов в идентификационных исследованиях по фонограммам речи, научный и технический прогресс обусловили разработку и внедрение в экспертную практику автоматизированной системы «Диалект» [5, 2].

Дальнейшее бурное развитие информационных и речевых технологий в последние десятилетия позволили создавать системы автоматической идентификации по фонограммам речи [2], что обеспечило возможность их использования в целях раскрытия преступлений.

1. Актуальность фоноскопического учёта

В последние годы значительно возросло количество преступлений, связанных с вымогательством, шантажом, рэкетом, телефонным терроризмом, которые, как правило, сопряжены с насилием, опасным для жизни и здоровья людей, или угрозой его применения. В дежурные части органов внутренних дел по телефонным каналам связи поступают анонимные сообщения о минировании вокзалов, аэропортов, промышленных предприятий, учебных заведений, административных зданий, жилых домов и иных мест общественного пользования. Сложно переоценить ущерб государству, связанный с простоем фабрик, заводов, железнодорожного транспорта, вызовом специальных служб и т.д.

В большинстве случаев при совершении указанных преступлений словесная форма является доминирующей. При этом автоматическими системами записи, оперативными работниками или потерпевшими производится фиксирование этой речевой информации на носители записи. Полученные фонограммы телефонных угроз, сообщений о готовящихся террористических актах, вымогательствах и других преступных действиях нередко являются единственным доказательством виновности преступников и источником розыскной информации. При наличии подозреваемого в совершении преступления необходимо проводить идентификационную фоноскопическую экспертизу. В том случае, когда оперативными действиями или иными мерами не удаётся установить личности преступников, возникает необходимость в проведении диагностических исследований фонограмм речи с целью получения информации об их региональной, половозрастной принадлежности, социально-культурном статусе и других индивидуализирующих личность характеристиках. От быстроты и эффективности извлечения этой информации из речевого сигнала порой зависит не только нормальное функционирование предприятий и транспорта, но также жизнь и здоровье людей.

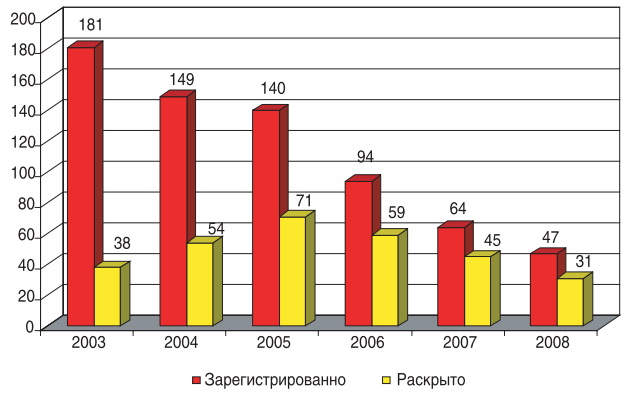

Рис.1. Количество зарегистрированных и раскрытых преступлений по ст. 340 УК РБ «Заведомо ложное сообщение об опасности»

Эффективно бороться с указанными видами преступлений возможно только при наличии базы образцов голоса потенциальных правонарушителей и автоматических систем поиска и диагностики лиц по фонограммам их речи, то есть путём создания и ведения фоноскопического учёта.

Постановка на фоноскопический учёт лиц, совершивших преступления, непосредственно сам отбор образцов их голоса и речи играет огромную профилактическую роль. Так, в Беларуси за последние пять лет, с момента начала постановки лиц на фоноскопический учёт, количество преступлений по ст. 340 УК РБ «Заведомо ложное сообщение об опасности» снизилось в 3,8 раза, при этом раскрываемость данных преступлений возросла с 21 до 66% (***рис. 1***).

2. Создание фоноскопического учёта в Беларуси

В целях раскрытия преступлений в экспертно-криминалистических подразделениях органов внутренних дел Беларуси создаются и ведутся фоноскопические учёты.

Учёт лиц по фонограммам их речи является новым видом криминалистических учётов не только в нашей республике, но и в странах ближнего и дальнего зарубежья. В данной области мы явились своего рода пионерами. Подобной системы фоноскопического учёта, при этом официально и публично созданной по решению правительства, как нам известно, пока нет ни в одной стране мира. Перенимать наш опыт к нам приезжали коллеги из России, а в прошлом году большую заинтересованность в создании системы фоноскопического учёта по нашей схеме проявил Казахстан.

Идея создания системы фоноучёта в нашей стране возникла в 1997 году, когда декретом Президента Республики Беларусь была введена в действие система неотложных мер по борьбе с терроризмом и иными особо опасными насильственными преступлениями. Однако реализовать эту идею без специального нормативного акта было практически невозможно. Поэтому мы решили инициировать подготовку проекта соответствующего постановления правительства. 16 октября 2001 года было принято Постановление Совета министров Республики Беларусь № 1507 «Об утверждении программы создания системы учёта лиц по фонограммам их речи». В начале 2002 года в структуре Государственного экспертно-криминалистического центра МВД Беларуси (ГЭКЦ) создан отдел фоноскопического учёта и диагностики. Для разработки проекта системы фоноскопического учёта была сформирована рабочая группа, в состав которой вошли представители ведущих научных учреждений республики. В ходе разработки проекта были проведены испытания ряда аппаратно-программных средств отечественного и зарубежного производства на предмет возможности их использования в системе фоноскопического учёта. Подготовленный указанной рабочей группой технический проект системы успешно прошёл научно-техническую экспертизу в Национальной академии наук Беларуси.

3. Структурная схема системы фоноскопического учёта

Фоноскопические учёты представляют собой систему регистрации, накопления, классификации, хранения и использования фонограмм и фоноскопической информации. Под фоноскопической информацией здесь понимается информация об особенностях голоса и речи человека и о его личности. Целью создания и использования фоноскопических учётов является установление и идентификация по фонограммам речи лиц, совершивших преступления. Объектами фоноскопических учётов являются фонограммы голоса и речи лиц, подозреваемых в преступлениях либо совершивших преступления, а также неустановленных преступников. На ***рис. 2*** приведена структурная схема системы фоноскопического учёта.

Система фоноскопического учёта является трёхуровневой (районный, областной и республиканский уровни) и состоит из фонотек образцов речи, автоматизированных банков фонограмм (АБФ) и автоматизированной информационно-поисковой системы (АИПС).

В экспертных подразделениях районного звена организуются фонотеки, куда заносятся образцы речи лиц, поставленных на фоноскопический учёт (учётные образцы). Далее информация стекается в экспертно-криминалистические подразделения областного уровня в автоматизированные банки фонограмм (АБФ). В АБФ, помимо учётных образцов установлен-

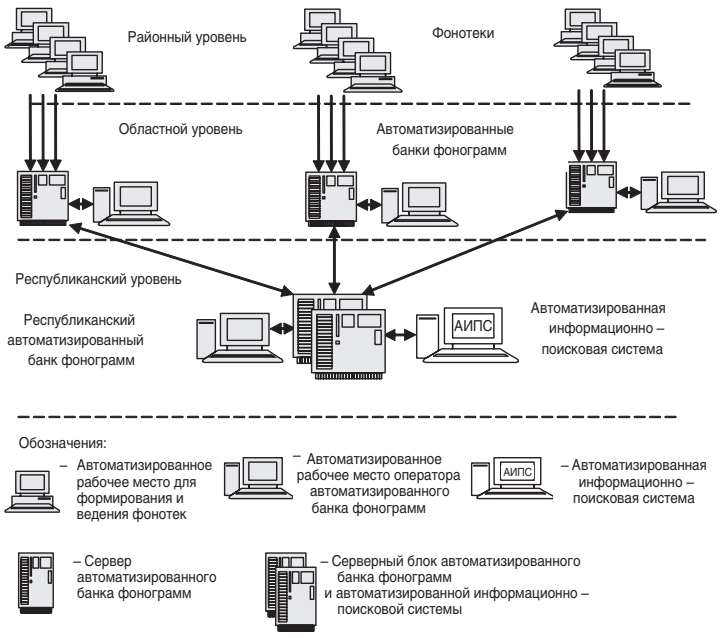

Рис. 2. Структурная схема системы фоноскопического учёта

Фото 1. АРМ для формирования и ведения фонотек

ных лиц, заносятся фонограммы голоса и речи неустановленных лиц, совершивших преступления. В основном, это анонимные телефонные сообщения о минировании, с угрозами применения насилия, вымогательством и т.п. Далее массивы фонограмм и регистрационных данных передаются в ГЭКЦ, где формируется республиканский АБФ и функционирует автоматизированная информационно-поисковая система, которая осуществляет поиск в массиве голосов по акустическим признакам.

3.1. АРМ для формирования и ведения фонотек

На районном уровне функционируют автоматизированные рабочие места (АРМ) для формирования и ведения фонотек (***фото 1***), на которых осуществляется постановка лиц на фоноскопический учёт.

В рамках договора «О научно-техническом сотрудничестве» между ГЭКЦ и Объединённым институтом проблем информатики Национальной академии наук Республики Беларусь разработано специализированное программное обеспечение для ведения фонотек и автоматизированных банков фонограмм, которое позволяет осуществлять отбор образцов голоса и речи с автоматическим контролем длительности и качества записываемого сигнала.

При постановке лица на фоноскопический учёт заполняется карточка учётного лица (***фото 2***), куда заносятся регистрационные данные.

Отбор образцов голоса и речи учётного лица проходит полностью в автоматическом режиме, то есть участие эксперта в процессе записи фонограммы, по большому счёту, не требуется. Сначала программа предлагает лицу, ставящемуся на учёт, перечень наводящих вопросов для изложения в свободной форме,

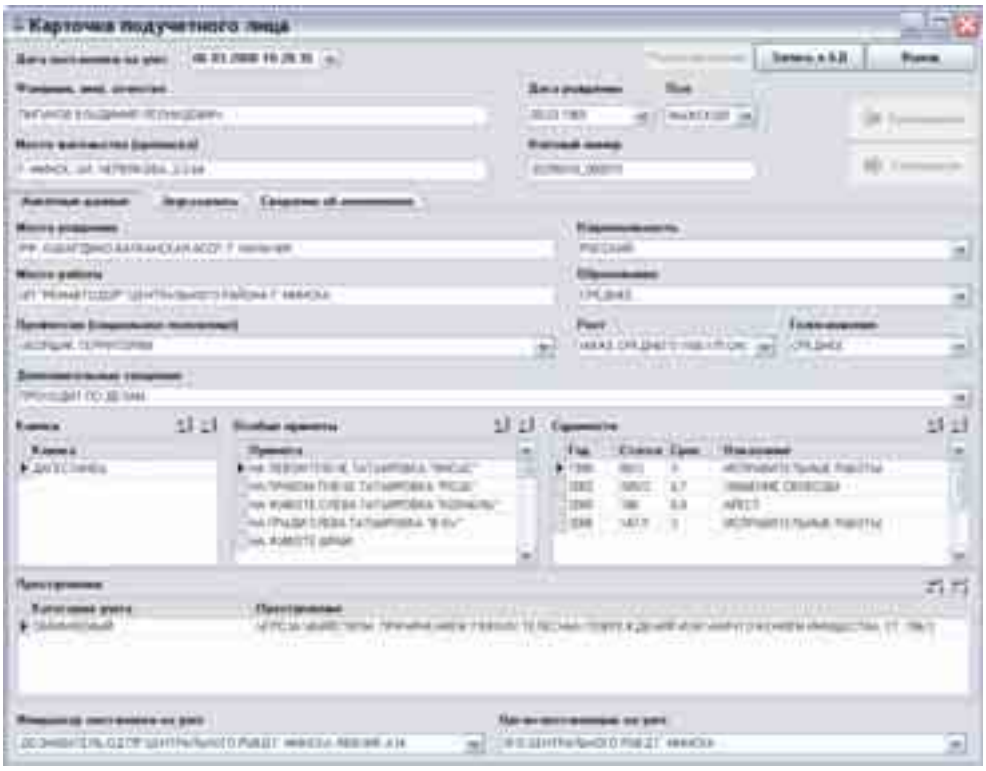

Фото 2. Карточка учётного лица

например, автобиографии, памятных событий либо рассказа на произвольную тему (***фото 3***). Программа контролирует общую длительность произнесённых речевых сигналов. Когда длительность достигает пяти минут, начинается второй этап записи — чтение специального фонетически сбалансированного текста (***фото 4***). Для не умеющих читать лиц, у которых отбираются образцы, предусмотрено программное озвучивание предлагаемого для чтения текста. Лицо, чьи образцы голоса записываются, должен повторить каждую предлагаемую фразу, причём программа автоматически контролирует громкость и качество произношения.

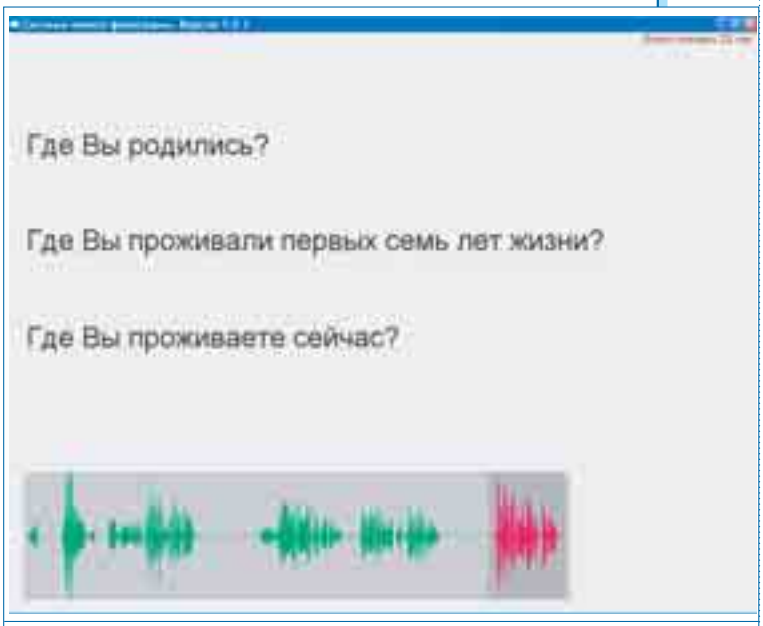

Фото 3. Первый этап отбора образцов речи — ответы на наводящие вопросы

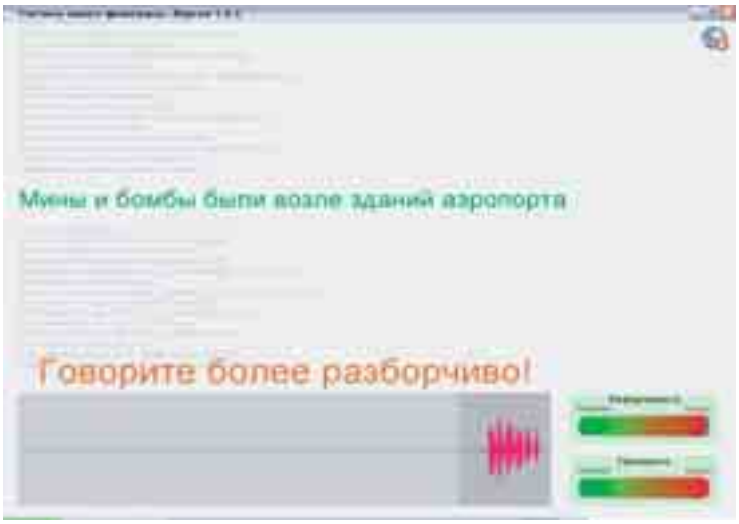

Фото 4. Второй этап отбора образцов речи — чтение специального фонетически сбалансированного текста

Фонотеки используются сотрудниками органов внутренних дел также в целях:
- предоставления потерпевшим (свидетелям) образцов устной речи лиц, поставленных на учёт в фонотеки, для прослушивания и опознания правонарушителей по голосу;
- использования учётных образцов устной речи для проведения фоноскопических экспертиз в случаях невозможности изъятия экспериментальных образцов речи;
- поиска лиц по регистрационным данным карточки учётного лица.

3.2. Автоматизированное рабочее место оператора АБФ

Автоматизированные рабочие места оператора АБФ функционируют в экспертно-криминалистических подразделениях областного звена и предназначены для систематизации и хранения информации, поступающей с районного уровня. АБФ формируются из фонотек районного звена и фонограмм речи неустановленных преступников, поступающих из оперативных подразделений. В свою очередь, республиканский банк фонограмм формируется из АБФ областного звена.

Программное обеспечение для формирования АФБ разработано на основе ПО для ведения фонотек и состоит из серверной и клиентской частей. Программное обеспечение позволяет производить отбор образцов речи, вести учёт как установленных, так и неустановленных лиц, осуществлять поиск лиц по регистрационным данным, производить автоматизированный экспорт и импорт информации, поступающей из фонотек и оперативных подразделений.

3.3. Автоматизированная информационно-поисковая система

Далее информация поступает в автоматизированную информационно-поисковую систему фоноскопического учёта (АИПС), которая функционирует на базе программного обеспечения «Трал», разработанного Центром речевых технологий (Россия, г. С.-Петербург), и позволяет систематизировать и хранить учётную информацию, а также осуществлять поиск лиц по регистрационным данным и акустическим признакам их голосов (***фото 5***).

Основными задачами использования АИПС являются:
- установление принадлежности анонимного сообщения одному из известных лиц, состоящих на фоноучёте;
- установление принадлежности голоса и речи известного лица одному из состоящих на учёте неизвестных лиц;
- установление принадлежности анонимного сообщения одному из состоящих на учёте неизвестных лиц.

После осуществления поиска по акустическим признакам система «Трал» выдает список лиц, голоса которых наиболее близки к проверяемой фонограмме по своим характеристикам.

Для точного установления нахождения (либо отсутствия) в АБФ проверяемого лица проводятся идентификационные исследования методами аддитивного, лингвистического и инструментального анализа. При этом лица из списка, выданного системой «Трал», попарно сравниваются с проверяемым лицом. Для проведения указанных исследований в ГЭКЦ используются следующие аппаратно-программные средства:

- компьютерная речевая лаборатория CSL 4500 (США);
- аппаратно-программный комплекс «MEDAV» (ФРГ) на базе рабочей станции «SGI O2»;
- аппаратно-программный комплекс криминалистического исследования фонограмм «ИКАР-Лаб II» (ЦРТ, г. С.-Петербург).
- программное обеспечение для криминалистической идентификации по фонограммам устной речи «Фонэкси» (***фото 6***) производства ООО «Целевые технологии» (Россия, г. Москва).

Фото 5. Вид программного обеспечения «Трал»

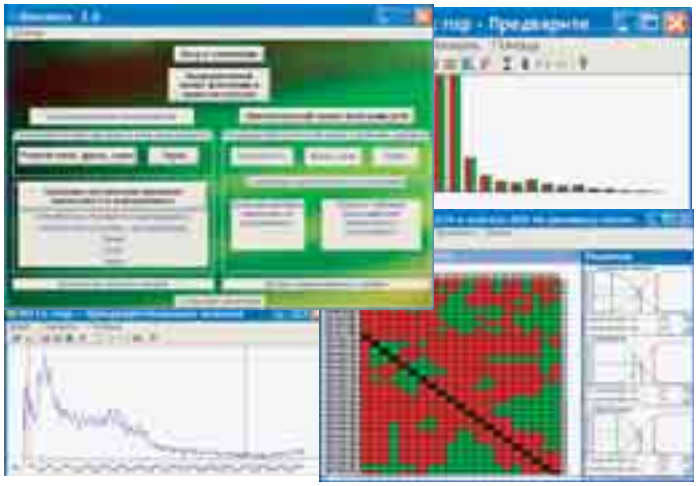

Фото 6. Вид программного обеспечения «Фонэкси»

4. Направления и перспективы развития фоноскопического учёта

В настоящее время фоноскопические учёты находятся на стадии активного развития, совершенствуются их научно-методические аспекты и аппаратно-программные средства. По предложениям и замечаниям экспертно-криминалистических подразделений, возникающим в практике формирования и ведения фоноучёта, постоянно ведётся работа по его совершенствованию. Эксперты ГЭКЦ активно взаимодействуют с научными учреждениями республики, разработчиками программных продуктов, используемых в системе фоноскопического учёта, с целью дальнейшего развития и повышения эффективности её использования. Постоянно дорабатывается и совершенствуется программное обеспечение, внедряются дополнительные методы идентификации, ведутся разработки новых программных средств.

Основными направлениями развития системы фоноскопического учёта с целью повышения её эффективности являются:

- повышение качества записи речевой информации, поступающей по телефонным каналам связи, а также получаемой в ходе оперативно-розыскных мероприятий;
- улучшение качества постановки лиц на фоноскопический учёт;
- повышение точности автоматической идентификации лиц по фонограммам их речи;
- уменьшение минимальной длительности фонограмм при обеспечении высокой эффективности поиска;
- повышение эффективности поиска фонограмм с низким отношением сигнал/шум;
- уменьшение влияния каналов записи фонограмм на эффективность поиска;
- постоянное увеличение объёма автоматизированного банка фонограмм;
- сокращение времени поиска фонограмм в больших массивах;
- разработка и внедрение в АИПС автоматической системы установления личностных характеристик по голосу и речи неизвестного лица.

В перспективе внедрение системы установления личностных характеристик по голосу и речи позволит не только оперативно получать розыскную информацию, но и обеспечит возможность сужать круг поиска фонограмм, осуществлять его по выборке, соответствующей установленным личностным характеристикам.

Заключение

Указанный выше перечень направлений развития системы фоноскопического учёта не является исчерпывающим, это только те основные направления, которые явно усматриваются на современном этапе развития системы фоноскопического учёта. При этом каждое из этих направлений по своим масштабам, разнообразию проблем и путей их решения достойно отдельного, более подробного, изучения и рассмотрения.

Литература

1. *Hollien H.* The Acoustics of Crime. The New Science of Forensic Phonetics. — New York: Plenum Press, A Division of Plenum Publishing Corporation, 1990.
2. *Каганов А.Ш., Майлис Н.П., Михайлов В.Г., Брызгунова Е.А., Коваль С.Л., Портнова Т.Е., Зубов Г.Н., Столбов М.Б., Кринов С.Н., Байчаров Н.В., Кураченкова Н.Б., Линьков А.Н., Попов Н.Ф., Никонов А.В.* Современные методы, технические и программные средства, используемые в криминалистической экспертизе звукозаписей: Методическое пособие для экспертов / Под ред. Гуржей Т.И. — М.: ГУ РФЦСЭ, 2003.
3. *Белкин Р.С. и др.* Криминалистика. М.: Юридическая литература. 1968. — 695 с.
4. *Копелев Л.* Утоли моя печали. Мемуары. — М.: Слово, 1991.
5. *Попов Н.Ф., Линьков А.Н., Кураченкова Н.Б., Байчаров Н.В.* Идентификация лиц по фонограммам русской речи на автоматизированной системе «Диалект»: Пособие для экспертов / Под ред. А.В. Фесенко. — М.: В/ч 34435, 1996.

Лячканов Сергей Евгеньевич —

Начальник шестого управления Государственного экспертно-криминалистического центра Министерства внутренних дел Республики Беларусь, полковник милиции. Радиоинженер, юрист. С 1997 года занимается экспертной деятельностью в области фоноскопии. Соискатель учёной степени кандидата технических наук в ОИПИ НАН Беларуси.

Технология VoiceXML и её приложения

Е.Б. Никитин,
аспирант

VoiceXML является мощной и гибкой веб-технологией в области синтеза и распознавания речи. В данной статье описывается история создания, преимущества и способы применения стандартного XML-формата для описания интерактивных диалогов между человеком и компьютером VoiceXML.

Abctract

VoiceXML is a powerful and agile web-technology in the text-to-speech and voice recognition field. The history of its creation, advantages and applications are described in the article. VoiceXML stands for a standard XML format for specifying interactive voice dialogues between a human and a computer.

Введение

VoiceXML — это открытый стандартный язык разметки для голосовых приложений, разработанный организацией VoiceXML Forum [1], которая была создана в 1999 году и объединяет 44 фирмы, среди которых Ericsson, Siemens, France Telecom, Novell, Samsung Electronics, Sun Microsystems, IBM, AT&T, Lucent Technologies и Motorola [2–11]. VoiceXML предназначен для создания и передачи ориентированных на Web персонализированных интерактивных сервисов с речевым ответом и для обеспечения телефонного и речевого доступа к интегрированным базам данных центров обработки вызовов (call center databases), а также к информации на Web-узлах и в интрасетях.

История VoiceXML началась в 1995 году, когда работники AT&T Research Dave Ladd, Chris Ramming, Ken Rehor и Curt Tuckey [12] собирались в неформальной обстановке и обсуждали, как Интернет повлияет на телефонию. Они решили, что нужно создать систему, которая посредством голосового браузера позволяла бы предоставлять контент и услуги пользователям обычных телефонов. Для этого был создан проект AT&T Phone Web project. Потом Lucent откололся от AT&T, и каждая фирма начала разрабатывать свой собственный стандарт. Chris Ramming остался работать в AT&T, Ken Rehor оказался в Lucent, а Dave Ladd и Curt Tuckey перешли в Motorola. К началу 1999 года AT&T и Lucent имели несовместимые разновидности языка разметки звонков (Phone Markup Language — PML), у Motorola был новый стандарт VoxML, другие компании также экспериментировали с идеями интеграции контента и сервисов Интернет в телефонию. Для функционирования голосовых веб-сервисов

требовалась разработка стандартного языка. Создатели Phone Web поспособствовали тому, что AT&T, Lucent и Motorola основали VoiceXML Forum. Вскоре к ним присоединилась IBM. К августу 1999 небольшая команда инженеров Форума выпустила VoiceXML 0.9, в котором были объединены лучшие особенности предыдущих языков, а также добавлены новые идеи — особенно важным было добавление тонального набора (Dual-Tone Multi-Frequency, DTMF) [13]. После этого сообщество расширилось, язык был значительно дополнен новыми функциями — и в марте 2000 года был обнародован VoiceXML 1.0. На следующий день появилось около двадцати различных приложений этого языка.

Вскоре после выхода VoiceXML 1.0 Форум вступил в World Wide Web Consortium (W3C) [14], благодаря чему выпускались промежуточные версии VoiceXML 2.0, которые в марте 2004 дошли до финального этапа создания рекомендаций. Рекомендация VoiceXML 2.1 увидела свет 19 июня 2007 года [15]. Это последняя версия на данный момент.

1. Процесс синтеза речи

Синтез речи производится системой TTS (Text-To-Speech), которая принимает SSML-документы (Speech Synthesis Markup Language) [16], а на выходе даёт голосовой поток.

Процесс переработки XML-файла в голосовое сообщение состоит из шести шагов.

1) Разбор XML: процесс разбора XML используется для извлечения содержания и дерева документа из входящего текстового файла. Структура, тэги и атрибуты, полученные на этом шаге, влияют на все последующие. Словами в SSML не могут быть тэги разметки: например, последовательность «black<break/>berry» обработчик воспримет как два слова: «black» и «berry», — а не как одно слово с паузой в середине. Разбиение на более мелкие элементы может изменить интерпретацию обработчиком.

2) Анализ структуры: от структуры документа зависит, как он будет прочитан. Например, существуют речевые шаблоны, связанные с абзацами и предложениями.

- *С поддержкой разметки:* P- и S-элементы (P — paragraph, абзац; S — sentence, предложение), определённые в SSML, однозначно задают структуру документа.

- *Без разметки:* если эти элементы не использованы в документе или его частях, обработчик сам распознаёт структуру, анализируя текст, часто опираясь на пунктуацию и другие особенности языка.

3) Нормализация текста: в языках существуют специальные конструкции, которые требуют перевода текста из письменной (орфографической) формы в устную. Нормализация текста — это автоматический процесс, который производится обработчиком синтеза. Например, в русском языке текст «$200» произносится как «двести долларов». Аналогично, «1/2» может произноситься как «одна вторая», «половина», «первое февраля» и т.д.

После этого шага текст для перевода в устную речь полностью преобразуется в лексемы. Что стоит за лексемой, зависит от конкретного языка. Так, в английском языке лексемы обычно разделены пробелом и являются словами.

- *С поддержкой разметки:* элемент **say-as** может использоваться для однозначного задания типа конструкций и разрешения неопределённостей.

- *Без разметки:* не размеченный элементами **say-as** текст разбирается автоматически, однако при этом следует ожидать больших трудностей из-за неопределённостей наподобие описанного выше примера с «1/2».

4) Перевод текста в фонемы: после того, как обработчик синтеза определил набор слов, которые должны быть сказаны, необходимо определить произношение каждого слова. Произношение слов может быть описано как последовательность фонем, которые являются звуковыми элементами языка. Например, в английском слово «read» может произноситься как «рэд» или как «рид», в зависимости от временной формы глагола.

- *С поддержкой разметки:* элемент **phoneme** позволяет задавать фонемные последовательности для любого слова, что предоставляет полный контроль над произношением. Также можно использовать элемент **say-as**, чтобы текст считался именем собственным, что позволит обработчику применить специальные правила для определения произношения. Эти элементы являются особенно полезными для акронимов и аббревиатур.

- *Без разметки:* в случае отсутствия элемента phoneme обработчик применяет автоматические функции по определению произношения. Это производится за счёт поиска слов в словаре и применения правил для определения других произношений. Обработчики синтеза речи разработаны так, чтобы производить перевод текста в фонемы, так что большинство слов могут быть обработаны автоматически. Альтернативным способом является изменение текста перед отправкой его обработчику таким образом, чтобы избежать многозначности, например, заменить слово «read» на «reed». При этом надо учитывать, что такой текст лучше нигде не отображать, потому что он грамматически некорректен.

5) Просодический анализ: просодема — это набор особенностей речи, которые включают в себя интонацию, ритм, паузы, скорость речи, ударения и другие. Воспроизведение просодем, близких к естественному языку, является важным, чтобы речь звучала естественно и корректно.

- *С поддержкой разметки:* элементы **emphasis**, **brake** и **prosody** могут быть использованы создателем документа для указания обработчику правильных просодем.

- *Без разметки:* в случае отсутствия этих элементов обработчик сам подберёт необходимые просодемы. Это достигается за счёт анализа структуры документа, синтаксиса и другой информации.

6) Генерация сигнала: фонемы и просодемы используются обработчиком для генерации аудиосигнала. Существует много разных подходов к этому шагу.

- *С поддержкой разметки:* элемент **voice** позволяет создателю документа запрашивать особенный тип голоса (например, голос взрослого мужчины). Элемент audio дает возможность вставлять предварительно записанные аудиофрагменты в выходной поток.

2. Процесс распознавания речи

Распознавание речи осуществляется системой STT (Speech-To-Text). Из-за большого количества возможных произношений и акцентов STT-процесс является зачастую сложной и неоднозначной процедурой. Для упрощения и повышения надёжности работы системы STT применяются грамматики. Грамматика распознавания речи — это набор паттернов, которые подсказывают системе распознавания речи, что сейчас может сказать пользователь. Например, если вы звоните в автоматический справочник, система спросит имя человека, с которым вас нужно соединить. После этого система запустит распознаватель речи, предоставив ему грамматику распознавания речи. Эта грамматика содержит имена людей в справочнике и наиболее частые варианты ответов звонящих. Speech Recognition Grammar Specification (SRGS — спецификация грамматики распознавания речи) [17] является стандартом W3C для описания грамматик распознавания речи.

SRGS определяет два альтернативных, но логически эквивалентных синтаксиса: один основан на XML, другой использует расширенный формат BNF (Backus-Naur Form) [18]. На практике чаще используется XML.

Если бы система распознавания речи возвращала просто строку со словами, сказанными пользователем, голосовому приложению пришлось бы выполнять большую работу по извлечению семантических значений этих слов. Поэтому SRGS-грамматики могут быть оформлены тэгами, в случае использования которых строится семантический результат. SRGS не описывает содержание ярлыков, потому что это сделано в дополнительном стандарте SISR (Semantic Interpretation for Speech Recognition) [19]. VoiceXML тесно связан со стандартами SRGS и SISR.

Пользовательский агент — это обработчик грамматики, который на входе принимает речь и сопоставляет её с грамматикой, а на выходе даёт соответствующий результат распознавания.

Распознаватель речи — это пользовательский агент, использующий следующие входные и выходные параметры:

- *Вход A*: грамматика или множество грамматик. Эти грамматики информируют распознаватель о словах и паттернах слов, которые надо слушать.
- *Вход B*: аудиопоток, который может содержать речь, соответствующую грамматикам.
- *Выход*: описание результатов, которое содержит детальную информацию о распознанном речевом входе. Формат и детали результата являются свободными. Для информативности большинство систем распознавания включают в результаты хотя бы транскрипцию слов, которые были распознаны. Также голосовому браузеру могут быть переданы ошибки и другая информация о работе.

3. Особенности разработки под VoiceXML

Как говорилось ранее, VoiceXML является приложением языка XML и имеет структуру дерева, по которому пользователь может перемещаться с помощью голосовых команд. Первые VoiceXML-платформы строго требовали, чтобы определение типа документа (Document Type Definition — DTD) объявлялось непосредственно перед начальным тэгом <vxml>. Современные VoiceXML-браузеры намного более гибки в этой части, так что теперь нет необходимости

в дополнительном объявлении DTD. Более того, для большей гибкости приложений рекомендуется вообще нигде в коде не указывать DTD. Неотъемлемым компонентом любого VoiceXML-приложения является система распознавания и синтеза речи, установленная на сервере. Доступны подобные продукты от таких известных фирм, как IBM [20], Microsoft [21], Speechtech [22], Voxeo [23]. Классификация тэгов VoiceXML (таблица 1) проста и логична.

Таблица 1

Типы тэгов VoiceXML

Назначение	Тэг	Назначение	Тэг
Определение приложения	<vxml>	Определение грамматик	<grammar>
	<meta>		<rule>
Диалоги	<form>		<ruleref>
	<menu>		<token>
Элементы ввода формы	<field>		<one-of>
	<record>		<item>
	<subdialog>		<tag>
	<transfer>		<example>
Элементы контроля формы	<initial>		<lexicon>
	<block>	Контроль интерпретатора	<property>
Меню	<menu>	Контейнеры	<block>
	<choice>		<filled>
	<grammar>		<if>
	<enumerate>		<catch>
Поля	<field>		<error>
	<enumerate>		<help>
	<grammar>		<noinput>
	<option>		<nomatch>
	<filled>		<object>
	<help>	Объявление и установка переменных	<var>
	<noinput>		<assign>
	<nomatch>		<clear>
Субдиалоги	<subdialog>	Процедурная логика	<if>
	<param>		<else>
	<return>		<elseif>
Контроль перехода в диалогах	<goto>	Логика сценариев	<script>
	<submit>	Создание аудиовыхода	<audio>
	<link>		<prompt>
	<choice>		<reprompt>
Контроль синтеза речи	<break>		<value>
	<voice>		<enumerate>
	<emphasis>	События	<throw>
	<prosody>	Прерывание сессии	<disconnect>
	<phoneme>		<exit>
	<say-as>	Отправка и получение данных	<submit>
	<sub>	Исправление ошибок	<log>
	<mark>	Обработка событий	<catch>
	<s>		<error>
	<sentence>		<help>
	<p>		<noinput>
	<paragraph>		<nomatch>

Таблица 2

Описание тэгов VoiceXML

Тэг	Описание
<assign>	Устанавливает значение переменной
<audio>	Проигрывает аудиозапись пользователю
<block>	Содержит исполняемый код (не интерактивный)
<break>	Элемент SSML, вставляет паузу в выходной аудиопоток
<catch>	Перехватывает событие
<choice>	Определяет элемент меню
<clear>	Очищает одну или несколько переменных формы
<disconnect>	Прекращает телефонную сессию
<else>	Отмечает начало ИНАЧЕ-части внутри элемента <if>
<elseif>	Отмечает начало ИНАЧЕ_ЕСЛИ-части внутри элемента <if>
<emphasis>	Элемент SSML, меняет ударения в речевом выходе
<enumerate>	Генерирует аудиовыход, в котором перечисляются все пункты меню
<error>	Перехватывает ошибки
<example>	Фраза-пример, которая соответствует правилам грамматики
<exit>	Выход из сессии
<field>	Объявляет поле ввода в форме
<filled>	Содержит действия, которые должны выполниться после заполнения формы
<form>	Предоставляет информацию и собирает данные
<goto>	Переход
<grammar>	Определяет грамматику распознавания речи
<help>	Помощь
<if>	Оператор условия
<initial>	Объявляет начальную логику в момент входа в форму
<item>	Элемент XML-грамматики входа, указывает опциональную или повторяющуюся информацию от пользователя
<lexicon>	Элемент XML-грамматики входа, указывает источник информации о произношении
<link>	Указывает на общий для всех диалогов переход
<log>	Записывает информацию по отладке в журнал звонка
<menu>	Позволяет пользователю выбрать один из вариантов
<meta>	Определяет метаданные парой имя-значение
<noinput>	Перехватывает события отсутствия ввода
<nomatch>	Перехватывает события несовпадения
<object>	Всегда выбрасывает исключение о неподдерживаемом объекте
<one-of>	Элемент XML-грамматики входа, указывает на альтернативные данные от пользователя
<option>	Определяет опцию в <field>
<p>, <paragraph>	Элемент SSML, определяет параграфы в тексте
<param>	Определяет параметры в элементе <subdialog>
<phoneme>	Элемент SSML, определяет фонетическое произношение
<prompt>	Отправляет текст в очередь на генерацию звукового выхода
<prosody>	Элемент SSML, меняет речевой выход
<record>	Записывает аудиофрагмент
<reprompt>	Повторяет запрос в случае повторного посещения поля после события
<return>	Возвращает из поддиалога
<rule>	Элемент XML-грамматики входа, определяет правило грамматики

<ruleref>	Элемент XML-грамматики входа, определяет отношение к другому правилу грамматики
<s>, <sentence>	Элемент SSML, определяет часть текста как предложение
<say-as>	Элемент SSML, определяет, как должно быть произнесено слово
<script>	Указывает блок Javascript в клиентской части
<speak>	Ключевой элемент для отдельного SSML-документа
<sub>	Атрибуты этого тэга предоставляют альтернативный текст
<subdialog>	Обращение к другому диалогу как поддиалогу текущего
<submit>	Отправляет значения на сервер
<tag>	Элемент XML-грамматики входа, указывает, как интерпретировать ввод пользователя
<throw>	Создаёт событие
<token>	Элемент XML-грамматики входа, указывает, какие слова будут произнесены пользователем
<transfer>	Переадресует звонок
<value>	Вставляет значение выражения в выходной аудиопоток
<var>	Объявляет переменные
<voice>	Элемент SSML, запрашивает изменения в голосе
<vxml>	Содержит VoiceXML-код

Рассмотрим классический пример базовой программы. Сначала необходимо указать стандартный заголовок, с которого начинаются все документы VoiceXML:

```
<?xml version=«1.0» encoding=«UTF-8»?>
```

Затем необходимо указать версию конкретного документа. В случае если вы хотите использовать функциональность самой новой версии 2.1 (например, элемент <data>), надо указать конкретную версию вашей программы: «2.1». Таким образом, получаем код:

```
<?xml version=«1.0» encoding=«UTF-8»?>
<vxml version = «2.1» >
  <form>
    <block>
    <prompt>
      Здравствуй, мир!
    </prompt>
    </block>
  </form>
</vxml>
```

Обратите внимание, что даже самый простой сценарий требует структуры кода. Тэг <form> отличается от аналогичного в HTML тем, что группирует секции ввода и вывода на вашей странице (необходимо помнить, что это «веб-страница» для телефона). В более сложных сценариях этот тэг может являться поименованной секцией, к которой можно перейти. Аналогично, <block> в нашем примере кажется чем-то неважным, потому что мы выполняем лишь одну функцию. Тем не менее, VoiceXML требует структуру внутри <form>, что и обеспечивает нам тэг <block>. Но как же VoiceXML сможет конвертировать «Здравствуй, мир»? Никаких дополнительных тэгов не надо, VoiceXML воспринимает любое значение, не ограниченное тэгами, как текст, который должен быть прочитан по телефону.

В VoiceXML определены два типа диалогов, которые позволяют взаимодействовать приложению с пользователем: формы и меню. Форма используется для представления или получения информации от пользователя. Меню является специализированной формой, которая требует сделать определённый выбор. Элемент **<form>** включает в себя директиву **<goto>**, которая говорит приложению, куда двигаться, в зависимости от пользовательского входа. Меню

использует элементы **<choice>** для определения дальнейшего пути на основе выбора пользователя. Следующий пример предлагает пользователю сделать выбор и вызывает различные файлы в зависимости от выбора.

```
<?xml version=«1.0» encoding=«UTF-8»?>
    <vxml version = «2.1»>
    <form id=«F1» scope=«document»>
     <catch event=«Event_1 Event_2»>
     <prompt>
       <value expr=«_message»/>
     </prompt>
     </catch>
     <menu id=«menu»>
     <prompt>
    Что такое НАНБ?
       Если вы думаете, что это Национальная академия наук Беларуси,
    скажите слово «академия» или нажмите «один».
      Если вы думаете, что это фрукт, скажите слово «фрукт» или нажмите «два».
     </prompt>
     <choice event=«Event_1» accept=«approximate»
          dtmf=«1»
          message=«Верно!»>
           наверное, это академия
     </choice>
     <choice event=«Event_2» accept=«exact»
          dtmf=«2»
          message=«Неправильно»>
           фрукт
     </choice>
     </menu>
    </form>
    </vxml>
```

Кроме прочего, можно обратить внимание на параметр accept тэга <choice>: у него есть значения exact и approximate. Exact требует полного совпадения, а approximate допускает совпадение даже по одному слову, так что в нашем примере, распознав слово Academy, программа признает выполненным условие «most definitely intended to be an Academy».

В данном примере были использованы всего семь элементов:

```
<vxml>
<form>
<catch>
<prompt>
<value>
<menu>
<choice>
```

Конечно, здесь не были описаны все возможности VoiceXML, но приведённые примеры могут помочь составить представление о том, как голосовые команды обрабатываются сервером. Конечно, такой тип приложений должен

учитывать масштабируемость и нагрузку на сервер, а традиционные серверные задания (доступ к базе данных, система сообщений, запросы и пр.) должны разрабатываться в виде строк выбора.

4. Возможности применения

Существует много уже реализованных применений VoiceXML.

1) Интерактивный голосовой ответ (IVR) и расширенное самообслуживание. Такие решения автоматизируют звонки, что позволяет повысить качество обслуживания, снизить затраты, сократить сроки возврата инвестиций и продвигать новые прибыльные возможности.

2) Контактный центр. Системы управления взаимоотношения с клиентами (CRM) — это клиенто-ориентированные стратегии бизнеса, целью которых является оптимизация прибыльности, оборотных средств и удовлетворения потребностей клиентов. Хотя цели во всех организациях одинаковы, реализация конкретных решений всегда уникальна.

3) Унифицированные коммуникации. К таким системам относится широкий спектр приложений, разработанных для преодоления барьера между разными типами связи: телефон, электронная почта, чат, веб и др. Многие компании инвестируют в развитие таких технологий для дальнейшего их внедрения в CRM-приложения. Наличие интегрированной в CRM-приложение кнопки «нажмите для звонка» предотвращает ошибки ручного набора и делает работу агентов более продуктивной. Такое приложение с интегрированными телефонными коммуникациями считывает номер, с которого производится звонок, и проверяет CRM-систему на совпадения. В случае если номер принадлежит клиенту, у агента открываются соответствующие записи. Это позволяет агентам приветствовать звонящих по имени и более продуктивно вести диалог, что приводит к сокращению времени разговора. С такими решениями среднее время каждого звонка снижается на 20–60 секунд.

4) Уведомления и напоминания. Автоматические мультимодальные системы уведомлений гарантируют, что клиенты, работники и другие заинтересованные стороны получат срочные сообщения, даже если их не ожидают. Уведомления и напоминания — развивающийся важный бизнес, который повышает лояльность клиентов и открывает возможности для увеличения прибылей. VoiceXML предоставляет следующие функциональные возможности:

a) обработка входящих и исходящих звонков;
b) отправка персонализированных сообщений на основе систем синтеза речи;
c) конференц-связь;
d) возврат звонка;
e) факс;
f) динамические меню;
g) наблюдение и составление отчётов;
h) облегчённая интеграция с вебом и электронной почтой.

Заключение

На данный момент VoiceXML является общепринятым языком разметки для голосовых приложений. Этому способствуют как функциональные факторы: простота, логичность, хорошее описание, доступность, — так и организационные: язык разрабатывается при поддержке W3C [14] и крупнейших мировых компаний — лидеров IT-отрасли.

Литература

1. VoiceXML Forum, http://www.voicexml.org/.
2. Ericsson, http://www.ericsson.com/.
3. Siemens AG, http://w1.siemens.com/.
4. NOVELL, http://www.novell.com/.
5. France Telecom, http://www.francetelecom.com/.
6. SAMSUNG, http://www.samsung.com/.
7. Sun Microsystems, http://www.sun.com/.
8. IBM, http://www.ibm.com/.
9. AT&T, http://www.att.com/.
10. Lucent Technologies, http://www.alcatel-lucent.com/.
11. Motorola, http://www.motorola.com/.
12. VoiceXML's History, Luann Martinez, 2009, http://www.voicexml.org/voicexml-tutorials/introduction/voicexmls-history/.
13. Harry Newton, Newton's Telecom Dictionary, 24th Edition: Telecommunications, Networking, Information Technologies, The Internet, Flatiron Publishing, 2008.
14. W3C — The World Wide Web Consortium, http://www.w3.org/.
15. Voice Extensible Markup Language (VoiceXML) 2.1, http://www.w3.org/TR/voicexml21/.
16. Speech Synthesis Markup Language (SSML) Version 1.0, http://www.w3.org/TR/speech-synthesis/.
17. Speech Recognition Grammar Specification Version (SRGS) 1.0, http://www.w3.org/TR/speech-grammar/.
18. Salmon, Backus-Naur Forms, Irwin Professional Publishing, 1992.
19. Semantic Interpretation for Speech Recognition (SISR) Version 1.0, http://www.w3.org/TR/semantic-interpretation/.
20. IBM WebSphere Voice, http://www-01.ibm.com/software/voice/.
21. Microsoft Speech Technologies, http://www.microsoft.com/speech/speech2007/default.mspx.
22. Speechtech s.r.o., http://www.speechtech.cz/.
23. Voxeo IVR Platforms, http://www.voxeo.com/.

Е.Б. Никитин —

окончил факультет прикладной математики Белорусского государственного университета, прошел инструкторские курсы Cisco CCNA в Киевском национальном университете имени Тараса Шевченко, аспирант дневной формы ОИПИ НАН Беларуси.