Речевые технологии

1/2010

Содержание

Ушёл из жизни *Вадим Георгиевич Михайлов.*

Доктор филологических наук, старший научный сотрудник филологического факультета МГУ им. М.В. Ломоносова. Учёный, посвятивший большую часть своих трудов проблемам речевых технологий.

В.Г. Михайлов — автор ряда теоретических и прикладных исследований в области измерений разборчивости и качества речи в каналах связи, ставших государственным стандартом. Человек большой эрудиции, общительный и отзывчивый товарищ, Вадим Георгиевич навсегда останется в памяти сотрудников лаборатории фонетики и речевой коммуникации филологического факультета МГУ, в которой он проработал 25 лет.

Профессор Л.В. Златоустова,
научный сотрудник С.А. Крейчи

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

Издательский дом «Народное образование».
Адрес редакции: 109341, Москва, ул. Люблинская, д. 157, корп. 2. Тел.: 8 (495) 979-54-27

Подписано в печать 25.03.2011. Формат 60x90/8. Бумага офсетная. Печать офсетная.
Печ. л. 14,0. Тираж 1000 экз. Заказ
Отпечатано в типографии НИИ школьных технологий. Тел.: 8 (495) 972-59-62.

Уважаемые коллеги!

Прошли 3 года. Журнал выдержал проверку временем и стал признанным источником значимой информации для «речевиков». Сложился собственный круг авторов и читателей, который постоянно расширяется. Особенно часто к журналу обращаются в университетских библиотеках, в том числе, и студенты. Радует, что на смену известным авторитетам в области речевых технологий подрастает молодёжь.

Постепенно мы приводим журнал в соответствие формальным требованиям ВАК и рассчитываем войти в список» в 2012-м году. В частности, журнал начинает выходить тиражом достаточным для того, чтобы доставляться во все крупные научно-технические библиотеки страны. Так что мы надеемся стать Вам полезными ещё и в этом качестве.

Эти изменения стали возможны благодаря Научно-техническому центру «Поиск-ИТ» и её Генеральному директору Алексею Евгеньевичу Любимову, взявшим на себя обязанности по управлению и развитию проекта. Издательский дом «Народное образование» и НИИ школьных технологий по-прежнему будут издателями и информационными партнёрами журнала.

Я желаю всем сотрудникам, авторам, читателям, благотворителям и подписчикам журнала профессиональных успехов и научных удач!

С уважением,
Харламов А.А.,
доктор технических наук

Дорогие читатели!

Научно-технический центр «ПОИСК-ИТ» уже длительное время специализируется на превращении интеллектуальных технологий обработки информации в прикладные системы. Это направление в последние годы стало особенно актуальным, здесь открываются совершенно новые перспективы и технические возможности. Но мы столкнулись с неожиданной проблемой: существует острейший дефицит разработчиков междисциплинарного профиля, на стыке техники, информатики, лингвистики, психофизиологии и некоторых других дисциплин. Уверен, что эта проблема очевидна и для других разработчиков интеллектуальных систем с речевыми приложениями. Поэтому мы приняли решение внести свою лепту в развитие специализированного журнала, ориентированного на популяризацию «речевого» направления информатики, в том числе, и в вузовской среде. Чтобы привлечь талантливую молодёжь в тематику речевых технологий, мы обеспечим поступление журнала во все вузы России, где ведётся подготовка по нужным нам направлениям. Мы приложим также усилия к тому, чтобы журнал стал более популярным и легко читаемым.

Мне хотелось бы надеяться, что нашему примеру последуют и другие успешные производители систем обработки информации. В частности, следующим шагом после обеспечения тиражности журнала, было бы логично сформировать достойный гонорарный фонд для авторов публикаций. По нашему предложению будет разработана и внедрена в жизнь программа поощрения авторов наиболее интересных статей на конкурсной основе по решению научной редакции и прямого голосования на сайте журнала. Это неправильно, что при наличии динамичного рынка продуктов и технологий по «речевой» тематике, специализированный отраслевой журнал выходит мизерным тиражом на средства энтузиастов.

Журнал создаёт и помогает поддерживать невидимую связь в нашем, пусть небольшом, но значимом «речевом» сообществе. В эту группу «фанатов» своего дела входят учёные, инженеры, программисты и люди бизнеса, а, главное, потребители результатов коллективной работы. От того насколько эта группа понимает и принимает друг друга будет зависеть результативность российских инновационных разработок в этой области. Для нас важен не только авторитет и значимость научных достижений, но и их практическая реализация в сложных и не очень комплексах и продуктах. Надеемся, что такая синергия приведет к ускорению научно-технического прогресса в конкретной отрасли — российских речевых технологиях, — перекинет прочный мост для движения российских инновационных разработок к потребителю как внутри страны, так и за ее пределами.

Научный уровень публикуемых статей — беспрестанная забота научной редакции журнала и ее главного редактора, и здесь требования не претерпят существенных изменений. В кратчайшие сроки редакция намерена осуществить все необходимые мероприятия, чтобы включить журнал в список научных изданий Высшей аттестационной комиссии Минобразования РФ.

Наиболее интересные разработки российских ученых, опубликованные в журнале, могут получить практическую реализацию в рамках нашего предприятия. Предполагается выделение персональных премий наиболее выдающимся студентам, ученым и разработчикам, работающим по темам НТЦ «ПОИСК-ИТ». Приветствуются аналогичные предложения и других разработчиков интеллектуальных информационных систем. Таким образом, журнал станет не только перекрестком для общения и обмена мнениями по тем или иным вопросам, но и инструментом реализации самых амбициозных планов энергичных людей.

С уважением,
Алексей Любимов

Практический алгоритм определения темпа речи для использования в контакт-центрах

С.Н. Никифоров,
главный инженер ООО «Нейрон-М»

Д.С. Никифоров,

И.И. Виторский,

М.С. Танюкевич,
студенты БГУИР

В статье приводится описание алгоритма работы программы определения темпа речи. Актуальность определения темпа речи в системах обработки телефонных вызовов (контакт-центрах) определяется необходимостью регулировать темп диалога оператора с клиентом. Необходимость контроля за темпом речи оператора обусловлена двумя факторами: временем диалога, так как стоимость минуты разговора с клиентом для крупных контакт-центров достаточно велика, а оператору необходимо обслужить как можно больше клиентов; второй фактор — комфортность обслуживания клиента. Результаты данной работы используются в контакт-центре справочно-информационной службы и позволили на 15% повысить эффективность трафика за счёт оптимизации темпа речи оператора.

Abstract
In the article the description of an algorithm of the programme of definition of the speech rate is given. Results of given job are used at the call-center service and have allowed on 15% increasing of effectiveness of traffic.

Введение

Определение темпа речи, а это не что иное, как сегментация непрерывного потока речи на слоги, определение и измерение гласных звуков, а также выделение пауз на фоне шумов, характерных для телефонной линии, — одна из основных задач распознавания речи. Известен целый ряд алгоритмов, использующих традиционную обработку речевого сигнала в частотной или временной области, выделяющих формантные характеристики [1], [2], [3]. В качестве альтернативных методов используются скрытые марковские модели (СММ). С целью достижения максимальной скорости работы и возможности использовать в многоканальных (десятки каналов) системах обработки вызовов в данной работе использовались комбинации алгоритмов, основанных на обработке речи во временной области. Такой подход позволил при минимальном использовании компьютерных ресурсов достичь приемлемой скорости работы параллельно в нескольких десятках каналов.

Общее описание используемых алгоритмов

Определение темпа речи основано на использовании двух алгоритмов: определении длительности пауз и выделении и оценке длительности слоговых сегментов в речевом сигнале. Локализация пауз проводится методом цифровой фильтрации в двух спектральных диапазонах, соответствующих локализации максимумов энергии для вокализованных и шумных (невокализованных) звуков полосовыми фильтрами четвёртого порядка, «взвешивания» кратковременной энергии речевого сигнала в двух частотных диапазонах с использованием прямоугольного окна длительностью 20 мс [1].

Определение длительности слоговых сегментов основано на слуховой модели, учитывающей спектральное распределение гласных звуков, фильтрации в двух взаимно коррелированных спектральных диапазонах. Принятие решения о принадлежности сегмента речи к слогу, содержащему гласный звук, и локализация гласного звука проводятся программно реализованной комбинационной логической схемой [5].

Заключение о скорости речи говорящего (темпе речи) производится на основании анализа обоими алгоритмами на интервале накопления информации: всего файла для режима «OffLine» или чтением потока (файла) с выводом результатов каждые 15 с для режима «OnLine».

В общем случае алгоритм определения темпа речи состоит из следующих этапов:

- Нормирование речевого сигнала. Обеспечивает выравнивание слабых (тихих) сигналов с целью исключения зависимости результатов измерения от громкости входного речевого сигнала.
- Выделение и измерение длительности пауз. Формирование первичных признаков темпа (алгоритм 1).
- Оценка длительности слоговых сегментов. Формирование главных признаков (алгоритм 2).
- Принятие решения о темпе речи.

Структура системы определения длительности пауз в непрерывном потоке речи *(алгоритм 1)*

1. Нормирование входного речевого сигнала

Входной речевой сигнал нормируется для исключения зависимости результатов измерений от амплитуды (громкости) записанного или вводимого сигнала.

Нормирование производится следующим образом:

— на интервалах длительностью 1 с производится поиск максимального абсолютного значения амплитуды;

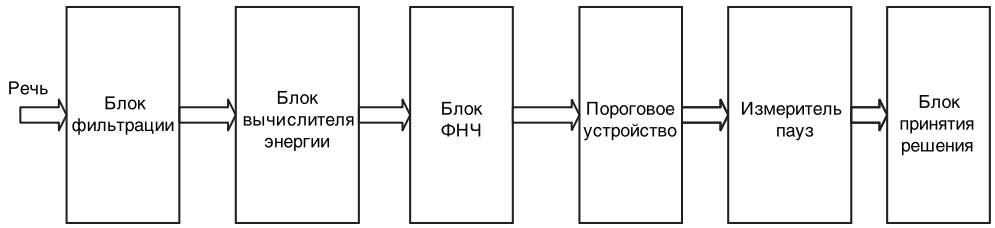
— находится среднее значение в полученном массиве;

— определяется коэффициент пересчёта, равный отношению максимально возможного значения амплитуды к найденному среднему значению;

— каждое значение входного сигнала умножается на коэффициент пересчёта.

2. Выделение и измерение длительности пауз

Метод основан на измерении мгновенной энергии в двух частотных диапазонах, соответствующих максимальному сосредоточению энергии вокализованных (диапазон частот 150 – 1000 Гц) и невокализованных (диапазон частот 1500 – 3 500 Гц) звуков. Структурная схема показана ниже.



2.1. Фильтрация

Формула типового рекурсивного звена фильтрации второго порядка в Z-области соответствует выражению [4]:

$$Y(Z) = (1 - K1 \times Z^{-1}) / (1 + 2K1 \times Z^{-1} - K2 \times Z^{-2}),$$

что эквивалентно разностному уравнению во временной области вида:

$$Y(n) = (2 \times Y1 - X1) \times K1 - Y2 \times K2 + X(n),$$

где $K1 = K \times \cos(2\pi \times Frq/Fd)$;

$K = 1.0 - \pi \times Pol/Fd$;

$K2 = K \times K$;

$X(n)$ — текущее значение входного сигнала;

$Y(n)$ — текущее значение выходного сигнала;

$Y1$ — значение выходного сигнала, задержанное на один период дискретизации;

$Y2$ — значение выходного сигнала, задержанное на два периода дискретизации;

Pol — полоса пропускания в Гц;

Fd — частота дискретизации в Гц;
Frq — средняя частота полосы фильтра в Гц.

Фильтр 4-го порядка реализуется путём каскадного последовательного соединения двух звеньев второго порядка указанного типа.

2.2. Расчёт мгновенной энергии речевого сигнала

Расчёт мгновенной энергии производится на интервалах (в окне длительностью 20 мс), что соответствует для частоты дискретизации ***Fd*** = 8000 Гц 160 отсчётам входного речевого сигнала [1].

Последовательность действий при вычислении мгновенной энергии следующая:
— вычисляется модуль ***Yнв = Abs (Yn)*** — выпрямление выходного сигнала фильтра,

— затем вычисляется значение мгновенной величины энергии в окне 20 мс (160 отсчётов) по формуле $Sn = M \times \sum_{1}^{160} Yнв \times Yнв,$

где ***Sn*** — значение мгновенной энергии в ***n***-м окне;
Yn — выходное значение фильтра;
Yнв — выпрямленное выходное значение;
M — масштабный коэффициент, ограничивающий переполнение. Мгновенная энергия рассчитывается в двух частотных диапазонах.

2.3. Расчёт ФНЧ

На третьем этапе сглаживаются (усредняются) результаты расчёта мгновенной энергии, для чего используется фильтр нижних частот (ФНЧ) первого порядка, соответствующий ***Z*** — уравнению ***Y(Z) = K / 1 – K × Z1*** или разностному уравнению вида ***Y(n) = (1–k)Y1–1+S(n),***
где ***Y(n)*** — текущее выходное значение ФНЧ;
S(n) — текущее входное значение ФНЧ (значение мгновенной энергии);
Y1 — задержанное на период дискретизации значение выходного сигнала;
K — коэффициент, определяющий постоянную времени или частоту среза ФНЧ.

2.4. Пороговое устройство

Пороговое устройство сравнивает текущее значение сглаженного значения средней энергии в заданной полосе с пороговым значением (определяется экспериментально), за начальный уровень может быть принято значение 50 мВ. За паузу принимается значение энергии меньше уровня порогов в обоих спектральных диапазонах. С этого момента начинается отсчёт длительности паузы.

2.5. Счётчик средней продолжительности пауз в файле

Средняя продолжительность паузы в обрабатываемом файле или на анализируемом участке определяется как сумма дин всех пауз, делённая на их количество $Tcc = 1/N \times (\sum_{1}^{Ni} Ti)$,

где Tcc — средняя длительность паузы;
N — количество пауз на анализируемом участке.

2.6. Блок принятия решения

Первичное заключение о темпе речи принимается исходя из следующих положений:
— при превышении средней длины паузы $\boldsymbol{Tcc}$ эталона темп считается медленным;
— при значении $\boldsymbol{Tcc}$, меньшем средней длины паузы эталона, темп считается быстрым;
— в противном случае — соответствующим эталону.

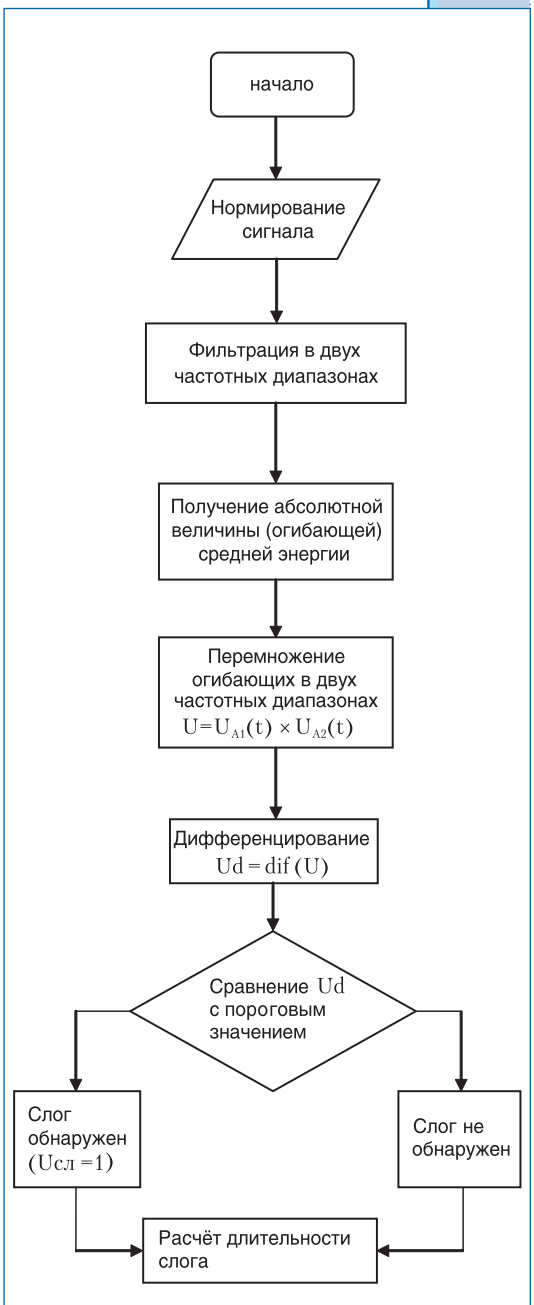
Оценка длительности слоговых сегментов (алгоритм 2)

Метод выделения признаков слоговых сегментов основан на формировании первичных параметров, использующих огибающие сигналов в частотных диапазонах $\boldsymbol{A1}$ = 800–2500 Гц и $\boldsymbol{A2}$ = 250–540 Гц. Результирующий параметр, который в дальнейшем используется для выделения признаков слогов, получается корреляционным методом и определяется так:

$$\boldsymbol{U_c(t) = U_{A1}(t)U_{A2}(t),}$$

где $\boldsymbol{U_{A1}(t)}$ — огибающая энергии в полосе частот $\boldsymbol{A1}$, а $\boldsymbol{U_{A2}(t)}$ — огибающая энергии в полосе $\boldsymbol{A2}$ [5]. Диапазон частот первого полосового фильтра, равный 250–540 Гц, выбран потому, что в нём отсутствует энергия высокоэнергетических фрикативных звуков типа /ш/ и /ч/, которые создают ошибочные слоговые ядра, а также сосредоточена значительная часть энергии всех звонких звуков, в том числе и гласных. Однако в этом диапазоне энергия сонорных звуков типа /л/, /м/, /н/ сравнима с энергией гласных, из-за чего определение слоговых сегментов только с учётом огибающей речевого сигнала в этом диапазоне сопровождается ошибками. Поэтому диапазон частот второго полосового фильтра выбран в пределах 800–2500 Гц, в котором энергия гласных звуков минимум в два раза превышает энергию сонорных звуков.

Благодаря операции умножения огибающих $\boldsymbol{U_{A1}(t)}$ и $\boldsymbol{U_{A2}(t)}$ в результирующей временной функции происходит усиление участков кривой в области гласных звуков из-за корреляции их энергий в обоих диапазонах. Кроме того, ошибочные максимумы энергии, предопределённые наличием в диапазоне 800–2500 Гц значительной части энергии фрикативных звуков, устраняются путём их умножения на



практически нулевое значение амплитуды фрикативных звуков в диапазоне 250–540 Гц.

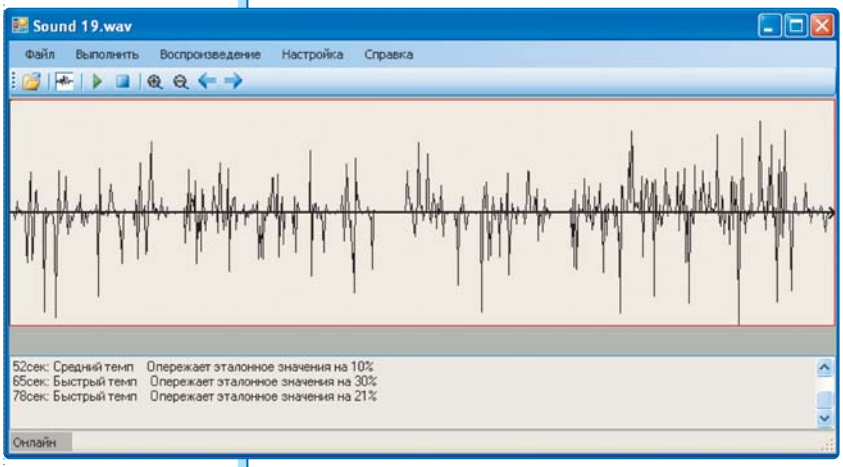
Последовательность операций при работе алгоритма следующая:

- Фильтрация сигнала двумя полосовыми рекурсивными фильтрами четвёртого порядка в диапазонах 250–540 Гц и 800–2500Гц соответственно.
- Детектирование выходных сигналов фильтров для получения огибающих.
- Перемножение огибающих выходных сигналов фильтров.
- Дифференцирование результирующего сигнала.
- Сравнение полученного сигнала с пороговыми напряжениями и выделение логического сигнала, соответствующего наличию слогового сегмента.
- Расчёт длительности слогового сегмента.

Алгоритм работы приведён на стр 7.

Механизм принятия решения о темпе речи

Принятие решения о темпе речи основывается на результате расчёта длительности пауз и слоговых сегментов. При этом реализуется следующая комбинационная логика:



— паузы длинные, слоги длинные — темп медленный. Критерием «длинные» является отклонение длительности от эталонных на 30%;

— паузы короткие или отсутствуют, слоги короткие — темп быстрый. Критерием «короткие» является отклонение длительности от эталонных на 30%;

— паузы длинные, слоги короткие — темп быстрый, т.е. приоритетным является анализ слогов, при этом выводится предупреждение о длинных паузах;

— паузы короткие или отсутствуют, слоги длинные — темп медленный.

Основной интерфейс программы показан слева.

Результаты тестирования программы определения темпа речи

1. Общая оценка качества работы программы

1.1. Оценка работы в режиме OffLine

п/п	*Имя Wav-файла*	*Характеристика файла*	*Результат измерения темпа (интегральная оценка по файлу)*	*Оценка работы программы**
1	1_enh	Темп средний, ближе к замедленному, мужской	Медленный, отставание от эталона –39%	+
2	2_enh	Темп медленный, паузы большие, мужской	Медленный, отставание от эталона –56%	+
3	Bistro	Темп быстрый, мужской	Быстрый, опережение +25%	+
4	Dictor	Темп средний, диктор	Средний, опережение +9%	+

		радио, мужской		
5	Dim2	Темп средний, мужской, нечёткая дикция	Средний, отставание –2%	+
6	Gromko	Темп средний, мужской, громко	Средний, отставание –8%	+
7	Ira	Темп переменный, женский	Средний, отставание –8%	+
8	Medlenno	Темп медленный, мужской	Медленный, отставание –45%	+
9	Pause2	Темп средний, длинные паузы, мужской	Средний, отставание –11%, предупреждение о длинных паузах	+
10	Radio	Темп средний, мужской, диктор радио	Средний, отставание 0%	+
11	Tixo	Темп средний, мужской, тихо	Средний, отставание –23%	+
12	Шепот	Темп средний, женский, очень тихо	Средний, отставание –6%	+
13	AA	Мужской, слоги, гласные короткие, паузы средние	Средний, отставание –5%	+
14	AAA	Мужской, слоги, гласные длинные, паузы средние	Медленный, отставание –117%	+
15	Sound20	Темп средний, мужской, тихо, нечётко	Средний, отставание –7%	+
16	Sound19	Темп средний, мужской	Средний, отставание –2%	+
17	Sound10	Темп замедленный, мужской	Средний, отставание –28%	+
18	Sound5	Темп быстрый, мужской	Быстрый, опережение +20%	+
19	Etalon	Темп средний, мужской	Средний, опережение +1%	+
20	F1_10	Темп средний, женский	Средний, отставание –4%	+

* «+» — соответствует. «–» не соответствует.
** — Эталонный файл — Etalon4.wav с параметрами: паузы — 313 мс, слоги — 98 мс.

1.2. Оценка работы в режиме OnLine. Измерение параметров записи при различных уровнях громкости

п/п	*Имя Wav-файла*	*Характеристика файла*	*Результат измерения длительности слогов программой (мс)*	*Оценка работы программы*
1	11_norm	Темп средний, мужской, громкость средняя (эталон)	143	+
2	11_gromko	Темп средний, мужской, громкость выше на 50% от эталона	144	+
3	11_tiho	Темп средний, мужской, громкость ниже на 50% от эталона	135	+

2. Измерение параметров записи при длинных паузах

п/п	*Имя Wav-файла*	*Характеристика файла*	*Результат измерения средней длительности пауз вручную (мс)*	*Результат измерения средней длительности программой (мс)*	*Оценка работы программой (мс)*
1	Paise2	Темп средний, мужской, громкость средняя, длинные паузы	2 142	2 230 Темп средний	+

* — эталонный файл — Etalon4 с параметрами: паузы — 313.

Заключение

Программа определения темпа речи выполнена в двух вариантах:
— Тестовый модуль для работы в режиме работы с файлами.
— Динамическая библиотека, предназначенная для встраивания разработчиками систем обработки вызовов в конечный продукт.

Внедрение программы в контакт-центр справочно-информационной службы позволило на 15% повысить эффективность трафика за счёт оптимизации темпа речи оператора.

Алгоритм, не требующий больших вычислительных ресурсов, и оптимизированный по времени и объёму программный код позволяют использовать результаты данной работы во встраиваемых микропроцессорных системах обработки речевых сигналов.

Литература

1. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов. М.: Радио и связь, 1981. С. 113–119.
2. Сапожков М.А., Михайлов В.Г. Вокодерная связь. М.: Радио и связь,1983. С. 156–158.
3. Дегтярев Н.П. Параметрическое и информационное описание речевых сигналов. Минск: Объединённый институт проблем информатики Национальной академии наук Беларуси, 2003. С.62–63.
4. Лобанов Б.М., Цирульник Л.И. Компьютерный синтез и клонирование речи. Минск: Белорусская наука, 2008. С.60–63.
5. Быков Н.М. и др. Надёжный метод выделения слоговых сегментов в речевом сигнале // Автоматика и информационно-измерительная техника. 2007. № 1.

Никифоров Сергей Никонорович —
главный инженер ООО «Нейрон-М», г. Минск.
Сфера интересов: цифровая обработка сигналов, синтез и распознавание речи.

Никифоров Дмитрий Сергеевич —
студент БГУИР, г. Минск, сфера интересов: цифровая обработка сигналов, синтез и распознавание речи.

Виторский Иван Игоревич —
студент БГУИР, г. Минск, сфера интересов: распознавание речи.

Танюкевич Михаил Сергеевич —
студент БГУИР, г. Минск, сфера интересов: распознавание речи.

Субполосная обработка сигналов: эффективность и применение в речевых технологиях

А.А. Петровский,
кандидат технических наук, доцент

Приводится краткое введение в банки цифровых фильтров, даются основные определения и понятия: полное восстановление или перфективная реконструкция, параунитарный банк, субполосное кодирование, скорость передачи. Рассматривается общий случай схемы субполосного кодера. Выводятся оценки ошибки реконструкции сигнала, а также оптимальное распределение бит по каналам. Доказывается эффективность субполосного кодирования по отношению к полнополосному кодированию. Показано, что если входной сигнал имеет нормальный закон распределения, коэффициенты децимации в каналах равны, банк фильтров — ортогональный, то субполосный кодер обеспечивает равное или лучшее качество по сравнению с полнополосным кодером при любом входном сигнале. Эффективность применения субполосной обработки речи показана на примере системы редактирования шума и кодирования речевого сигнала в вейвлет области.

1. Введение в банки фильтров

1.1. Определения

Банк фильтров — цифровая система, состоящая из секций анализа и синтеза, называемых банком фильтров анализа и банком фильтров синтеза *(рис. 1)*. Входной сигнал $x(n)$, представленный последовательностью отсчетов, разбивается при помощи фильтров секции анализа $H_k(z)(k=0,1,\ldots,M-1)$ на M субполосных составляющих, которые в идеальном случае в частотной области не перекрываются. Операции, выполняемые секцией синтеза, являются обратными операциями секции анализа. Подобрав соответствующим образом набор фильтров секции синтеза $F_k(z)(k=0,1,\ldots,M-1)$, можно восстановить исходный сигнал $y(n)$ из его субполосных компонент [1].

Банк фильтров относится к классу многоскоростных систем цифровой обработки сигналов [1–4], в которых частота дискретизации различна в разных точках системы.

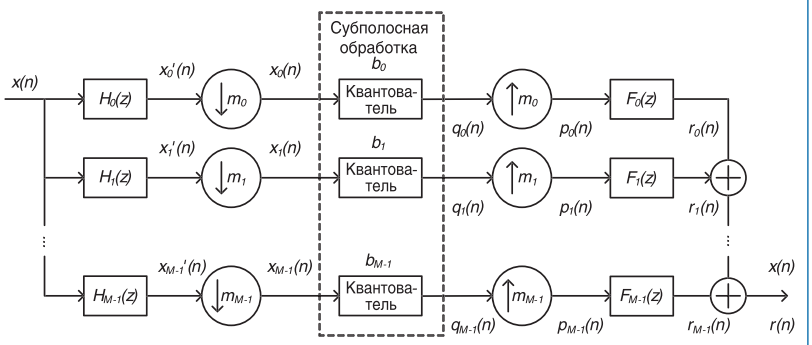


Рис. 1. Банк фильтров: система анализа/синтеза сигнала

Значение коэффициента темпа поступления отсчетов в канале (коэффициента децимации m_k) задаётся утверждением о дискретизации сигнала в зависимости от ширины частотной полосы канала B_k и его положения в банке фильтров. Оригинальная частота дискретизации f_s сигнала в k-м канале теоретически может быть уменьшена в $m_k \le f_s/B_k$ раз. Равенство в данном случае означает, что канал максимально децимирован:

$$\sum_{k=0}^{M-1} \frac{1}{m_k} = 1. \qquad (1.1)$$

Таким образом, в частотно-временном представлении сигнала исключена избыточность, т.е. сумма отсчётов во всех каналах соответствует количеству отсчётов в исходном сигнале. Банк фильтров считается передецимированным, если $\sum_{k=0}^{M-1} \frac{1}{m_k} > 1$, т.е. достаточно, чтобы хотя бы в одном канале коэффициент децимации не был равен числу каналов. Более сложные случаи позиционирования субполос каналов в банках фильтров рассмотрены в [4–6].

Соотношение между z-преобразованием сигналов на входе и выходе M-канального банка фильтров, изображённого на *рис. 1*, следующее [1]:

$$Y(z) = \sum_{k=0}^{M-1} F_k(z) \frac{1}{m_k} \sum_{l=0}^{m_{k-1}} H_k\left(zW_{m_k}^l\right) X\left(zW_{m_k}^l\right) \qquad (1.2)$$

где $W_{m_k} = e^{-j2\pi/m_k}$. Анализ данного выражения показывает, что в банке возможны искажения входного сигнала: амплитудные, фазовые и отражения частотных характеристик (элайзинг), обусловленные наличием дециматоров и интерполяторов. Величина искажений зависит как от частотных характеристик канальных фильтров, так и выбора коэффициентов децимации m_k. Анализ искажений, возникающих в структуре банка фильтров, проще анализировать в максимально децимированном банке фильтров, для которого соотношение (1.2) значительно упрощается:

$$Y(z) = \sum_{k=0}^{M-1} T_k(z) X\left(zW_M^k\right), \qquad (1.3)$$

где $T_k(z)$ — передаточная функция k-го канала:

$$T_k(z) = \frac{1}{M} \sum_{l=0}^{M-1} F_l(z) H_l\left(zW_M^k\right). \qquad (1.4)$$

Выходной сигнал $y(n)$ системы анализа-синтеза банка фильтров будет свободен от элайзинговой составляющей $X(zW_M^k)$, $k>0$ в случае, если

$$T_k(z) = 0, \text{для } 1 \le k \le M. \qquad (1.5)$$

В банке фильтров, для которого справедливо условие (1.5), остаются только амплитудные и фазовые искажения, которые определяются согласно следующему выражению:

$$\frac{Y(z)}{X(z)} = T_0(z) = \frac{1}{M}\sum_{l=0}^{M-1} F_k(z)H_k(z). \qquad (1.6)$$

Очевидно, что для получения перфективной реконструкции входного сигнала $x(n)$ банком фильтров, передаточная функция искажений $T_o(z)$ должна принять форму простого звена задержки с некоторым масштабированием амплитуды:

$$T_0(z) = cz^{-\Delta}, c \ne 0, \Delta \in \Xi. \qquad (1.7)$$

Полное восстановление или перфективная реконструкция — свойство банка цифровых фильтров, заключающееся в том, что сигнал, прошедший через схему анализа-синтеза, идентичен входному с точностью до задержки. Для этого фильтры синтеза должны подавлять наложение частотных характеристик (элайзинг) и устранять амплитудные и фазовые искажения [1].

Параунитарный (ортогональный) банк фильтров (ПУБФ) — банк фильтров, у которого передаточные функции анализирующих и синтезирующих фильтров и их соответственно смещенные версии ортогональны друг другу. Фильтры синтеза в параунитарных банках являются транспонированными версиями фильтров анализа [1]:

$$F_k(z) = H_k^T(z^{-1}). \qquad (1.8)$$

При соблюдении этого условия обеспечивается возможность перфективной реконструкции банком фильтров входного сигнала $x(n)$ пусть $x[n] = [x_o[n] \dots x_{M-1}[n]]^T$ будет входным вектором и $y[n] = [y_o[n] \dots y_{M-1}[n]]^T$ будет соответствующим выходным вектором с $M \times N$ параунитарной передаточной матрицей $A(z)$. Пусть $S_{xx}(e^{jw})$ будет $M \times M$ СПМ-матрица входного вектора $x[n]$. Заметим, что СПМ i-го входного компонента $x_i[n]$ является i-й элемент $S_{xx}(e^{jw})$. Поэтому дисперсия $x_i[n]$ составит

$$\int_{-\pi}^{\pi} \left(S_{xx}\left(e^{jw}\right)\right)_i \frac{dw}{2\pi}. \qquad (1.9)$$

Усредненная дисперсия входного сигнала будет

$$\frac{1}{N}\sum_{i=0}^{N-1}\sigma_i^2 = \frac{1}{N}\int_{-\pi}^{\pi} tr\left(S_{xx}\left(e^{jw}\right)\right)\frac{dw}{2\pi}, \qquad (1.10)$$

где $tr(B)$ — след матрицы B.

СПМ вектора на выходе определяется как

$$S_{yy}\left(e^{jw}\right) = A\left(e^{jw}\right)S_{xx}\left(e^{jw}\right)A^H\left(e^{jw}\right), \qquad (1.11)$$

где A^H — матрица, эрмитово транспонированная к матрице A.

Усредненная дисперсия выхода составит

$$\frac{1}{N}\int_{-\pi}^{\pi} tr\left(S_{yy}\left(e^{jw}\right)\right)\frac{dw}{2\pi} = \frac{1}{N}\int_{-\pi}^{\pi} tr\left(A\left(e^{jw}\right)S_{xx}\left(e^{jw}\right)A^H\left(e^{jw}\right)\right)\frac{dw}{2\pi}. \quad (1.12)$$

Так как $tr(AB) = tr(BA)$, это упрощает выражение (1.12)

$$\frac{1}{N}\int_{-\pi}^{\pi} tr\left(S_{xx}\left(e^{jw}\right)A^H\left(e^{jw}\right)A\left(e^{jw}\right)\right)\frac{dw}{2\pi}.$$

Так как матрица $A(z)$ — параунитарная, $A^H(e^{jw})A(e^{jw}) = I$, следовательно

$$\frac{1}{N}\int_{-\pi}^{\pi} tr\left(S_{xx}\left(e^{jw}\right)\right)\frac{dw}{2\pi} \quad (1.13)$$

или усреднённая дисперсия выхода равна усреднённой дисперсии входа. Этот факт показывает, что в параунитарной системе энергия сохраняется.

Банк фильтров можно представить в полифазной форме, если передаточные функции секций анализа и синтеза записать в виде соответствующих векторов:

$$\begin{aligned}\mathbf{H}(z) &= [H_0(z)\ H_1(z)\ \ldots H_{M-1}(z)]^T,\\ \mathbf{F}(z) &= [F_0(z)\ F_1(z)\ \ldots F_{M-1}(z)]^T,\end{aligned} \quad (1.14)$$

то тогда можно выбрать такие полифазные матрицы анализа

$$\mathbf{E}(z) = \begin{bmatrix} E_{0,0}(z) & \cdots & E_{0,M-1}(z) \\ \vdots & \ddots & \vdots \\ E_{M-1,0}(z) & \cdots & E_{M-1,M-1}(z) \end{bmatrix} \quad (1.15)$$

и синтеза

$$\mathbf{D}(z) = \begin{bmatrix} D_{0,0}(z) & \cdots & D_{0,M-1}(z) \\ \vdots & \ddots & \vdots \\ D_{M-1,0}(z) & \cdots & D_{M-1,M-1}(z) \end{bmatrix}, \quad (1.16)$$

Вектора передаточных функций секций анализа и синтеза можно представить следующим образом [1]:

$$\begin{aligned}\mathbf{H}(z) &= \mathbf{E}(z^M)\left[1\ z^1\ \ldots z^{-(M-1)}\right]^T,\\ \mathbf{F}(z) &= \left[z^{-(M-1)}\ z^{-(M-2)}\ \ldots 1\right]^T\mathbf{D}(z^M).\end{aligned} \quad (1.17)$$

Для получения перфективной реконструкции на компоненты полифазных матриц накладывается дополнительное ограничение:

$$\mathbf{D}(z)\mathbf{E}(z) = cz^{-\Delta}\mathbf{I}, c \neq 0, \Delta \geq 0, \quad (1.18)$$

где c — ненулевая константа; Δ — задержка, выраженная целым числом интервалов дискретизации, вносимая секциями анализа-синтеза; I — единичная матрица. На *рис. 2* показана полифазная структура банка фильтров.

Традиционно банки фильтров разделяют на банки с равнополосными и неравнополосными каналами, ортогональные, биортогональные, двухканальные и многоканальные и т.д. Каждый фильтр банка цифровых фильтров образует канал. Поэтому говорят об M-канальном банке фильтров. Сигнал в канале называется субполосой, отсюда название «субполосная фильтрация» или «субполосное кодирование» [3, 4].

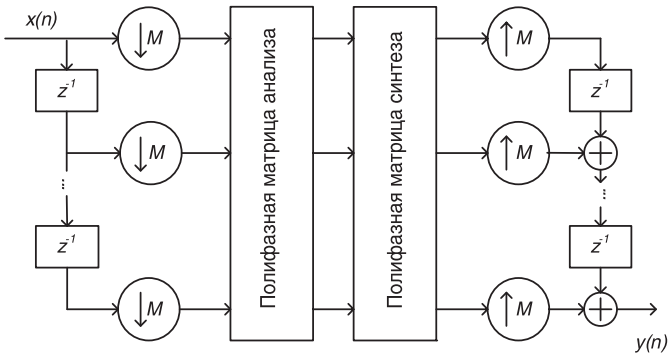


Рис. 2. Полифазная структура банка фильтров

1.2. Ошибка реконструкции сигнала

Рассматривается субполосный кодер, представленный на рис. 1.1, состоящий из $k = 0, ..., M-1$ каналов. Каждый канал имеет анализирующий фильтр $H_k(z)$, синтезирующий фильтр $F_k(z)$, дециматор/интерполятор с коэффициентом передискретизации m_k и b_k-битный квантователь. Входной сигнал $x(n)$ после фильтрации анализирующими фильтрами на рис. 1.1 обозначен как $x'(n)$, а канальные сигналы — $x_k(n)$. Субполосный кодер максимально децимирован, если выполняется условие (1.1).

Пусть входной сигнал $x(n)$ стационарный в широком смысле и имеет среднее значение, равное нулю. Следовательно, все последующие (производные) сигналы (включая шум квантования) также будут иметь среднее значение, равное нулю. Это предположение не является причиной возникновения каких-либо трудностей для применения субполосного кодирования к сигналам со средним значением, отличным от нуля, таких как изображения, так как система в этом случае эквивалентна системе с нулевым средним, полученной путем вычитания среднего значения из входного сигнала [7].

Следовательно, если $S_{xx}(e^{jw})$ является спектральной плотностью мощности (СПМ) сигнала $x(n)$, то его дисперсия определяется как [1]

$$\sigma_x^2 = \int_{-\pi}^{\pi} S_{xx}\left(e^{j\omega}\right)\frac{\partial\omega}{2\pi}.$$

Дисперсия сигнала $x'_k(n)$ на выходе анализирующего фильтра $H_k(z)$ будет определяться следующим выражением [1]

$$\sigma_k^2 = \int_{-\pi}^{\pi} S_{xx}\left(e^{j\omega}\right)\left|H_k\left(e^{j\omega}\right)\right|^2\frac{\partial\omega}{2\pi}$$

для $k = 0,1 \ldots, M-1$. Так как сигнал стационарен в широком смысле, то дисперсия не изменится после децимации [1,7], т.е. дисперсия $x_k(n)$ субполосы k равна σ_k^2. Квантователь представляется как модель с аддитивным шумом [8], т.е. выходом квантователя является сигнал $x_k(n) + q_k(n)$, где $x_k(n)$ — входной сигнал и $q_k(n)$ — шум квантования. Дисперсия шума квантования, как для равномерных квантователей, так и для квантователей, адаптированных под сигнал, определяется как [1,8]

$$\sigma_{q_k}^2 = \varepsilon_k 2^{-2b_k}\sigma_k^2, \qquad (1.19)$$

где ε_k — константа, характеризующая квантователь, которая зависит от функции плотности вероятности k-го субполосного сигнала.

Предположим, используются равномерные (многобитные) квантователи, тогда шумы квантования $q_k(n)$ различных каналов некоррелированы между собой и являются «белым» шумом [9]. СПМ шума $q_k(n)$ является константой σ_{qk}^2, так как это «белый» шум, и имеют нулевое среднее. Сигналы $p_k(n)$, полученные после интерполяции $q_k(n)$, определяются как

$$p_k(n) = \begin{cases} q_k(n/m_k), & \text{если } n \bmod m_k = 0 \\ 0 & \text{иначе} \end{cases},$$

и больше не являются стационарными в широком смысле, но имеют циклическую стационарность с периодом m_k [7]. Другими словами, если $n \bmod m_k = 0$, то СПМ сигнала $p_k(n)$ равна σ_{qk}^2, а СПМ оставшихся отсчетов равна нулю. Отсчеты шума $p_k(n)$ поступают на вход синтезирующего фильтра $F_k(n)$, на выходе которого получается сигнал $r_k(n)$, стационарный в широком смысле с периодом m_k. Усредненная на периоде m_k дисперсия сигнала $r_k(n)$ определяется как

$$\frac{1}{m_k}\left[\int_{-\pi}^{\pi}\sigma_{q_k}^2\left|F_k\left(e^{j\omega}\right)\right|^2\frac{\partial\omega}{2\pi} + 0 + \cdots + 0\right],$$

где $m_k - 1$ — нули, полученные из нулевых отсчетов сигнала шума $p_k(n)$. Ошибка реконструкции $r(n)$ равна сумме всех ошибок $r_k(n)$, что обусловлено линейностью секции синтеза банка фильтров. Так как $r_k(n)$ во всех каналах некоррелированы, то дисперсия их суммы равна сумме их дисперсий [10]

$$\sigma_r^2 = \sum_{k=0}^{M-1}\frac{\sigma_{q_k}^2}{m_k}\int_{-\pi}^{\pi}\left|F_k\left(e^{j\omega}\right)\right|^2\frac{\partial\omega}{2\pi}, \qquad (1.20)$$

где σ_r^2 — дисперсия ошибки реконструкции. Пусть $n_k = \int_{-\pi}^{\pi}\left|F_k\left(e^{j\omega}\right)\right|^2\frac{\partial\omega}{2\pi}$ означает нормы синтезирующих фильтров (заметим, что для КИХ-фильтра с коэффициентами импульсной характеристики $(h_o, h_1, \ldots, h_L)$ норма равна $(h_o^2 + h_1^2 + \ldots + h_L^2)$. Затем, используя уравнение (1.19), выражение (1.20) преобразуется к следующему виду

$$\sigma_r^2 = \sum_{k=0}^{M-1}\frac{\varepsilon_k 2^{-2b_k}\sigma_k^2 n_k}{m_k}. \qquad (1.21)$$

Для параунитарных систем [1] свойство отсутствия потерь (или энергетический баланс) говорит о том, что энергия выхода равна энергии входа (1.13). Вследствие этого факта, дисперсия ошибки реконструкции сигнала банком фильтров определяется следующим выражением:

$$\sigma_r^2 = \sum_{k=0}^{M-1}\frac{\varepsilon_k 2^{-2b_k}\sigma_k^2}{m_k}. \qquad (1.22)$$

Сравнивая уравнения (1.21) и (1.22) для ортогонального случая можем заметить, что в (1.22) отсутствует норма синтезирующего фильтра n_k. Это является следствием свойства отсутствия потерь, подразумевающего фильтры с единичной энергией или $n_k = 1$ для всех k [1]. Следует отметить также, что допущение об ортогональном банке фильтров заменяет допущение о равномерных (многобитных) квантователях и не обязательнодля получения данного результата. Поэтому, в смысле квантователя, результат более общий для ортогонального случая.

1.3. Скорость передачи субполосного кодера

Для субполосного кодера на *рис. 1* b_k определяет количество бит на отсчет для k-го квантователя. Однако благодаря коэффициенту децимации m_k данным квантователем квантуется один отсчет для каждого из m входных отсчётов. Поэтому скорость передачи квантователя равна b_k/m_k бит на входной отсчёт. Следовательно, усредненная скорость передачи субполосной системы можно определить как

$$b = \sum_{k=0}^{M-1} \frac{b_k}{m_k} \text{ бит/отсчет.} \qquad (1.23)$$

Полнополосный кодер просто квантует вход $x(n)$, используя b-битный квантователь. Поэтому из уравнения (1.22) его дисперсия ошибки квантования (которая также является дисперсией ошибки реконструкции) определяется как

$$\sigma_q^2 = \varepsilon 2^{-2b} \sigma_x^2, \qquad (1.24)$$

где ε определяется как и ранее.

Эффективность субполосного кодирования вычисляется как отношение дисперсии ошибки реконструкции полнополосного кодера (или импульсно-кодовой модуляции — ИКМ) σ_q^2 к дисперсии ошибки реконструкции субполосного кодера σ_r^2 с аналогичной скоростью передачи данных [8].

2. Оценка эффективности субполосного кодирования

2.1. Оптимальное распределение бит по каналам субполосного кодера

Проблема оптимального распределения бит в каналах состоит в нахождении $b_o, \ldots, b_{M-1}$, которые минимизируют дисперсию ошибки реконструкции σ_r^2 в уравнении (1.21), удовлетворяя ограничению (1.23). Данная минимизация с ограничением (1.23) может быть решена с использованием метода множителей Лагранжа. Для этого определяется целевая функция

$$C = \sum_{k=0}^{M-1} \frac{\varepsilon_k 2^{-2b_k} \sigma_k^2 n_k}{m_k} + \lambda \left(\sum_{k=0}^{M-1} \frac{b_k}{m_k} - b \right),$$

где λ — множитель Лагранжа. Дифференцируя C по b_k и приравнивая к нулю, $(dC/db_k = 0)$ или

$$\frac{1}{m_k} \varepsilon_k 2^{-2b_k} \sigma_k^2 n_k (-2 \ln 2) + \lambda \frac{1}{m_k} = 0,$$

из которого следует, что

$$\frac{\partial}{\partial b_k} 2^{-2b_k} = \frac{\partial}{\partial b_k} e^{-2\ln 2 b_k} = -2\ln 2 e^{-2\ln 2 b_k}.$$

Таким образом,

$$2\ln 2 \varepsilon_k 2^{-2b_k} \sigma_k^2 n_k = \lambda,$$

или

$$2^{-2b_k} = \frac{\lambda}{2\,ln\,2\varepsilon_k\,\sigma_k^2 n_k}. \qquad (2.1)$$

После логарифмирования по основанию 2 обеих частей (1.25) получается, что

$$-2b_k = \log_2 \frac{\lambda}{2\ln 2 \varepsilon_k\,\sigma_k^2 n_k},$$

или

$$\sigma_r^2 = \sum_{k=0}^{M-1} \frac{\varepsilon_k 2^{-2b_k} \sigma_k^2}{m_k}. \qquad (2.2)$$

Данное выражение действительно для всех значений $k = 0, \ldots, M-1$. Подставляя уравнение (2.2) в ограничение (1.23), получается, что

$$b = \sum_{k=0}^{M-1} \frac{1}{m_k} \left[\frac{1}{2}\log_2 \frac{2\ln 2}{\lambda} + \frac{1}{2}\log_2(\varepsilon_k \sigma_k^2 n_k)\right] =$$
$$= \frac{1}{2}\log_2 \frac{2\ln 2}{\lambda} \sum_{k=0}^{M-1} \frac{1}{m_k} + \frac{1}{2}\sum_{k=0}^{M-1} \log_2 (\varepsilon_k \sigma_k^2 n_k)^{\frac{1}{m_k}}. \qquad (2.3)$$

Первая сумма здесь равна 1 на основании уравнения (1.16). Вторая сумма логарифмов может быть записана как логарифм произведения. Используя эти упрощения, средняя скорость передачи в субполосной системе равна

$$b = \frac{1}{2}\log_2 \frac{2\ln 2}{\lambda} + \frac{1}{2}\log_2 \prod_{i=0}^{M-1} (\varepsilon_i \sigma_i^2 n_i)^{\frac{1}{m_i}}, \qquad (2.4)$$

или

$$\frac{1}{2}\log_2 \frac{2\ln 2}{\lambda} = b - \frac{1}{2}\log_2 \prod_{i=0}^{M-1} (\varepsilon_i \sigma_i^2 n_i)^{\frac{1}{m_i}}. \qquad (2.5)$$

Подставляя данный результат в уравнение (2.2), получается следующее распределение бит по каналам:

$$b_k = b - \frac{1}{2}\log_2 \prod_{i=0}^{M-1} (\varepsilon_i \sigma_i^2 n_i)^{\frac{1}{m_i}} + \frac{1}{2}\log_2(\varepsilon_k \sigma_k^2 n_k) =$$
$$= b + \frac{1}{2}\log_2 \frac{\varepsilon_k \sigma_k^2 n_k}{\prod_{i=0}^{M-1} (\varepsilon_i \sigma_i^2 n_i)^{\frac{1}{m_i}}}, \qquad (2.6)$$

для всех значений $k = 0, \ldots, M-1$.

Это и есть оптимальное распределение бит по каналам субполосного кодера. Заметим, что для ортогонального случая оптимальное распределение бит получается путем подстановки $n_k = 1$ для всех k. Результирующая скорость передачи — действительная величина и может быть отрицательной.

2.2. Дисперсия минимальной ошибки реконструкции сигнала

На основании оптимального распределения бит (уравнение (2.6)) следует, что

$$2^{-2b_k} = 2^{-2b}\frac{\prod_{i=0}^{M-1}(\varepsilon_i\sigma_i^2 n_i)^{\frac{1}{m_i}}}{\varepsilon_k\sigma_k^2 n_k},$$

для всех $k = 0, \ldots, M-1$. Подставляя это выражение в уравнение (1.22) и принимая во внимание (1.1), получается, что дисперсия минимальной ошибки реконструкции сигнала равна

$$\sigma_r^2 = \sum_{k=0}^{M-1}\frac{1}{m_k}2^{-2b}\prod_{i=0}^{M-1}(\varepsilon_i\sigma_i^2 n_i)^{\frac{1}{m_i}} = 2^{-2b}\prod_{i=0}^{M-1}(\varepsilon_i\sigma_i^2 n_i)^{\frac{1}{m_i}}\sum_{k=0}^{M-1}\frac{1}{m_k} = \\ = 2^{-2b}\prod_{i=0}^{M-1}(\varepsilon_i\sigma_i^2 n_i)^{\frac{1}{m_i}}, \quad (2.7)$$

2.3. Оценки эффективности субполосного кодирования

Эффективность субполосного кодирования можно определить как отношение дисперсии ошибки реконструкции полнополосного кодера σ_q^2 (1.24) к дисперсии ошибки реконструкции субполосного кодера σ_r^2 (1.31) с аналогичной скоростью передачи данных

$$G = \frac{\varepsilon 2^{-2b}\sigma_x^2}{2^{-2b}\prod_{i=0}^{M-1}(\varepsilon_i\sigma_i^2 n_i)^{\frac{1}{m_i}}} = \frac{\varepsilon\sigma_x^2}{\prod_{i=0}^{M-1}(\varepsilon_i\sigma_i^2 n_i)^{\frac{1}{m_i}}}. \quad (2.8)$$

Некоторые специальные случаи оценки эффективности субполосного кодирования рассматриваются ниже.

Предположим, что входной сигнал $x(n)$ является гауссовым (отсчеты имеют гауссову функцию плотности вероятности). Тогда известно, что выход линейной системы с гауссовым входом также является гауссовым [10]. Таким образом, все субполосные сигналы будут гауссовыми, что приведет к $\varepsilon = \varepsilon_0 = \varepsilon_1 = \ldots \varepsilon_{M-1}$ (при условии многобитного квантователя). Следовательно, эффективность кодирования составит

$$G = \frac{\varepsilon\sigma_x^2}{\left(\varepsilon^{\frac{1}{m_0}+\cdots+\frac{1}{m_{M-1}}}\right)\prod_{i=0}^{M-1}(\sigma_i^2 n_i)^{\frac{1}{m_i}}} = \frac{\sigma_x^2}{\prod_{i=0}^{M-1}(\sigma_i^2 n_i)^{\frac{1}{m_i}}}.$$

Пусть коэффициенты децимации будут равны $m_o = m_1 = \ldots m_{M-1}$. Из уравнения (1.1) следует, что каждый коэффициент равен M и эффективность субполосного кодирования составит

$$G = \frac{\varepsilon\sigma_x^2}{[\prod_{i=0}^{M-1}(\sigma_i^2 n_i)]^{\frac{1}{M}}}.$$

В случае ортогонального банка фильтров $n_k = 1$ как было обозначено ранее, эффективность субполосного кодера будет определяться как

$$G = \frac{\varepsilon\sigma_x^2}{\prod_{i=0}^{M-1}(\varepsilon_i\sigma_i^2)^{\frac{1}{m_i}}}. \qquad (2.9)$$

Далее, свойство отсутствия потерь, обозначенное ранее, может быть применено к матрице полифазного анализа (1.15). Так как входом полифазного анализа является $x(n)$, то усреднённая дисперсия входа будет равна σ_x^2. Пусть σ_k^2 обозначает дисперсию выхода $k = 0, \ldots, M-1$ субполос, тогда на основании свойства отсутствия потерь следует, что

$$\sigma_x^2 = \sum_{k=0}^{M-1}\frac{\sigma_k^2}{m_k}. \qquad (2.10)$$

Подставляя равенство (2.10) в соотношение (2.9), получаем, что эффективность субполосного кодирования может быть выражена следующим образом:

$$G = \frac{\varepsilon\sum_{i=0}^{M-1}\frac{\sigma_i^2}{m_i}}{\prod_{i=0}^{M-1}(\varepsilon_i\sigma_i^2)^{\frac{1}{m_i}}}. \qquad (2.11)$$

Анализ (2.11) показывает, что более простое выражение оценки эффективности субполосного кодирования получается при одновременном рассмотрении следующих допущений [11–15]: входной сигнал имеет нормальный закон распределения, коэффициенты децимации в каналах равны, банк фильтров — ортогональный. При этом эффективность субполосного кодирования определяется отношением среднего арифметического к средне геометрическому неотрицательных величин σ_i^2:

$$G = \frac{\frac{1}{M}\sum_{i=0}^{M-1}\sigma_i^2}{(\prod_{i=0}^{M-1}\sigma_i^2)^{\frac{1}{M}}}. \qquad (2.12)$$

Так как среднее арифметическое больше или равно среднему геометрическому, следовательно, эффективность субполосного кодирования $G \geq 1$, т.е. субполосный кодер обеспечивает равное или лучшее качество по сравнению с полнополосным кодером при любом входном сигнале. Величина, обратная выражению (2.12), также известна как мера пологости спектра (SFM) [16, 17]. Часто в задачах кодирования речи используется взвешенная на порог маскирования величина SFM (перцептуально взвешенная SFM — PSFM) [17, 18].

3. Применение в речевых технологиях субполосной обработки сигналов

3.1. Структура кодера-редактора шумов речевого сигнала

Предлагается комбинированная система редактирования шумов и кодирования речевого сигнала без специального процессора повышения качества речи на основе критического дерева пакета дискретного вейвлет преобразования (ПДВП) $CB - WPD: (l,n) \in E_{cv}, l = \overline{0,5}$ *(рис. 3)* и вычисления порога восприятия речевого сигнала человеком. Разработка ориентирована на частоту дискретизации 8кГц и обработка введётся в 17 барках [19].

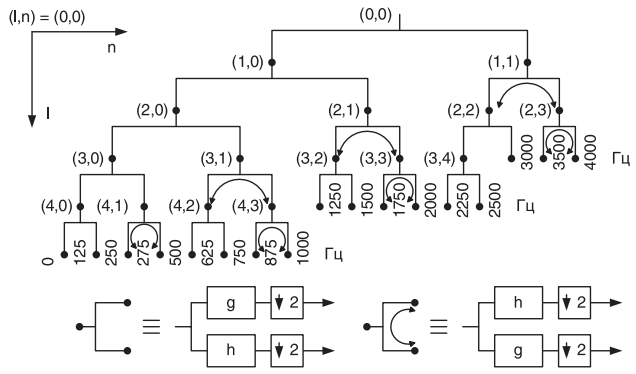


Рис. 3. Структура критического дерева ПДВП

Предполагается, что чистый речевой сигнал $x(t)$ и окружающий его шум $n(t)$ статистически независимы и стационарны в широком смысле (где t обозначает временной индекс). Зашумленный временной сигнал $y(t) = x(t) + n(t)$ преобразуется в вейвлет область на основе ПДВП. Вейвлет коэффициенты зашумленного речевого сигнала определяются следующим образом:

$$Y_{l,n}(k) = \langle y, \psi_{l,n,k} \rangle, (l,n) \in E_{CB}, k \in \mathrm{Z}. \qquad (3.1)$$

где k временной индекса вейвлет коэффициента в субполосе обработки (l, n).

На *рис. 4* показана схема обработки речевого сигнала в одной из ветвей $CB - WPD: (l,n) \in E_{cv}, l = \overline{0,5}$ (соответствующей ей частотной полосе *(рис. 3)*) комбинированной системы редактирования шума и кодирования речевого сигнала. Оценка порогов маскирования выполняется в вейвлет области в соответствии с алгоритмом, показанным в [20]. В данной работе используется наиболее общее психоакустически мотивированное правило спектрального взвешивания [21]. Вейвлет коэффициенты $\hat{X}_{l,n}(k)$, отредактированные от шума входного сигнала, поступают на схему кодирования и квантования и далее формируется пакет данных

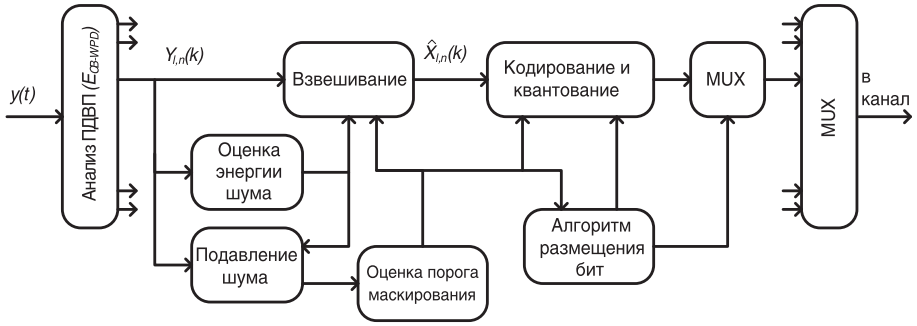


Рис. 4. Структура кодера-редактора шумов речевого сигнала на базе ПДВП $CB - WPD: (l,n) \in E_{cv}, l = \overline{0,5}$

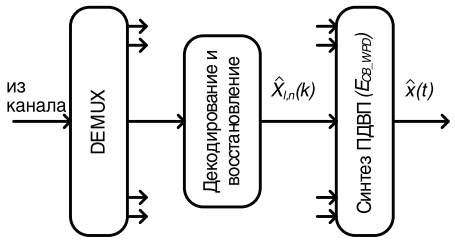


Рис. 5. Структура декодера речевого сигнала на базе ПДВП

для передачи в канал. Схема кодирования и квантования реализована в соответствии с [22].

Структура декодера показана на *рис. 5*, где закодированные вейвлет коэффициенты $\hat{X}_{l,n}(k)$ декодируются и восстанавливаются в каждой субполосе (l, n). Синтез сигнала выполняется на основе обратного ПДВП в соответствии со структурой дерева $^{E}SB_{WPD}$:

$$\hat{x}(t) = \sum_{(l,n)\in E, k\in Z} \hat{X}_{l,n}(k)\hat{\psi}_{l,n,k}(t), \qquad (3.2)$$

где $\{\hat{\psi}_n(e): n \in Z_+\}$ — множество ортогональных вейвлет функций ПДВП $\{\psi_n(e): n \in Z_+\}$, где $\{\hat{\psi}_{l,n,k}(t) = 2^{-l/2}\hat{\psi}_n(2^{-l}t - k)$.

3.2. Эксперимент

Для кодера со скоростью передачи 4–6 кбит/с экспериментальные результаты приведены на *рис. 6:* (а) чистый речевой сигнал с частотой дискретизации $f_s = 8$ кГц и его спектрограмма, (б) зашумленный речевой сигнал с $SNR = 5$ дБ и его спектрограмма, (в) отредактированный речевой сигнал от шума и его спектрограмма, (г) реконструированный речевой сигнал декодером и его спектрограмма.

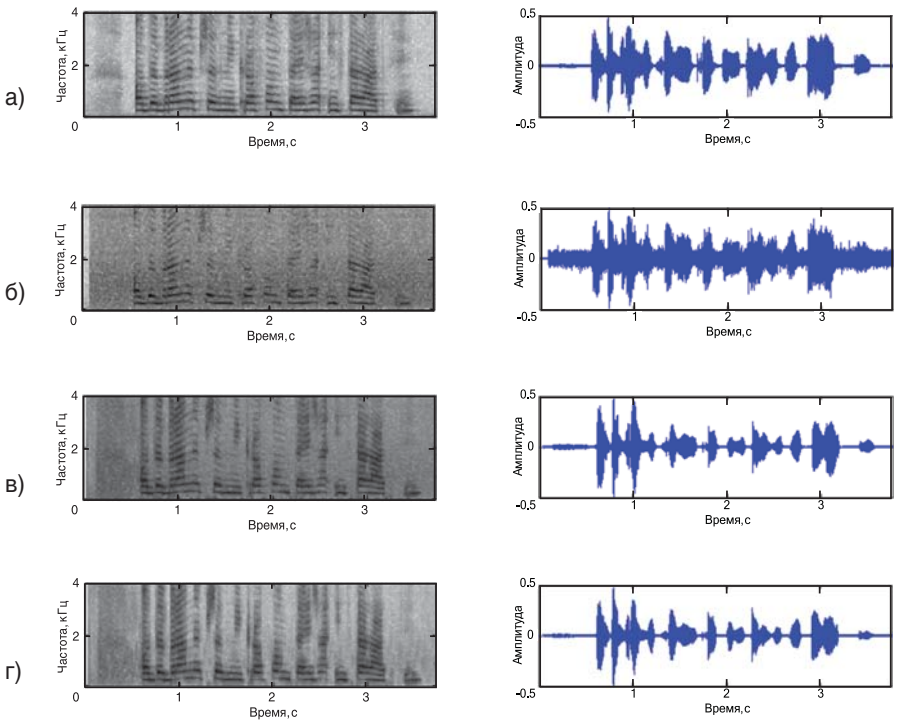


Рис. 6. Результаты обработки речевого сигнала в кодере-редакторе шумов: а) чистый речевой сигнал; б) зашумленный речевой сигнал; в) отредактированный речевой сигнал от шума; г) реконструированный речевой сигнал декодером

Достоинством данной системы субполосной обработки речевых сигналов является возможность комбинирования двух процессов перцептуальной обработки в субполосах: редактирование акустического шума в речевом сигнале и последующая его компрессия. Обработка ведётся в области вейвлет коэффициентов, причём порог маскирования расчитывается один раз для обоих процессов.

Литература

1. *Vaidynathan P.P.* Multirate systems and filter banks, Prentice Hall: Englewood Cliffs, 1993.
2. *Crochiere R.E., Rabiner L.* Multirate digital signal processing, Prentice-Hall, Engle-wood Cliffs, NJ, USA, 1983.
3. *Витязев В.В.* Цифровая частотная селекция сигналов // Радио и связь. М., 1993.
4. *Piotrowski A., Parfieniuk M.* Cyfrowe banki filtrow: analiza, synteza i implementacja dla systemow multimedialnych, Politechnika Bialostocka, Bialystok, 2006.
5. *Vaughan R.G., Scott N.L., White D.R.* The theory of bandpass sampling, IEEE Trans. Signal processing, 1991. V. 39. №. 9. P. 1973–1984.
6. *Parfieniuk M., Petrovsky A.* Simple rule of selection of subsampling ratios for warped filter banks, in Proc. VIII Int. conf. «Modern communication systems», Naroch, Belarus, 2003. P. 130–134. Special Issue of Trans. Belarussian Engineer Academy, № 1(15)/3.
7. *Sathe V.P., Vaidynathan P.P.* Effects of multirate systems on the statistical properties of random signals. IEEE Transactions Signal Processing, 41(1), January 1993. P. 131–146.
8. *Jaynt N.S., Noll P.* Digital coding of waveforms, Prentice Hall: Englewood Cliffs, 1984.
9. *Uzun N., Haddad R.A.* Cyclostationary modeling, analysis, and optimal compensation of quantization errors in subband codecs. IEEE Transactions Signal Processing, 43(9), September 1995. P. 2109–2119.
10. *Papoulis A.* Probability, random variables, and stchastic processs, McGraw-Hill: Tokyo, 1984.
11. *Soman A.K., Vaidynathan P.P.* Coding gain in paraunitary analysis/synthesis systems. IEEE Transactions Signal Processing, 41(5), May 1993. P. 1824–1835.
12. *Djokovic I., Vaidynathan P.P.* On optimal analysis/synthesis filters for coding gain maximization. IEEE Transactions Signal Processing, 44(5), May 1996. P. 1276–1279.
13. *Calvagno G., Mian G.A., Rinaldo R.* Computation of the coding gain for subband coders. IEEE Transactions Communication, 44(4), April 1996. P. 475–487.
14. *Kok C.W., Nguyen T.Q.* Multirate filter banks and transform coding gain. IEEE Transactions Signal Processing, 46(7), July 1998. P. 2041–2044.
15. *Gosse K., Duhamel P.* Perfect reconstruction versus MMSE filter banks in source coding. IEEE Transactions Signal Processing, 45(9), September 1997. P. 2188–2202.
16. *Spanias A., Painter T., Atti V.* Audio signal processing and coding, Wiley-Interscience, NJ, USA, 2007.
17. *Bosi M., Goldberg R.E.* Introduction to digital audio coding and standards, Springer Science+Business Media, USA, 2003.
18. *Петровский А.А., Белявский К., Петровский Ал.А.* Перцептуальное кодирование аудио и речевых сигналов: Доклады БГУИР. 2004. № 1(5). С. 73–91.
19. *Petrovsky A.A., Bielawski K., Petrovsky Al.A.* Combined system for acoustic echo and Noise reduction based on the psychoacousticaly motivated multirate filter bank // Mittweida, Germany: IWKM, 2000, Journal of the University of Applied Sciences Mittweida. P. 33–41.
20. *Петровский А.* Построение психоакустической модули в области вейвлет-коэффициентов для перцептуальной обработки звуковых и речевых сигналов // Речевые технологии. 2008. № 4. С. 61–71.
21. *Петровский Ал.А., Боровић А., Парфенюк М.* Дискретное преобразование Фурье с неравномерным частотным разрешением в перцептуальных системах редактирования шума в речи // Речевые технологии. 2008. № 3. С. 16–26.
22. *Петровский Ал.* Перцептуальный кодер звука на базе вейвлет преобразования с динамической трансформацией частотно-временного плана // Цифровая обработка сигналов. 2009. № 4. С. 48–58.

Петровский Алексей Александрович —

кандидат технических наук, доцент. Работает в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники», кафедра Электронных вычислительных машин.
Закончил учреждение образования «Белорусский государственный университет информатики и радиоэлектроники», специальность — «Проектирование и технология электронных вычислительных средств». Сфера интересов — цифровая обработка сигналов: многоскоростная обработка, анализ/синтез банков фильтров, проектирование проблемно-ориентированных средств вычислительной техники реального времени для систем мультимедиа.
Член общества AES.

Комплексная система верификации ключевых слов на основе метода опорных векторов

А.М. Сорока,
БГУ, Минск, Беларусь

Одной из основных сложностей, влияющих на точность методов поиска ключевых слов (ПКС), остаётся проблема декодирования акустически схожих пар спутывания, точность решения которой в значительной степени влияет на общую эффективность системы [1]. Это свидетельствует о необходимости дополнительного анализа пар спутывания, представляющих конкурирующие версии. Кроме этого, при использовании оценок апостериорных вероятностей для принятия решений в системах ПКС высокая точность поиска сопряжена с высоким уровнем ложных тревог, для уменьшения которого необходимо использовать специализированные алгоритмы верификации. В статье рассмотрен алгоритм верификации потенциальных ключевых слов на основе метода опорных векторов (МОВ).

Основные методы верификации речевых данных

В системе поиска ключевых слов неизбежны ошибки, связанные в первую очередь с особенностями и функциональным состоянием диктора, шумами окружающей среды, помехами и т.д. Использование решётки слогов позволяет проанализировать максимальное количество вариантов последовательностей, содержащих искомые ключевые слова, и тем самым максимизировать вероятность правильного обнаружения [2]. Однако такие процедуры приводят к увеличению количества ложных тревог. Использование алгоритмов верификации найденных ключевых слов позволит в значительной степени снизить чувствительность системы к словам, не входящим в словарь, т.е. уменьшить вероятность ложной тревоги.

Существует три наиболее распространённых метода верификации:

1) Метод, основанный на использовании эталонов. Метод применяется только в условиях, когда ключевые и «неключевые» слова фиксированы на этапе создания системы. Любое изменение словаря и условий использования приводит к необходимости

полного переобучения системы, что неприемлемо для большинства практических приложений.

2) Статистический метод. Эффективность метода сильно зависит от выбранного признака верификации и классификатора. Наиболее распространённые признаки — соотношение правдоподобия, апостериорная вероятность слова, акустическая вероятность нормализации; параметры лингвистической модели, акустическая устойчивость, контекст ключевых слов. В качестве классификаторов могут использоваться байесовские классификаторы, нейронные сети и СММ.

3) Ассоциативный метод на основе статистических данных и шаблонов правил. Пример такого подхода — алгоритм TBL, предложенный Э. Брилем [3].

Наиболее распространённый и эффективный — подход верификации данных с использованием статистических моделей. Главная причина успешного использования СММ — существование итеративных методов, гарантирующих сходимость оценки параметров. Однако поскольку СММ генерируется на основе традиционной статистической теории распределения вероятностей, СММ может корректно описывать модель только в том случае, если есть достаточное количество обучающих данных. Кроме этого, СММ позволяет точно идентифицировать данные разных типов только тогда, когда в пространстве выборок область распределения данных различных типов или не перекрывается, или перекрывается незначительно.

Искусственные нейронные сети представляют интересный и важный класс классификаторов, успешно использованный для анализа и распознавания речи. Применение ИНС позволяет преодолеть многие недостатки, свойственные СММ, однако использование ИНС для задач верификации имеет ряд недостатков. Наиболее значимые недостатки — сложность определения оптимальной топологии модели, медленная сходимость во время обучения и тенденция к чрезмерной адаптации данных.

С точки зрения верификации ключевых слов в динамических условиях наиболее перспективным выглядит классификатор на основе МОВ. Даже при небольшом количестве обучающих данных, при помощи МОВ в пространстве признаков можно найти самую оптимальную гиперплоскость для создания классификатора, что обуславливает её широкое применение на настоящий момент. МОВ также эффективно применяется для задач распознавания речи, изображений и др. Впервые для обработки речи МОВ был использован А. Ганапавираем в 1998 г. В его работе в качестве предварительного препроцессора для сегментации по фонемам была использована СММ, затем производилось принудительное выравнивание по фиксированной длине и масштабу (3:4:3), а для последующего распознавания речи была использована гибридная модель на основе СММ и МОВ [4]. Результаты применения такой системы показали, что она эффективнее, чем изолированная модель СММ.

Наиболее удачны бинарные МОВ, способные разделять данные сложной и схожей конфигурации на два класса, что делает их весьма эффективными для решения задач верификации и идентификации близкорасположенных данных. В связи с этим в данной статье предлагается использовать МОВ для верификации ключевых слов в системах ПКС.

Верификация ключевых слов с использованием меры достоверности на интервале

Представим речевой сигнал на входе системы поиска ключевых слов в виде последовательности наблюдений $O = \{O_1, O_2, \ldots, O_T\}$. Поскольку ключевые слова уже найдены, то известно начало и конец каждого слова, и параметры их СММ могут быть оценены на основе решётки. Обозначим такую СММ ключевого слова, как Δ, и определим меру достоверности как количественную величину совпадения O и Δ. Другими словами, мера достоверности определяется вероятностью генерации последовательности O на основе модели Δ.

Для моделирования речевых сигналов на основе СММ компоненты вектора признаков предполагаются независимыми, и для каждого состояния плотность распределения вероятностей наблюдений можно представить смесью гауссовых распределений, задающей вероятность вектора наблюдений O_t в состоянии j как:

$$b_j(o_t) = \prod_{d=1}^{D}\left[\sum_{m=1}^{M_s} c_{jdm} N\left(o_{dt}; \mu_{jdm}; \sigma_{jdm}\right)\right],$$

где D — размерность вектора признаков речевого сигнала; M_s — размерность ГС для компоненты d; c_{jdm}, σ_{jdm}, μ_{jdm} — соответственно весовое значение, среднее значение и дисперсия для m-й компоненты смеси, предположительно описывающуюся нормальным распределением:

$$N(o_{dt}; \mu_{jdm}; \sigma_{jdm}) = \frac{1}{\sqrt{2\pi}\sigma_{jdm}} e^{-(o_{dt}-\mu_{jdm})^2 \Big/ 2\sigma_{jdm}^2}$$

Пусть j-е состояние модели Δ определяется участком последовательности $O^k = \{O_k, O_{k+1}, \ldots, O_K\}$, тогда значение меры достоверности CM_j можно определить следующим образом:

$$CM_j = \frac{\sum_{t=k}^{K}\sum_{m=1}^{M_s}\sum_{d=1}^{D} F\left(o_{dt}; \mu_{jdm}, \sigma_{jdm}\right)}{(K-k)M_d S},$$

где

$$F(o_{ti}; \mu_{jdm}, \sigma_{jdm}) = \begin{cases} 1, & (o_{dt} - \mu_{jdm}) \in [\mu_{jdm} - k\,\sigma_{jdm}, \mu_{jdm} + k\,\sigma_{jdm}] \\ 0, & \check{c}\acute{\imath}\acute{r} \div \acute{l} \end{cases},$$

где k — управляющий интервалом достоверности параметр. Определим меру достоверности для каждого ключевого слова Δ путём нормализации значений для каждого состояния:

$$CM_1 = \frac{1}{N}\sum_{j=1}^{N} CM_j \quad \text{, где } N \text{ — количество состояний СММ.}$$

Верификация ключевого слова может происходить путём сравнения полученной меры достоверности с некоторым пороговым значением, как правило, выбранным эмпирически.

Верификация ключевых слов на основе нормализации длительности состояния

Один из недостатков меры достоверности, представленной выше, — отсутствие её нормализации на длину состояния СММ, вследствие этого возможны ситуации, когда

состояние с малой длительностью затеняет результаты для более длительной последовательности наблюдений.

Пусть есть СММ ключевого слова $\lambda = \{N_1, \pi, A, B\}$, где N — число состояний СММ $S = \{S_1, S_2, \ldots S_N\}$, π — матрица начальных вероятностей, A и B — матрицы вероятностей переходов и наблюдений соответственно. Обозначим время входа и выхода системы из состояния i как $b[i]$ и $e[i]$ соответственно, тогда нормализованную меру достоверности можно определить следующим образом:

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

Согласно формуле Байеса заменив $\log p(o_t \mid s_i)$ на $\log p(s_i \mid o_t)$ получим ещё одно выражение для меры достоверности:

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

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

Представленные меры достоверности представляют собой среднее значение акустической вероятности в рамках СММ [5].

Верификация ключевых слов 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 [6].

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

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

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

$$L_\Lambda(o_t) > L_{j_2}(o_t) > \ldots > L_{j_k}(o_t) > \ldots > L_{j_{N(o_t)}}(o_t)$$

Пусть характеристическое значение $L_{Kw}(o_t) = L_{jk}(o_t)$ для модели ключевого слова Δ_{Kw} занимает в отсортированной последовательности позицию k, тогда динамический рейтинг на O_t-м фрейме можно определить 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

$$G(L_k(o_t) - L_{Kw}(o_t)) = \begin{cases} 0, & L_k(o_t) \le L_{Kw}(o_t) \\ 1, & \check{c}\acute{i}\,\acute{r} \div \acute{l} \end{cases}$$

Введём меру достоверности на основе динамического рейтинга как обобщённое характеристическое значение в рамках всей анализируемой длительности сигнала:

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

Этот метод пороговый, при этом для всех ключевых слов может быть установлено одинаковое пороговое значение.

Вычисление динамического рейтинга для ключевых слов не приводит к существенному увеличению вычислительной сложности системы по сравнению с алгоритмами на основе акустической меры достоверности. В процессе декодирования Витерби значение вероятностей для выражения (1) уже рассчитаны, поэтому вычисление $Q(o_t)$ и последующая нормализация приводят к дополнительным DT сложениям и вычитаниям, а также $T+1$ делению. Однако этот алгоритм обладает рядом существенных преимуществ. Во-первых, эффективность верификации ключевых слов на основе динамического рейтинга имеет более стабильный характер, особенно в условиях шума. Это обусловлено тем, что имеющийся в речевых данных шум влияет только на абсолютные значения акустических вероятностей и не оказывает практически никакого влияния на расположение характеристических значений при сортировке. 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 выше, для всех ключевых слов может быть установлено одинаковое пороговое значение, и кроме этого, наличие шума в исходных данных не влияет на абсолютное значение порога.

Верификация ключевых слов на основе машины на опорных векторах

Все методы верификации ключевых слов, основанные на моделях фонем и мерах достоверности, обладают общими недостатками. Во-первых, эти методы пороговые, процедура определения порога имеет, как правило, эмпирический характер и сопряжена с проведением многочисленных экспериментов. Во-вторых, алгоритм верификации ключевых слов на основе единичной меры доверительности, как и любая пороговая система, обладает ограничениями по улучшению его характеристик. При сохранении высокой точности распознавания вероятность ложной тревоги может быть уменьшена только до какого-то фиксированного значения, характерного для конкретной меры доверительности.

Для решения проблемы можно использовать алгоритмы верификации ключевых слов на основе нескольких мер достоверности одновременно. Существует много методов объединения нескольких мер доверительности, например, логическое соединение, линейный дискриминант Фишера, метод нейронной сети и т.д. Предположим, что каждому ключевому слову поставлено в соответствие u мер доверительности $CM_1, CM_2, \ldots CM_u$, каждой из которых соответствует своё пороговое значение $TH_1, TH_2, \ldots TH_u$. В простейшем случае значения мер достоверности $cm_1, cm_2, \ldots cm_u$ сравниваются с соответствующими порогами, и если $cm_1 > TH_1, cm_2 > TH_2, \ldots cm_u > TH_u$, то это слово определяется как правильно распознанное, иначе принимается решение об ошибке распознавания. Введение

различных весовых коэффициентов позволяет гибко настраивать систему верификации, однако улучшение характеристик такой системы тоже ограничено, поскольку данный метод также пороговый. Для преодоления этих недостатков при построении комплексной системы верификации ключевых слов был использован классификатор на основе МОВ, обладающий значительными достоинствами при объединении мер достоверности.

Для верификации ключевых слов определим два класса. Если выбранный речевой фрагмент — ключевое слово, он относится к классу 1 ($y_1=+1$), если фрагмент представляет ложную тревогу — то к классу 2 ($y_1=-1$). В качестве входных параметров класса выберем меры доверительности CM_1, CM_2,... CM_u. При обучении системы одна из важных процедур — определение значение параметра c, а также выбор целевой функции f. В процессе обучения для ключевых и неключевых слов было предложено ввести различные весовые коэффициенты для c: параметры bC и $(1-b)C$ соответственно. В качестве целевой функции предложено использовать:

$f=\alpha k_1-(1-\alpha)k_2$, где k_1 — мера правильной классификации первого класса, k_2 — мера правильной классификации второго класса, $0,5<\alpha<1$ — пороговое значение.

Алгоритм верификации ключевых слов на основе мер достоверности с использованием МОВ дополнительно требует $O(mn^u)$ вычислений, где m — количество ключевых слов в словаре, n — размер выборки для обучения, u — количество мер достоверности. Поскольку число мер достоверности при верификации ключевых слов невелико, то вычислительная сложность алгоритма МОВ также достаточно невелика и лежит в пределах, позволяющих разрабатывать системы поиска в реальном масштабе времени. При правильном выборе функции ядра метод верификации ключевых слов на основе МОВ способен обеспечить высокую точность.

Гибридная система верификации ключевых слов

В результате декодирования на основе решётки и сети спутывания система ПКС генерирует оценки различных мер достоверности, сравнение которых с порогом позволяет принять решение о наличии или отсутствии искомых ключевых слов в потоке слитной речи. Однако использование такого подхода приводит к высокому уровню ложных тревог [8]. Уменьшение этого уровня за счёт изменения порога приводит к снижению вероятности правильного обнаружения.

Для устранения этого недостатка, присущего всем пороговым классификаторам, был использован гибридный подход на основе МОВ для верификации ключевых слов, позволяющий принимать многокритериальные решения. Этот подход реализован следующей структурой системы ПКС *(рис. 1)*.

Отличительная особенность предложенной структуры — то, что на выходе сети спутывания оцениваются меры достоверности и значений апостериорной вероятности, из анализа которых принимается решение о наличии пары спутывания. Если такая пара спутывания обнаруживается, то для вычисления апостериорной вероятности слогов спутывания используется гибридная модель [2].

Далее значения мер достоверности и апостериорных вероятностей поступают на вход блока многокритериального принятия решений, реализованного на основе МОВ.

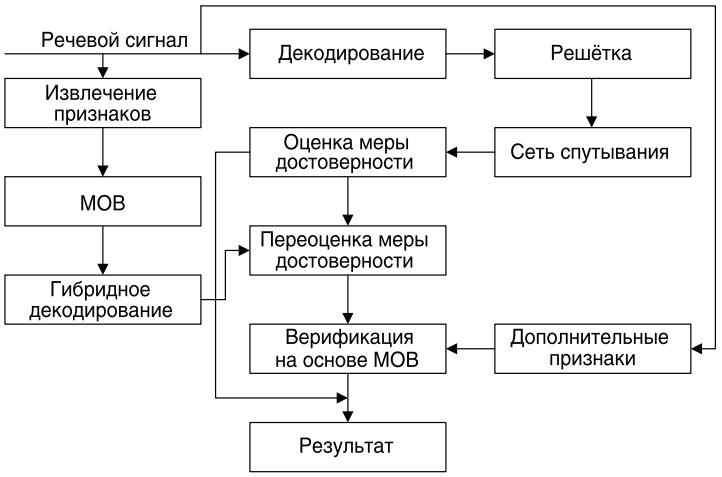


Рис. 1. Структура гибридной системы поиска ключевых слов

Алгоритм верификации ключевых слов на основе гибридной системы СММ-МОВ состоит из следующей последовательности действий:

1) Речевые данные поступают на систему декодирования, которая генерирует решётку слогов, которая затем преобразуется в сеть спутывания.

2) В каждом множестве спутывания фиксируются три слога с максимальными вероятностями и оценивается мера достоверности потенциального ключевого слова. Если полученная мера достоверности больше порогового значения, то происходит переход к этапу 4.

3) Создаётся объединённый вектор признаков, с помощью МОВ классификатора идентифицируются слоги спутывания, для которых переоценивается апостериорная вероятность.

4) Верификация ключевых слов.

Экспериментальное исследование

Для экспериментального исследования была составлена база речевых сигналов, содержащая записи русскоязычных интернет-радиостанций, общим объёмом 11.4 Гб. Для верификации ключевых слов использован классификатор на основе МОВ с ядром в виде гауссовой радиальной функции. Наилучшие характеристики верификации обеспечиваются при использовании следующих параметров классификатора на основе МОВ: $b = 0{,}96$, $\alpha = 0{,}6$ или $b = 0{,}99$, $\alpha = 0{,}6$. Для этих значений параметров проведено экспериментальное исследование по определению эффективности верификации на основе МОВ, объединяющей три различные меры доверительности — динамический рейтинг, нормализацию состояний и достоверность на интервале, так и для каждой меры доверительности в отдельности. Полученный результат приведён в *таблице 1*.

Таблица 1

Эффективность предложенного метода верификации ключевых слов

Pd,%	*FAR,%*			
	МОВ	Динамический рейтинг	Нормализация состояний	Достоверность на интервале
85.7	15.1	21.2	35.4	43.3

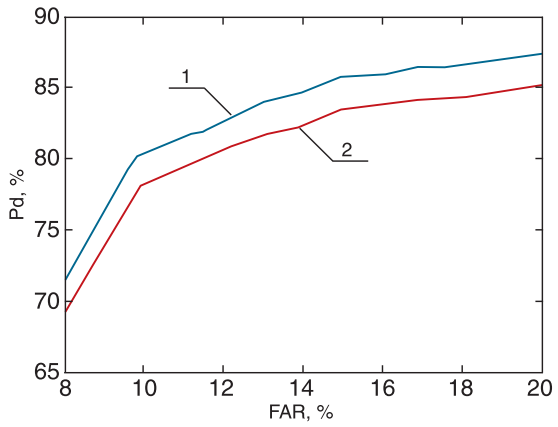


Рис. 2. Эффективность системы ПКС с использованием блока верификации ключевых слов (1) и без него (2)

Как видно из *таблицы 1*, для одних и тех же значений точности поиска Pd использование МОВ позволяет добиться меньших значений вероятности ложных тревог.

Разработанный на основе МОВ блок классификации использован для верификации ключевых слов в качестве дополнительной процедуры, на вход которой подаются результаты поиска ключевых слов на основе сети спутывания. На *рис. 2* представлена эффективность системы ПКС на основе сети спутывания, с дополнительным блоком верификации и без него.

Как видно из анализа рабочих характеристик, использование дополнительно блока верификации ключевых слов позволяет снизить вероятность ложной тревоги на 4.9% в зависимости от порога, используемого в сети спутывания.

Заключение

В статье рассмотрен метод верификации ключевых слов на основе метода опорных векторов, с использованием гибридной модели декодирования пары спутывания. Результаты экспериментального исследования показали, что использование блока верификации ключевых слов позволяет снизить уровень ложных тревог в среднем на 4.9% при фиксированной точности обнаружения.

Литература

1. *Рабинер Л.Р., Шафер Р.В.* Цифровая обработка речевых сигналов. М.: Радио и связь, 1981. С. 113–119.
2. *Сапожков М.А., Михайлов В.Г.* Вокодерная связь. М.: Радио и связь,1983. С. 156–158.
3. *Дегтярев Н.П.* Параметрическое и информационное описание речевых сигналов. Минск: Объединённый институт проблем информатики Национальной академии наук Беларуси, 2003. С.62–63.
4. *Лобанов Б.М., Цирульник Л.И.* Компьютерный синтез и клонирование речи. Минск: Белорусская наука, 2008. С.60-63.
5. *Быков Н.М. и др.* Надёжный метод выделения слоговых сегментов в речевом сигнале // Автоматика и информационно-измерительная техника. 2007. № 1.

Сорока Александр Михайлович —

научные интересы — методы и алгоритмы обработки цифровых сигналов, теория метода опорных векторов, смешанные гауссовы модели.

Алгоритм двухэтапного распознавания фонем русского языка

А.М. Сорока,
БГУ, Минск, Беларусь

Одним из основных подходов к решению задачи распознавания слитной речи является метод распознавания на основе классификации минимальных речевых единиц. Как правило, в качестве минимальных речевых единиц выбираются фонемы либо дифоны, в силу наилучшего соотношения размеров словаря минимальных речевых единиц и точности распознавания. Однако признаковые описания акустических реализаций фонем крайне неравномерно распределены в пространстве признаков. При этом различие между близкорасположенными реализациями нивелируется значительным различием между остальными реализациями. Такие близкорасположенные пары минимальных речевых единиц получили название «пар спутывания».

Введение

Существует несколько методов разрешения пары спутывания, которые основаны на построении лингвистических решёток и сетей спутывания, а также использования лингвистических моделей, учитывающих контекстные связи [1]. Эти методы обладают рядом очевидных недостатков, в числе которых — высокая трудоёмкость алгоритмов и необходимость создания лингвистической модели. В статье предлагается алгоритм двухэтапного распознавания фонем на основе метода опорных векторов с построением признакового описания на основе вейвлет-преобразования, который позволяет избежать чрезмерного возникновения пар спутывания за счёт более точной классификации отдельно взятой фонемы.

Метод опорных векторов

Метод опорных векторов (МОВ) впервые был предложен Вапником [2]. Этот метод в процессе обучения непрерывно минимизирует эмпирический риск. Использование Вапником в качестве эвристики выбора разделяющей гиперплоскости

предположения о минимизации ожидаемого риска путём максимизации отступов классов привело к высокой обобщающей способности алгоритма. В настоящее время МОВ успешно используется во многих областях.

Предположим, что у нас имеется множество объектов X, заданных при помощи n-мерных вещественных векторов x, где $x \in R^n$ и множество классов $Y = \{-1+1\}$. Объекты, для которых известно точное соответствие между признаковым описанием и классом, называются прецедентами. Множество прецедентов, используемых для настройки классификатора, называется обучающей выборкой, а сам процесс настройки — обучением классификатора.

Построим линейный пороговый классификатор:

$$a(x) = sign(\sum_{j=1}^{n} w_j x_j - w_0) = sign(< w, x > - w_0), \qquad (1)$$

где $w = (w^1, \ldots, w^n) \in R^n$, $w_0 \in R$. Уравнение $< w, x > = w_0$ задаёт разделяющую гиперплоскость, при этом w — вектор нормали к данной гиперплоскости, w_0 — расстояние от гиперплоскости до начала координат.

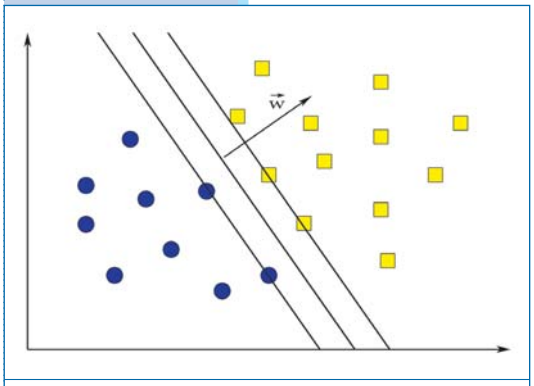


Рис. 1. Случай линейной разделимой обучающей выборки

Случай, при котором прецеденты линейно разделимы в признаковом пространстве, показан на *рис. 1*.

Рассмотрим функционал ошибок:

$$Q(w, w_0) = \sum_{i=1}^{l} [y_i(< w, x_i > - w_0) < 0] \qquad (2)$$

Если существует такая разделяющая гиперплоскость $<w,x> = w_0$, что функционал (2) обращается в ноль, следовательно, множество объектов X является линейно разделимым на два класса. Очевидно, что в таком случае существует бесконечное число разделяющих гиперплоскостей. Вапником введено [2] понятие оптимальной разделяющей гиперплоскости — такой гиперплоскости, которая максимально удалена от границ обоих классов. Алгоритмы построения оптимальной разделяющей гиперплоскости с использованием метода Лагранжа могут быть найдены в специальной литературе [Ч]. Итоговый вид классификатора в данном случае может быть описан следующим выражением:

$$a(x) = sign(\sum_{i=1}^{l} \lambda_i y_i \langle x_i, x \rangle - w_0) \quad , \qquad (3)$$

где $\lambda_i \in R, i = 1 \ldots m$ — коэффициенты Лагранжа.

На практике класс задач с линейно разделимой выборкой встречается крайне редко. Для решения проблемы классификации линейно неразделимых выборок Кортес и Вапник [3] предложили метод опорных векторов с мягким зазором. Фактически, они вводят неотрицательную величину ошибки классификации. Теперь проблема оптимизации представляет задачу минимизации ошибки классификации. В таком случае оптимальная разделяющая

гиперплоскость определяется вектором w, который минимизирует следующий функционал:

$$\begin{aligned} (w \cdot x_i) + b \geq +1 - o, \quad & if \quad y_i = +1 \\ (w \cdot x_i) + b \leq -1 + o, \quad & if \quad y_i = -1 \end{aligned} \qquad (4)$$

Здесь $\xi = (\xi_1, ..., \xi_m)$ — вектор двойственных переменных, C — константа.

Этот подход не единственный для решения задачи в случае, если исходная выборка не является линейно разделимой в исходном признаковом пространстве. Предположим, что существует пространство более высокой, чем исходное, размерности, в котором исходная выборка окажется линейно разделимой (*рис. 2*). Переход от исходного пространства признаков X к новому пространству H может быть выполнен при помощи некоторого преобразования $\psi : X \to H$.

Таким образом, классификатор будет описываться следующим выражением:

$$a(x) = sign(< w, \psi(x) > - w_0), \qquad (5)$$

где (w, w_0) задают разделяющую гиперплоскость в расширенном пространстве. В таком случае итоговый вид классификатора записывается в следующем виде:

$$a(x) = sign(\sum_{i=1}^{l} \lambda_i y_i \langle \psi(x_i), \psi(x) \rangle - w_0) \qquad (6)$$

Анализируя выражение (6), можно видеть, что нет необходимости в явном виде задавать функцию отображения $\psi : X \to H$. Пусть существует функция $K(x_i, x_j)$ такая, что $K(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle$. В таком случае итоговый вид классификатора приобретает следующий вид:

$$a(x) = sign(\sum_{i=1}^{l} \lambda_i y_i K(x_i, x) - w_0) \qquad (7)$$

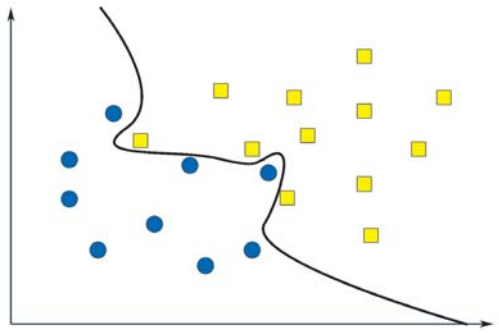


Рис. 2. Использование расширенного пространства

Функция $K(x_i, x_j)$ получила название ядра или ядерной функции. Стоит отметить тот факт, что здесь показано не единственное применение ядерной функции — данный класс функций получил широкое практическое применение.

Наиболее часто используются следующие ядерные функции:

линейная: $K(x, y) = x \cdot y$,

полиномиальная: $K(x, y) = (x \cdot y + 1)^d$, где d — степень полинома,

радиальная базисная Гауссова функция (RBF): $K(x, y) = \exp(-\frac{|x - y|^2}{2д^2})$,

где δ — ширина функции Гаусса.

Построение векторов признаков на основе вейвлет-преобразования

Для построения векторов признаков акустических сигналов в системах распознавания речи широко используются мелчастотные кепстральные коэффициенты (МЧКК) [5]. Однако, как показывают практические исследования [6], использование этого подхода не обеспечивает достаточной точности классификации акустических сигналов, что может быть обусловлено близостью векторов признаков в признаковом пространстве. В статье предложены два алгоритма извлечения векторов признаков на основе вейвлет преобразования, обладающего более высокой способностью к выделению локальных частотно-временных особенностей сигнала в сравнении с традиционным кратковременным Фурье-преобразованием.

В статье рассматриваются два алгоритма извлечения векторов признаков для речевых сигналов на основе вейвлет-анализа [6].

Первый алгоритм фундируется возможностью провести сегментацию и распознавание фонемы посредством визуального анализа графического представления результатов вейвлет-преобразования. Этот способ построения векторов признаков (ВП1) основан на методах смежной дисциплины — распознавания графических образов и может быть описан следующей последовательностью действий. Графический вейвлет образ сегментируется на участки, соответствующие одному периоду в квазипериодической трактовке вейвлет образа, далее в каждом сегменте детектируются резкие характерные изменения с использованием алгоритма детектора Харриса. Следующий шаг — нормализация координат полученных характерных точек. Для формирования вектора признаков характерные точки представляются в виде смеси двумерных Гауссовых распределений [7]:

$$p(x) = \sum_{j=1}^{K} w_j p(x \mid C^j), \qquad (8)$$

где w_j — весовой коэффициент,

$$p(x \mid C^j) = \frac{1}{(2\pi)^{-n/2} \left|\Sigma_y\right|^{-1/2}} e^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}, \qquad (9)$$

x — тестируемый вектор, C^j — предполагаемый кластер, K — количество компонент в смеси, Σ_j — диагональная матрица вида $\Sigma_j = \begin{bmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{22}^2 \end{bmatrix}$.

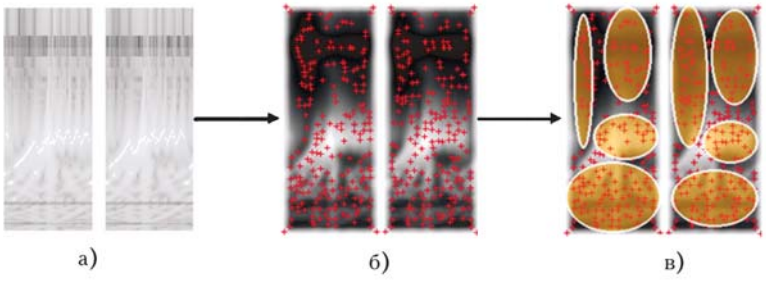


Рис. 3. Формирование вектора признаков с использованием методов анализа изображений — сегментация исходного изображения (а), нахождение характерных точек (б), аппроксимация распределения ключевых точек с использованием смеси Гауссовых распределений

Вектор признаков для заданного образа может быть описан следующим выражением:

$$x = \left(\mu_1^1, \mu_2^1, \sigma_{11}^1, \sigma_{22}^1, ..., \mu_1^K, \mu_2^K, \sigma_{11}^K, \sigma_{22}^K\right) \qquad (10)$$

Метод продемонстрирован на *рис. 3.*

Во втором случае (ВП2) для формирования вектора признаков вейвлет-образ акустического сигнала разбивается на $3N$ прямоугольных окон, в каждом из которых находится усреднённая энергия S_{ij}, $i = 1...N$, $j = 1...3$. В данном случае вектор признаков описывается следующим выражением:

$$x = \left(S_{12}, \quad ... \quad S_{N2}, \quad \Delta_1, \quad ... \quad \Delta_N\right), \qquad (11)$$

параметры $\Delta_i = S_{i3} - S_{i1}$ введены для учёта динамических процессов в начале и конце фонемы, обусловленных эффектами редукции и коартикуляции. Алгоритм представлен на *рис. 4*.

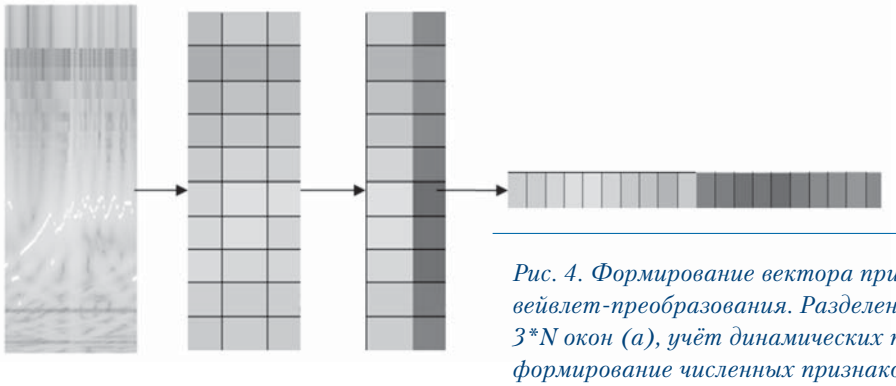


*Рис. 4. Формирование вектора признаков с использованием вейвлет-преобразования. Разделение вейвлет-образа на 3*N окон (а), учёт динамических процессов (б), формирование численных признаков (в)*

Двухэтапный метод распознавания фонем

В статье рассматривается двухэтапный метод распознавания фонем. Метод состоит из следующих этапов. На первом этапе производится классификация фонем по акустически схожим группам с использованием многоклассового классификатора на основе метода опорных векторов. Многоклассовый классификатор формируется из набора бинарных классификаторов, каждый из которых обучен по принципу «один против всех». На втором этапе производится классификация фонем внутри группы. Многоклассовый классификатор на втором этапе строится по принципу «каждый против каждого», что не влечёт за собой увеличения вычислительных затрат в силу малости групп, однако способствует более точному распознаванию.

Разделение фонем на акустически схожие группы определено эмпирическим путём на основе анализа ошибок распознавания многоклассовым классификатором. В данном случае классификатор строился по принципу «один против всех», при этом положительными прецедентами считались только реализации данной фонемы, а отрицательными — все остальные. Для каждой фонемы определялись наиболее частотные неверные результаты классификации, на основе которых делалось предположение о наличии пары спутывания. Далее была проведена глобализация пар спутывания с целью определения акустически схожих групп фонем. Итоговое разбиение фонем на акустически схожие группы представлено на *рис. 5.*

К’	Г’		К’					Й	И, Ы
К	Г		К						
		Ч		Щ					Э
				Ш	Ж				
Т’	Д’			С’	З’	Р’	Н’	Л’	А
		Ц		С	З	Р			
Т	Д						Н	Л	О
П’	Б’			Ф’	В’		М’		У
П	Б			Ф	В		М		

Рис. 5. Группы акустически схожих фонем

Экспериментальное исследование

Для определения характеристик разработанного метода было проведено экспериментальное исследование. Подготовлена база акустических реализаций фонем от разнополых дикторов на основе свободного речевого корпуса русского языка VoxForge [4] и акустической базы, подготовленной на кафедре радиофизики Белорусского государственного университета г. Минска. Общий объём базы составил 4500 фонем, в среднем по 100 реализаций каждой фонемы русского языка. Также была подготовлена тестовая выборка объёмом 100 реализаций на каждую из четырёх фонем: [а, м, н, д].

Для определения характеристик разработанных методов проведено сравнительное тестирование метода построения векторов признаков на основе мел-частотных кепстральных коэффициентов (МЧКК) и предложенных методов. Для эксперимента сформирована обучающая выборка из 4000 звуков различных фонем русского языка, из которых 700 соответствуют фонеме [а] и тестовая выборка из 300 звуковых реализаций фонемы [а].

Точность классификации с использованием алгоритмов ВП1, ВП2 и МЧКК составила 60%,82% и 80% соответственно. Для эксперимента по классификации близкорасположенных в признаковом пространстве фонем сформирована обучающая выборка из 1000 звуков гласных фонем и тестовая выборка из 100 звуков фонемы [а]. В данном эксперименте точность классификации с использованием алгоритмов ВП1, ВП2 и МЧКК составила 76%, 92% и 82% соответственно.

Оптимальные параметры алгоритма классификации определены с использованием методов кросспроверки и поиска по сетке. Вектора признаков формировались на основе алгоритма ВП2. Результаты точности классификации фонетической группы и классификации фонемы внутри группы приведены в *табл. 1*.

Таблица 1

Результаты эксперимента по точности классификации фонемы

	[а]	[м]	[н]	[д]
Точность определения группы,%	99	92	93	92
Точность определения фонемы внутри группы,%	90	94	90	94

Также проведено экспериментальное исследование точности классификации фонемы с использованием предложенного метода и одноэтапного распознавания многоклассовым классификатором. Результаты исследования приведены в *табл. 2*.

Таблица 2

Суммарная точность предложенного алгоритма и алгоритма классификации с использованием НС

	[а]	[м]	[н]	[д]
Точность определения группы,%	99	92	93	92
Точность определения фонемы внутри группы,%	90	94	90	94

Анализ результатов данных экспериментов показал, что точность предложенного алгоритма превышает точность одноэтапного алгоритма в среднем на 6%.

Заключение

В статье рассматриваются два алгоритма построения векторов признаков для акустических сигналов на основе вейвлет-преобразования. Использование первого метода (ВП1) не показало практически значимых результатов, что может быть обусловлено некорректным моделированием распределения характерных точек на вейвлет-образе. В то же время использование второго метода (ВП2) показало результаты, превосходящие результаты использования традиционно используемых методов формирования векторов признаков МЧКК на 2% при классификации фонем в общем случае и на 10% при классификации близкорасположенных в признаковом пространстве фонем.

Также в данной статье рассматривается двухэтапный метод классификации фонем русского языка на основе метода опорных векторов. Точность предложенного метода превосходит точность одноэтапного метода в среднем на 6%. Использование данного алгоритма в качестве алгоритма предварительной классификации фонем позволяет уменьшить количество пар спутывания.

Литература

1. Алиев Р.М., Янь Ц., Хейдоров И.Э. Поиск ключевых слов с использованием решётки фрагментов слов // Компьютерная лингвистика и интеллектуальные технологии: Сб. материалов ежегод. междунар. конф. «Диалог 2009», Бекасово, 27–31 мая 2009 г. / Рос. фонд фундам. исслед., Моск. гос. ун-т; Редкол.: А.Е. Кибрик [и др.]. М., 2009. С. 351–354.
2. Vapnik V. The nature of statistical learning theory [M] // New York. Springer-Verlag, 1995.
3. Cortes C., Vapnik V. Support-vector networks // Machine Learning. Vol. 20. № 3. 1995.
4. Шмырев Н.В. Свободные речевые базы данных VoxForge.org // Сб. трудов международной конференции «Диалог 2008». 2008. С. 585–588.
5. Huang X., Acero A. Spoken Language Processing: a guide to theory, algorithm, and system development. New Jersey: Prentice-Hall Inc. Upper Saddle River, 2001.
6. Siafarikas M., Mporas I., Ganchev T., Fakotakis N. Speech Recognition using Wavelet Packet Features // Journal of Wavelet Theory and Applications. 2008. V. 2. № 1. P. 41–59.
7. Rennie J. A short tutorial on using expectation-maximization with mixture models // www.ai.mit.edu/people/jrennie/writing/mixtureEM.pdf, 2004.

Параллельная архитектура системы синтеза русской речи с представлением данных в XML-формате

Киселёв В.В.,

Соломенник М.В.,

кандидат технических наук

К современным системам синтеза речи реального времени предъявляются особые требования не только по качеству звучания синтезированной речи, но и к скорости обработки больших входных данных, что может достигаться путём параллельной обработки текста на многоядерных или кластерных системах. Кроме того, архитектура синтезатора должна обладать универсальным методом обмена информацией, как между своими модулями, так и с внешними приложениями. В статье даётся анализ параллельной архитектуры системы синтеза русской речи нового поколения, основанной на обмене информацией в XML-формате и способе формирования синтезированного сигнала — Unit Selection.

Abstract

The modern real time text-to-speech (TTS) system demands not only special requirements for quality of synthesized speech sounding, but also requirements for large data amount processing speed that can be achieved by using parallel text processing on multicore processors or on computer clusters. Moreover, architecture of a synthesizer should have a universal data exchange method, to communicate between internal system modules and between engine and external applications. This paper deals with analysis of parallel architecture of new generation of Russian speech synthesis system which bases on data exchange in XML format and which uses Unit Selection technique of signal synthesis.

Введение

Системы синтеза речи в последнее время привлекают всё больше и больше внимания как исследователей, разработчиков, конечных пользователей, так и производителей крупного телекоммуникационного оборудования и других компаний, интегрирующих речевые технологии в свои разработки различного назначения. Помимо основного требования к таким системам — качество и естественность синтезированного сигнала — предъявляются требования по производительности, возможности обрабатывать большие массивы входных данных, возможности безотказной работы 24 часа в сутки. Повышение производительности работы систем с помощью параллельной обработки информации стало очень актуальным в последнее время в связи с появлением дешёвых вычислительных платформ, основанных на многоядерной архитектуре [1], а также развитии кластерных вычислений.

В первой части статьи будут рассмотрены возможности организации параллельных вычислений с конвейерной и сетевой архитектурой [2]. Далее будет рассмотрена модель внутренних данных системы, представленных в XML-формате, и в заключение рассмотрены преимущества и ограничения рассматриваемой архитектуры.

Рис. 1. Структура системы синтеза русской речи

Многомодульность системы синтеза речи позволяет частично заменять отдельные элементы, а также с помощью адаптеров использовать соответствующие элементы других систем. Структура системы, состоящая из модулей, имеющих унифицированный интерфейс обмена данными, позволяет эффективно организовать параллельную обработку данных как на одной многопроцессорной машине, так и на компьютерном кластере.

1. Организация параллельной обработки данных

Архитектура системы представляет собой структуру, представленную на *рис. 1*.

Поступающий на вход системы текст проходит предварительную обработку в одном из процессоров выбранного формата (txt, doc, rtf, odt, SSML[1], SAPI XML и т.д.), *рис. 2*.

[1] Speech Synthesis Markup Language, основанный на XML язык разметки для синтеза речи.

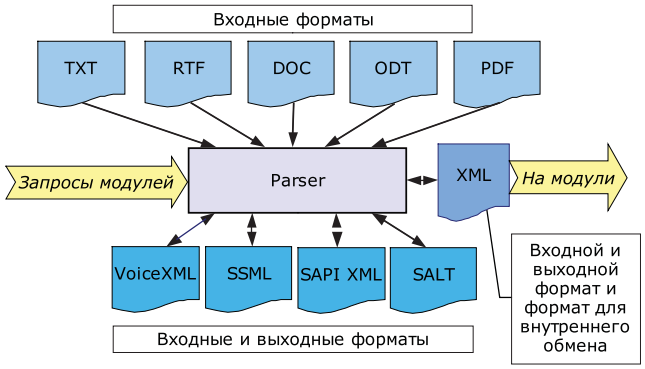


Рис. 2. Схема работы первичного модуля Parser.

Сам процесс синтеза речи разделён на четыре этапа: лингвистический, просодический, фонетический и акустический, которые обрабатывают входной поток последовательно. Кроме того, каждый этап разбит на несколько подэтапов. Обработкой данных на каждом из подэтапов занимается отдельный модуль, что позволяет организовать параллельную обработку данных на разных этапах, а также заменять каждый элемент синтеза речи другой его реализацией в процессе эксплуатации.

1.1. Конвейерная обработка данных

Модульная структура системы синтеза речи и возможность потоковой обработки данных, предоставляемая XML, а также возможность независимой обработки порции данных каждым из модулей, позволяют организовать эффективную конвейерную обработку (*рис. 3*).

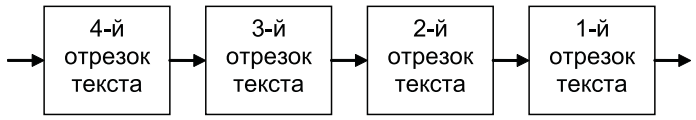


Рис. 3. Схема конвейерной обработки данных в системе синтеза русской речи

Конвейерная обработка данных может быть организована как на одной многопроцессорной платформе, так и в сети. В последнем случае возможность сетевого обмена данными включается в сами модули (например, с помощью библиотеки передачи сообщений MPI[2], которая позволяет организовать параллельную обработку как в сети, так и в многопроцессорной среде) или реализуется в виде внешних заглушек, обменивающиеся текстовой информацией со стандартными модулями и получающие от них управляющие сигналы.

1.2. Сетевая обработка данных

Сетевая обработка данных (*рис. 4*) предполагает обработку отдельных блоков текста на разных машинах, соединённых через стандартную сеть и, возможно, объединённых в блейд-сервер. В то же время отдельная машина может иметь несколько процессорных элементов, на которых текст обрабатывается конвейерным способом. Такая обработка может быть организована так же как и конвейерная обработка в случае сетевой реализации, но в этом случае необходима отдельная синхронизация отдельных частей текста, которая может осуществляться посредством дополнительных XML-тегов. Сетевая обработка данных имеет смысл, если обрабатываются большие объёмы входного

[2] Message Passing Interface, интерфейс передачи сообщений, программный интерфейс для передачи информации между компьютерами, выполняющими одну задачу.

текста и скорость обмена данными по сети заметно больше, чем скорость обработки одной части текста.

2. Обмен данными в XML-формате

В качестве формата межмодульного взаимодействия в системе синтеза речи используется XML-формат данных [3]. Анализ преимуществ XML-формата для межмодульного взаимодействия в системах синтеза речи даётся в [4].

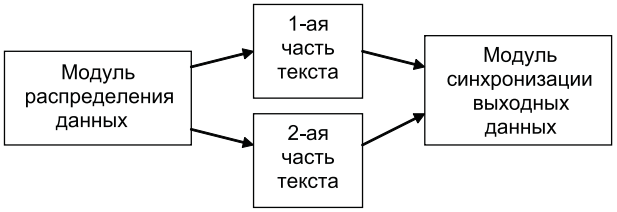


Рис. 4. Схема сетевой обработки данных в системе синтеза русской речи

2.1. Характеристики модели представления данных в XML- формате

Несомненным преимуществом XML является возможность потоковой обработки информации с использованием SAX[3]-парсера, таким образом, новые данные могут обрабатываться по мере их поступления. Некоторые ограничения накладывает специфика синтеза речи, а именно невозможность обработки информации на некоторых этапах длиной менее предложения.

Преимуществом XML является возможность сквозного прохода информации, если не требуется её изменения. Таким образом, специфические для некоторых форматов теги могут передаваться без изменений и при необходимости использоваться на последующих этапах синтеза.

Кроме того, XML-представление данных может быть легко трансформировано в любой удобный вид представления, что особенно важно на этапе отладки и экспертной оценки работы системы лингвистами.

2.2. Движение данных в системе синтеза речи

На вход системы синтеза поступает текст в формате XML, прошедший предварительную обработку и обрамлённый тегом «text».

2.2.1. Лингвистический процессор

Лингвистический процессор системы синтеза речи принимает на вход данные, полученные в результате работы предварительного процессора. В качестве выходных данных данный модуль готовит структурированную информацию о грамматических, синтаксических и семантических характеристиках элементов текста. На выходе модуля во входной XML добавляются теги: «sentence», «word», «letter», «dictitem».

[3] Simple API for XML, способ последовательного чтения XML-файлов.

2.2.2. Просодический процессор

Просодический процессор дополняет входные данные следующими элементами:
— места членения на интонационные единицы;
— длительности пауз;
— место логического ударения (интонационному центру);
— номер интонационного контура.

Данный модуль добавляет и корректирует следующие теги: «pause», «stress», «intonation».

2.2.3. Фонетический процессор

Фонетический процессор дополняет входные данные следующими элементами:
— транскрипцией;
— расставленными индексами гласных в соответствии со степенью редукции;
— аллофонами.

Данный модуль добавляет и модифицирует следующие теги: «phonemes», «allophones», «pause».

2.2.4. Акустический процессор

Акустический процессор на вход получает данные в XML-формате, созданные на предыдущих этапах. Выходом акустического процессора служит собственно синтезированный сигнал. Кроме того, этот модуль добавляет и модифицирует теги: «allophone», «word» (добавляется атрибут длительности звучания слова). Ниже в качестве примера приведён фрагмент XML-предложения «Привет мир!» для слова «мир», получаемый на выходе системы, после синтеза звукового сигнала.

```
<text>
  <sentence>
…
    <word dur="14866">
      м<letter char="м" reduct="-1"/><phoneme ph="m&apos;"/>
      <allophone ph="m&apos;" spread="normal" code="m&apos;a"
FO_INIT="112" OtN="1" FO1="112" En="0" Rm="71"/>
      <stress type="1" power="2"/>и<letter char="и"
reduct="0"/><phoneme ph="и"/>
      <allophone ph="i0" spread="normal" code="m&apos;i0r"
FO_INIT="112" OtN="3" FO1="80" FO2="80" FO3="80" En="0" Rm="59"/>
      p<letter char="p" reduct="-1"/><phoneme ph="p"/>
      <allophone ph="r" spread="normal" code="_r_k" FO_INIT="80"
OtN="1" FO1="80" En="0" Rm="130"/>
      <dictitem weight="10" form="10" genesys="6" yo_place1="0"
yo_place2="0" stress_dict="1" stress_additional_dict="0" seman-
tics1="0" semantics2="0" subpart_of_speech="1"/>
      </word>
      <intonation type="1"/>
      <pause type="long" time="600"/>
      </sentence>
</text>
```

Заключение

Рассмотренная модульная архитектура, позволяющая параллельно обрабатывать данные разных этапов синтеза, даёт возможность настраивать систему на работу как в многопроцессорной среде, так и в сети. Параллельная обработка данных позволяет значительно ускорить синтез в реальном масштабе времени и при обработке больших объёмов текста. Выбранный в качестве формата данных для межмодульного взаимодействия XML-формат даёт возможность лёгкой модификации системы синтеза и предоставляет мощное средство контроля результатов обработки входных данных на любом из этапов синтеза.

Фактором, существенно влияющим на скорость обработки данных, является сериализация данных при обмене данными между отдельными модулями, а также дисковые операции. Быстрый SAX-парсер позволяет минимизировать затраты на разбор XML-данных. Дисковые операции следует минимизировать путём сокращения обращений к используемым базам данных и вывода служебной информации.

В целом рассмотренная модель позволяет создать эффективную, легко расширяемую систему синтеза.

Литература

1. Sutter H. The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software. // Dr. Dobb's Journal, 30(3), March 2005. P.202–210.
2. Воеводин В.В., Воеводин Вл.В. Параллельные вычисления. СПб.: БХВ-Петербург, 2004.
3. http://www.w3.org/XML/
4. Schroder M., Breuer S. XML Representation Languages as a Way of Interconnecting TTS Modules. // Proc. ICSLP'04 Jeju, Korea, 2004. P. 1889–1892.

Киселев Виталий Владимирович —

директор ООО «Речевые технологии», г. Минск. С 1999 г. профессионально занимается системами синтеза и распознавания речи, диалоговыми речевыми системами. Автор более 30 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.

Сорока Александр Михайлович —

кандидат технических наук, ведущий инженер ООО «Речевые технологии», г. Минск. В 1993 году окончил Минский радиотехнический институт (Факультет вычислительной техники). Опыт работы в области речевых технологий с 2007 года. Область научных интересов: параллельные вычисления, речевые технологии (синтез речи).

Об автоматическом определении эмоций по речи

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

Возможность автоматически определять эмоции по голосу и речи человека необходима для развития успешных диалоговых систем. Идентификация эмоционального состояния человека востребована в телекоммуникационной сфере, в индустрии развлечений, обучении, медицине и других сферах. В статье представлен краткий обзор подходов к исследованию эмоционального состояния человека по его речи, а также приведены примеры реализованных программ для определения эмоций.

Abstract
Possibility of automatic detection emotions a voice and speech of speakers is necessary for the development of dialog systems. Identification of emotion of person is claimed in telecommunications, business solutions, training, medicine and other spheres. In this paper the review of some approaches to research of emotional condition of speaker speech and as examples of realized programs for detection emotions are resulted is presented.

Эмоции и речь тесно взаимосвязаны и играют огромную роль в общении, поэтому автоматическая и объективная диагностика эмоционального состояния человека по его речи представляет большой практический интерес. Возможность распознавать эмоции в речи важна как для исследования самой речи и эмоций, так и для улучшения качества обслуживания клиентов, например в контакт и call-центрах. Различные научные и коммерческие организации занимаются исследованием этого феномена.

Что же такое эмоции? Существует множество определений. Вот некоторые из них. Эмоции — сильные психические состояния, связанные обычно с возбуждением или высоким уровнем энергии и дающие начало чувствам и страстям. Также чувствами или эмоциями называют переживание человеком своего отношения к тому, что он познаёт и делает, к другим людям и к самому себе.

Эмоции бывают положительными или отрицательными. Удивление, эйфория, гнев, страх различаются по степени положительности либо отрицательности.

Эмоции дают нам информацию о том, как говорящий оценивает ситуацию и какие ответные реакции можно от него ожидать. Известно, что мысли и эмоции влияют на дыхание, выражение лица, положение тела, тон и темп голоса. Голос выражает любые сильные эмоции, он принимает музыкальный мелодичный характер, меняясь по громкости, тембру и высоте звука. Темп речи и её ритмическое членение с помощью пауз и логического ударения также имеют выразительное значение, помогающие уловить эмоции человека.

Важный канал для опознания эмоционального состояния человека — его речь. Она передаёт самые тонкие, деликатные эмоции. Скорость речи зависит от индивидуальных качеств и намерений говорящего. Тембр зависит от того, что говорит человек, какие чувства испытывает. Так, при раздражении тембр прерывисто-царапающий, при апатии — ленивоглухой, при радости — звонко-здоровый, при недоумении — оловянно-нерешительный, при гневе — прерывисто-разрывающий [1].

Т.В. Корнева и Е.Ф. Бажин ещё в 1977 г. установили, что различия в точности распознавания эмоций по голосу связаны в основном с модальностью эмоций [2]. Наименьшее количество ошибок при такой оценке испытуемые получили при идентификации гнева и ровного настроения. Средний балл их опознания в процентном соотношении составил соответственно 99,3 и 97,0. Другие эмоции оценивались хуже. Так, средний балл опознания сниженного настроения равнялся 75,8; тревоги — 81,4; апатии — 80,7; повышенного настроения — 79,5.

В.П. Морозов в 1991 г. ввёл термин «эмоциональный слух» — способность опознавать эмоции по речи и пению человека [3]. Между эмоциональным слухом и речевым слухом отсутствует корреляция. «Эмоциональная глухота» может встречаться и у людей с хорошо развитым восприятием речи. Любопытные данные были получены в отношении точности распознавания эмоций людьми разного возраста, пола и профессий. Испытуемые показали существенные различия в правильности понимания эмоций — от 10 до 95%. Выявлено, что музыканты и вокалисты обладают более развитым эмоциональным слухом. В связи с этим эмоциональный слух стал рассматриваться как один из критериев художественной одарённости, который стал использоваться на приёмных экзаменах в консерваторию.

Исследование В. Х. Манерова (1993) идентификации эмоций по речи показало, что основным признаком, используемым человеком при слуховом восприятии эмоционально обусловленных изменений речи, является степень речедвигательного возбуждения [4]. Определение вида эмоции, переживаемой говорящим, осуществляется слушающим менее успешно, чем определение степени эмоционального возбуждения. Наиболее точно опознаются базовые эмоции, затем удивление и неуверенность и хуже всего — презрение и отвращение. На точность опознания влияет способность диктора передавать в речи эмоциональные состояния. Существует тенденция лучшего распознавания положительных эмоций по сравнению с индифферентным и отрицательным эмоциональными состояниями.

Информация, используемая человеком при определении эмоций других людей, связана с так называемыми *«когнитивными схемами эмоций»*, т.е. с установлением того набора признаков, с помощью которого можно судить о наличии той или иной эмоции. Сопоставление совокупности наблюдаемых признаков со схемой позволяет идентифицировать эмоцию.

При этом предполагается, что ни один из признаков жёстко не привязан к определённой эмоции, а её идентификация осуществляется на вероятной основе. Эмоции других людей распознаются по внешним проявлениям эмоций: изменению речи и голоса, поведения,

ответные реакции. Учитываются также антецеденты, т.е. то, что предшествует и является причиной эмоций: ситуация в её взаимодействии с имеющейся у человека целью [5].

Учёные университета Эль-Паco (США) выбрали для изучения такое понятие, как **уровень уверенности высказывания**, т.е. насколько уверенно говорящий произносит то или иное высказывание. Идея исследования заключалась в том, чтобы создать модель прогнозирования уровня уверенности. Высказывания, на которых практиковалась модель, — это высказывания различных уровней уверенности, они взяты из речи носителей английского языка [6].

Уровень уверенности говорящего определяется **тоном и высотой голоса:**

- *явно высокий* — энтузиазм, радость, заинтересован и проявляет интерес;
- *высокий*, в широком диапазоне силы, тональности и высоты — гнев и страх, неуверенность;
- *чрезмерно высокий,* пронзительный — беспокойство;
- *мягкий и приглушённый*, с понижением интонации к концу каждой фразы — печаль, усталость;
- *форсирование звука* — напряжение, обман;
- в состоянии эмоционального возбуждения обычно *возрастает сила голоса*, изменяются его высота и тембр, но иногда сильное возбуждение может, наоборот, проявляться в *уменьшении силы голоса* (человек «шипит от ярости»).

Учёные Саутгемптонского университета (Великобритания) разработали компьютерные методы, позволяющие прогнозировать **ответную эмоциональную реакцию** говорящего. В ходе эксперимента было выявлено, что просодическая информация помогает в автоматическом определении степени раздражённости человека. Наиболее полную информацию о внутреннем психоэмоциональном состоянии человека может дать анализ его связной речи: расстановка логических ударений, скорость произнесения слов, конструкция фразы, наличие таких отклонений от нормы, как неуверенный или неверный подбор слов, обрывание фраз на полуслове, изменение слов, появление слов-паразитов, исчезновение пауз и т.д.

В результате было выделено:

- *быстрая речь* — очевидная взволнованность, страстное желание убедить или уговорить кого-то;
- *медленная речь* — высокомерие, усталость, угнетённое состояние;
- *прерывистая речь* — неуверенность;
- *лаконичность и решительность речи* — явная уверенность;
- *заикание* — напряжённость или обман;
- *нерешительность в подборе слов* — неуверенность в себе или намерение внезапно удивить чем-то;
- *появление речевых недостатков* (повторение или искажение слов, обрывание фраз на полуслове) — несомненное волнение, но иной раз и желание обмануть;
- *опускание речевых пауз* — напряжение;
- *слишком удлинённые паузы* — незаинтересованность или несогласие;

- *появление в речи пауз, заполняемых словами-паразитами* — нерешительность и затруднение в выражении мысли, поиск выхода из положения;
- *возрастание числа тривиальных наборов слов*, проговариваемых быстрее, чем обычно, — эмоциональное возбуждение, напряжение;
- *умолкание или скупость в словах* — обида.

Распознавание эмоционального состояния человека представляет огромный интерес. Проблема автоматического распознавания эмоционального состояния говорящего по голосу на данный момент не решена. Существующие системы различаются списками распознаваемых эмоций, типами используемых баз данных, акустическими параметрами и их производными, а также алгоритмами классификаторов; эти различия делают результаты распознавания впрямую несопоставимыми.

Так как эмоции и мысли влияют на дыхание, выражение лица, положение тела, тон и темп голоса, то определять эмоции можно по выражению лица, по речи и голосу. Уже созданы некоторые программы для определения эмоций по выражению лица. Так, например, учёные из Университета короля Хуана Карлоса (Испания) разработали систему, способную различать выражения лиц в режиме реального времени. На скорости 30 кадров в секунду программа анализирует выражение лица человека и классифицирует его в соответствии с шестью заложенными в неё шаблонами: гнев, отвращение, страх, счастье, печаль и удивление. Анализу может подвергаться как лицо целиком, так и его часть. Для идентификации выражения лица система использует базу данных Cohn-Kanade, содержащую 333 варианта выражения лиц различных людей. Вероятность совпадения с базой — 89%. Система может работать и в неблагоприятных условиях, на неё не влияет ни изменение освещённости, ни движение пользователя [7].

Создаются компьютерные программы, позволяющие определять эмоции по речи человека. Проводятся работы по компьютерному детектору эмоций по голосу **(Voice-Stress Analysis)** на основе анализа стресса. Такие современные системы находят применение в США в государственных и правоохранительных органах [8].

Создана ещё одна интересная компьютерная программа, позволяющая выявить и проанализировать в диалоге эмоциональное состояние собеседника по его речи — ***детектор любви***. Научно доказано наличие глубинной связи между чувствами человека и особенностями его речи. Богатая палитра эмоций и оттенков настроения выражается в тончайших модуляциях нашего голоса. А эта компьютерная программа анализирует особенности голоса, исследует диапазон эмоций говорящего, определяет степень концентрации внимания, уровень смущения и волнения [9].

В 2006 году один из южнокорейских операторов запустил мобильный сервис анализа голоса, который основан на системе голосового анализа и действует как детектор эмоций, делая заключения **об уровне честности** участников разговора. В течение разговора анализируются различные звуки, которые попадают в микрофон абонента, и делается заключение об их эмоциональном статусе. В конце разговора абонент получает сообщение с графиком правдивости, где показан уровень стресса и число неточных ответов и попыток сменить тему. Происходит анализ, который учитывает, как определённая мозговая активность влияет на специфические особенности голоса. Это позволяет определить и измерить широкий спектр эмоций, используя различные оценки составляющих эмоций, строить оценку правдивости любого утверждения, сделанного участниками разговора [10].

Среди коммерческих организаций, активно использующих и разрабатывающих автоматические модули оценки эмоционального состояния, можно выделить такие компании, как Nemesysco Ltd. [11], Nice Systems Ltd [12], Центр Речевых технологий [13]. Модули нашли своё практическое применение в call-центрах при анализе разговоров как оператора, так и клиента. Как правило, компании не ограничиваются двумя эмоциональными состояниями. Например, компания Nemesysco Ltd. может опознавать до 16-ти эмоциональных состояний с различными числовыми значениями каждого состояния — от удовлетворённости, расстройства или злости до сомнения или неуверенности. Другие компании используют комплексный анализ по голосу и речи, применяя языко-зависимые технологии поиска ключевых слов.

Системы, распознающие эмоциональное состояние человека, могут быть применены в интерактивном телевидении, виртуальном обучении, при исследовании нарушений функций мозга, а также будут полезны людям, имеющим какие-либо речевые отклонения. Для развития успешных диалоговых систем необходимы исследования по выявлению эмоций человека по его речи. Понимание эмоций другого человека важно как для общения между людьми, так и при взаимодействии человека с системами искусственного интеллекта. Автоматическое распознавание речи и прогнозирование эмоций говорящего нашли бы активное применение, например, в телекоммуникационной сфере и индустрии развлечений, что помогло бы избегать конфликтных ситуаций и улучшить качество обслуживания клиентов.

Литература

1. Хаббард Л. Рон. Свободный человек // Способность. № 232.
2. http://www.emotionlabs.ru/content/66/.
3. http://cons-help.com/63/.
4. Джемс В. Психология. Часть II СПб: Изд-во К.Л. Риккера, 1911. С. 323–340.
5. Манёров В.Х., Шнейдер Е.М. Автоматическое распознавание эмоций по спектральным и интонационным признакам // Материалы доклада и сообщения 5-го Всесоюзного совещания-симпозиума цикла «Акустика речи и слуха». Одесса, 1989.
6. Frijda N.H. (1986). The emotions. Cambridge: Cambridge University Press.
7. Jaime C. Acosta and Nigel G. Ward. Responding to User Emotional State by Adding Emotional Coloring to Utterances. In Twelfth International Conference on Spoken Language Processing. ISGA, 2009.
8. http://www.voicestressanalysis.net/
9. http://www.membrana.ru/lenta/?26699
10. http://www.ukrpolygraph.org/2006/09/28/90
11. http://www.nemesysco.com
12. www.nice.com
13. http://www.speechpro.ru

Логические функции определения границ и интонационного типа пунктуационных синтагм

О.Г. Сизонов,
соискатель

Описываются правила пунктуационного синтагматического членения и интонационной разметки текста для синтеза русской речи по тексту. Правила учитывают не только типы знаков пунктуации в предложении, но и их ближайшее окружение в тексте. Это позволяет реализовать алгоритм синтеза интонационно-окрашенной речи, избегая монотонности «второго рода», т.е. частой повторяемости одинаковых интонационных конструкций.

Abstract

The rules of punctuation syntagmatic segmentation and intonation marking of a text for Russian text-to-speech synthesis are described. Not only the types of the signs of punctuation in offer, but their nearest surroundings in a text take the rules into account. It is allowed to realize an algorithm of a synthesis intonationally-avoiding monotony «of the second order» i.e. the fre-quent repeatability of the similar intonation constructions.

Введение

Среди объектов исследования лаборатории синтеза и распознавания речи Объединённого института проблем информатики НАН Беларуси метод мультиволнового синтеза речи по тексту занимает одно из приоритетных направлений как перспективный в плане достижения высокого качества синтезируемого сигнала, в том числе разборчивости и интонационной выразительности. Разработанная в лаборатории компьютерная модель мультиволнового синтеза речи по тексту учитывает специфику лингвистической обработки текстов, фонетическую и просодическую структуру русской речи, особенности артикуляторно-акустических явлений процесса речеобразования [1].

Интонационные характеристики синтезируемой речи, её разборчивость, выразительность и естественность обеспечиваются вычислением необходимых значений

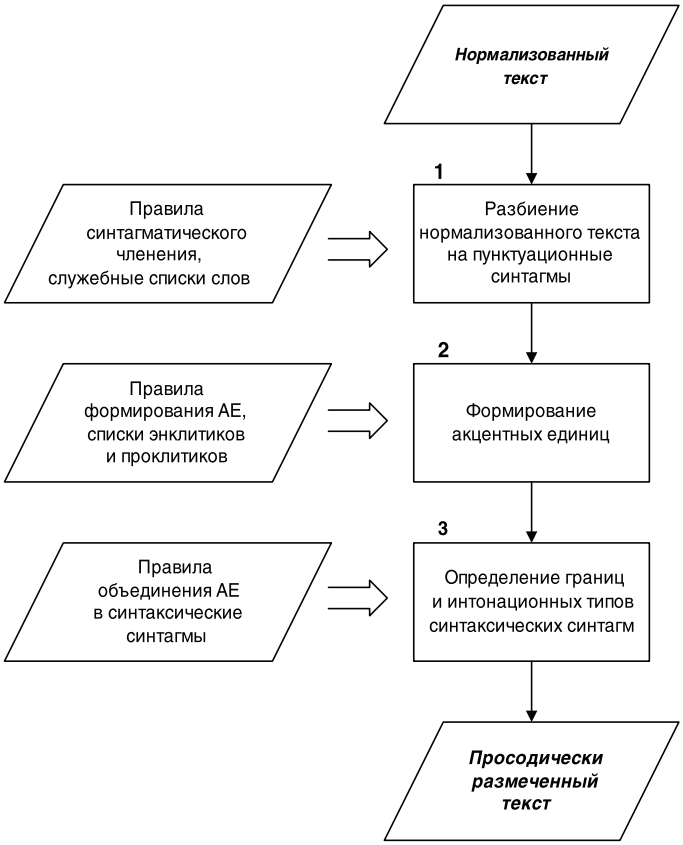


Рис. 1. Структурная схема просодического процессора

частоты основного тона (F_0), длительности звуков и пауз (T), амплитуды звука (A). Эта просодическая информация определяется на основе анализа определённых свойств входного текста, осуществляемого просодическим блоком, который на основе морфологической, синтаксической, пунктуационной информации подаваемого текста обнаруживает в нём минимальные интонационные единицы — синтагмы, определяет их интонационный тип. Маркированный текст является результатом работы, выходными данными просодического процессора.

Последовательность этапов, выполняемых процессором, представлена на *рис. 1*.

Предметом данной работы является формализация и описание правил просодической разметки, производимой просодическим процессором на первом этапе обработки текста — разбиении поступившего нормализованного текста на пунктуационные синтагмы, включая определение их интонационных типов. Правила описаны в виде логических функций, аргументами которых являются последовательность слов и пунктуационных символов нормализованного текста, служебные списки слов и морфологических признаков.

1. Перечень интонационных типов пунктуационных синтагм

Интонационные типы можно разделить на следующие категории:

— завершённости ***P***,
— незавершённости ***C***,
— восклицания ***E***,
— вопроса ***Q***.

Входящие в эти категории типы обозначаются символом категории и соответствующим индексом.

Все интонационные типы и определяющие их признаки приведены в таблицах 1–4.

Таблица 1

Интонационные типы завершённости

Обозначение	*Признак конца синтагмы*	*Комментарий*
P1	[:]	Интонация «двоеточия»
P2	[)]	Интонация «закрывающей скобки»
P3	[;]	Интонация «точки с запятой»
P4	[.] 1-й раз	Интонация «точки» каждой 1-й, 4-й, 7-й и т.д. идущей подряд синтагмы данного типа
P4_1	[.] 2-й раз	Интонация «точки» каждой 2-й, 5-й, 8-й и т.д. идущей подряд синтагмы данного типа
P4_2	[.] 3-й раз	Интонация «точки» каждой 3-й, 6-й, 9-й и т.д. идущей подряд синтагмы данного типа
P5	[…]	Интонация «многоточия»
P6	[.] + абзац	Интонация «абзаца»
P7	Сочинительный союз + [.]	Интонация «точки» при условии, что синтагма начинается с сочинительного союза и ей предшествует синтагма с интонацией незавершённости
P8	Вопросительный союз + [.]	Интонация «точки» при условии, что синтагма начинается с вопросительного союза и ей предшествует синтагма с интонацией незавершённости
P9	Подчинительный союз + [.]	Интонация «точки» при условии, что синтагма начинается с подчинительного союза и ей предшествует синтагма с интонацией незавершённости
P10	Причастие + [.]	Интонация «точки» при условии, что синтагма начинается с причастия и ей предшествует синтагма с интонацией незавершённости
P11	Деепричастие + [.]	Интонация «точки» при условии, что синтагма начинается с деепричастия и ей предшествует синтагма с интонацией незавершённости

Таблица 2

Интонационные типы незавершённости

Обозначение	*Признак конца синтагмы*	*Комментарий*
C1	Союз «И»	Интонация «И»
C2	«ИЛИ»	Интонация «ИЛИ»
C3	[,] 1-й раз	Интонация «запятой» для каждой 1-й, 4-й, 7-й и т.д. идущей подряд синтагмы данного типа
C3_1	[,] 2-й раз	Интонация «запятой» каждой 2-й, 5-й, 8-й и т.д. идущей подряд синтагмы данного типа
C3_2	[,] 3-й раз	Интонация «запятой» каждой 3-й, 6-й, 9-й и т.д. идущей подряд синтагмы данного типа
C4	[–]	Интонация «тире»
C5	[( ]	Интонация «открывающей скобки»
C6	[, –]	Интонация «запятая — тире»
C7	[,] + сочинительный союз	Интонация «запятой перед сочинительным союзом»
C8	[,] + вопросительный союз	Интонация «запятой перед вопросительным союзом»

C9	[,] + подчинительный союз	Интонация «запятой перед подчинительным союзом»
C10	[,] + причастие	Интонация «запятой перед причастием»
C11	[,] + деепричастие	Интонация «запятой перед деепричастием»

Таблица 3

Интонационные типы восклицания

Обозначение	*Признак конца синтагмы*	*Комментарий*
E1	[!] + восклицательное словоодна синтагма	Интонация «восклицания» последней или единственной синтагмы восклицательного предложения, содержащей восклицательное слово
E1_1	[!] + восклицательное слово 1-я синтагма	Интонация «восклицания» каждой нечётной (кроме последней) идущей подряд, содержащей восклицательное слово, синтагмы восклицательного предложения
E1_2	[!] + восклицательное слово 2-я синтагма	Интонация «восклицания» каждой чётной (кроме последней) идущей подряд, содержащей восклицательное слово, синтагмы восклицательного предложения
E2	[!] одна синтагма	Интонация «восклицания»последней или единственной синтагмы восклицательного предложения, не содержащей восклицательного слова
E2_1	[!] 1-я синтагма	Интонация «восклицания» каждой нечётной (кроме последней) идущей подряд и не содержащей восклицательного слова синтагмы восклицательного предложения
E2_2	[!] 2-я синтагма	Интонация «восклицания» каждой чётной (кроме последней) идущей подряд и не содержащей восклицательного слова синтагмы восклицательного предложения

Таблица 4

Интонационные типы вопроса

Обозначение	*Признак конца синтагмы*	*Комментарий*
Q1	[?] + вопросительное слово одна синтагма	Интонация «вопроса» последней или единственной синтагмы вопросительного предложения, содержащей вопросительное слово
Q1_1	[?] + вопросительное слово 1-я синтагма	Интонация «вопроса» каждой нечётной (кроме последней) идущей подряд, содержащей вопросительное слово, синтагмы вопросительного предложения
Q1_2	[?] + вопросительное слово 2-я синтагма	Интонация «вопроса» каждой чётной (кроме последней) идущей подряд, содержащей вопросительное слово, синтагмы вопросительного предложения

Q2	[?] Одна синтагма	Интонация «вопроса» последней или единственной синтагмы вопросительного предложения, не содержащей вопросительного слова
Q2_1	[?] 1-я синтагма	Интонация «вопроса» каждой нечётной (кроме последней) идущей подряд и не содержащей вопросительного слова синтагмы вопросительного предложения
Q2_2	[?] 2-я синтагма	Интонация «вопроса» каждой чётной (кроме последней) идущей подряд и не содержащей вопросительного слова синтагмы вопросительного предложения

2. Логические функции блока интонационной разметки пунктуационных синтагм

Основа синтагматического членения текста — выделение по его пунктуационным признакам предложений или их частей, называемых пунктуационными синтагмами [2].

Входные данные блока интонационной разметки пунктуационных синтагм:

$W=\{w_1, w_2, \ldots w_n\}$ — «последовательность слов и пунктуационных символов нормализованного текста, где n — число слов и знаков пунктуации в тексте.

Выходные данные блока интонационной разметки:

S — «последовательность синтагм, формируемых в процессе просодической обработки входного текста;

T — «последовательность интонационных типов синтагм, где T_j — интонационный тип синтагмы S_j.

Каждая функция F определяет, является ли i-й элемент множества W окончанием j-й синтагмы S соответствующего интонационного типа T_j.

2.1. Логические функции определения пунктуационных синтагм с интонацией «незавершённости»

2.1.1. [C1] «устанавливается после слова, за которым идёт союз «И» или «ДА».
[1]Пример: *Пьер уже три месяца выбирал карьеру[C1] и ничего не делал[P4].*

$$F_{/C1/} = (W_{i+1} \equiv «И») \vee (Wi+1 \equiv «ДА»)$$

2.1.2. [C2] «устанавливается после слова, за которым идёт союз «ИЛИ».
[2]Пример: *Маленькая княгиня не слыхала[C2] или не хотела слышать его слов[P4].*

$$F_{/C2/} = (W_{i+1} \equiv «ИЛИ»)$$

2.1.3. [C4] «устанавливается после слова, после которого встретился символ «-», отделённый с обеих сторон пробелами.
[3]Пример: *Этот пресловутый нейтралитет Пруссии[C4] «только западня[P4].*

$$F_{/C4/} = (W_{i+1} \equiv «-»)$$

2.1.4. [C5] «устанавливается после слова, после которого идёт символ «(«, перед которым может быть пробел.

[4]Пример: Богданыч[C5] (Богданычем называли полкового командира[P2]) вас оса*дил [P4].*

$$F_{/C5/} = (W_{i+1} \equiv «(»)$$

2.1.5. [C6] «устанавливается после слова, непосредственно после которого встречена последовательность символов «,-», разделённая или нет пробелов, если интонационный тип предыдущей синтагмы отличен от [C6].

[5]Пример: *Такая странная антипатия[C6], «думал Пьер[P2], «а прежде он мне даже очень нравился[P4].*

$$F_{/C6/} = (W_{i+1} \equiv «,») \wedge (W_{i+2} \equiv «-») \wedge (T_{j-1} \not\equiv C_6)$$

2.1.6. [C7] «устанавливается после слова, после которого идут символ «,» и сочинительный союз <M1>, разделённые пробелом.

[6]Пример: *Графиня хотела хмуриться[C7], но не могла[P7].*

$$F_{/C7/} = (W_{i+1} \equiv «,») \wedge (W_{i+2} \in M1)$$

2.1.7. [C8] «устанавливается после слова, после которого идут символ «,» и вопросительно-подчинительный союз <M2>, разделённые пробелом.

[7]Пример: *Генерал садился на лошадь[C8], которую подал ему казак[P8].*

$$F_{/C8/} = (W_{i+1} \equiv «,») \wedge (W_{i+2} \in M2)$$

2.1.8. [C9] «устанавливается после слова, после которого идут символ «,» и подчинительный союз <M3>, разделённые пробелом.

[8]Пример: *Предложение было слишком лестно[C9], чтобы отказаться[P9].*

$$F_{/C9/} = (W_{i+1} \equiv «,») \wedge (W_{i+2} \in M3)$$

2.1.9. [C10] «устанавливается после слова, после которого идут символ «,» и причастие, определяемое по списку суффиксов <M4>, разделённые пробелом.

[9]Пример: *Вейротер был австрийский генерал[C10], заменивший убитого Шмита[P10].*

$$F_{/C10/} = (W_{i+1} \equiv «,») \wedge (W_{i+2} \in M4)$$

2.1.10. [C11] «устанавливается после слова, после которого идут символ «,» и деепричастие, определяемое по списку окончаний <M5>, разделённые пробелом.

[10]Пример: *Остальная пехота поспешно проходила по мосту[C11], спираясь воронкой у входа[P11].*

$$F_{/C11/} = (W_{i+1} \equiv «,») \wedge (W_{i+2} \in M5)$$

2.2. Логические функции определения пунктуационных синтагм с интонацией «завершённость»

2.2.1 [P1] «устанавливается после слова, после которого идёт символ «:», перед которым может быть пробел.

[11]Пример: *Всё только одного желали[P1]: под предводительством государя скорее итти против неприятеля[P4].*

$$F_{/P1/} = (W_{i+1} \equiv \text{«:»})$$

2.2.2. [P2] «устанавливается после слова, если после него идёт символ «)», перед которым может быть пробел, или если непосредственно после слова встречена последовательность символов «,-», разделённая или нет пробелом, и интонационный тип предыдущей синтагмы [C6].

[12]Пример: *Богданыч[C5] (Богданычем называли полкового командира[P2]) вас осадил [P4].*

$$F_{/P2/} = [(W_{i+1} \equiv \text{«)»})] \vee [(W_{i+1} \equiv \text{«,»}] \wedge [(W_{i+2} \equiv \text{«–»}) \wedge (T_{j-1} \equiv C6)]$$

2.2.3. [P3] «устанавливается после слова, после которого идут символ «;», перед которым может быть пробел.

[13]Пример: *Оттепель и туман продолжались [P3]; за 40 шагов ничего не было видно[P4].*

$$F_{/P3/} = (W_{i+1} \equiv \text{«;»})$$

2.2.4. [P5] «устанавливается после слова, после которого идёт подряд три символа «.» (многоточие), перед которыми может быть пробел.

[14]Пример: *Только в Юхнове с Пелагеюшкой сошлись[P5]…*

$$F_{/P5/} = (W_{i+1} \equiv \text{«.»}) \wedge (W_{i+2} \equiv \text{«.»}) \wedge (W_{i+3} \equiv \text{«.»})$$

2.2.5. [P7] «устанавливается после слова, за которым встречен символ «.», если предыдущая синтагма имела интонационный тип [C7].

[15]Пример: *Графиня хотела хмуриться[C7], но не могла[P7].*

$$F_{/P7/} = (W_{i+1} \equiv \text{«.»}) \wedge (T_{j-1} \equiv C7)$$

2.2.6. [P8] «устанавливается после слова, за которым встречен символ «.», если предыдущая синтагма имела интонационный тип [C8].

[16]Пример: *Генерал садился на лошадь[C8], которую подал ему казак[P8].*

$$F_{/P8/} = (W_{i+1} \equiv \text{«.»}) \wedge (T_{j-1} \equiv C8)$$

2.2.7. [P9] «устанавливается после слова, за которым встречен символ «.», если предыдущая синтагма имела интонационный тип [C9].

[17] Пример: *Предложение было слишком лестно[C9], чтобы отказаться[P9].*

$$F_{/P9/} = (W_{i+1} \equiv \text{«.»}) \wedge (T_{j-1} \equiv C9)$$

2.2.8. [P10] «устанавливается после слова, за которым встречен символ «.», если предыдущая синтагма имела интонационный тип [C10].

[18]Пример: Вейротер был австрийский генерал[C10], заменивший убитого Шмита[P10].

$$F_{/P10/} = (W_{i+1} \equiv \text{«.»}) \wedge (T_{j-1} \equiv C10)$$

2.2.9. [P11] «устанавливается после слова, за которым встречен символ «.», если предыдущая синтагма имела интонационный тип [C11].

[19] Пример: Остальная пехота поспешно проходила по мосту[C11], спираясь воронкой у входа[P11].

$$F_{/P11/} = (W_{i+1} \equiv \text{«.»}) \wedge (T_{j-1} \equiv C11)$$

2.2.10. [P6] «устанавливается после слова, за которым встречен признак абзаца: последовательность из точки, переноса строки и табуляции.

[20] Пример: *Наташа стала надевать платье [P6].*

$$F_{/P6/} = (W_{i+1} \equiv \text{«.»}) \wedge (W_{i+2} \equiv \text{«↵«}) \wedge (W_{i+3} \equiv \text{«→»})$$

2.2.11. [P4] «устанавливается после слова, за которым встречен символ «.», за которым нет признака абзаца, и нет двух идущих подряд символов «.», и интонационный тип предыдущей синтагмы отличен от [C7], [C8], [C9], [C10], [C11], [P4], [P4_1].

[21] Пример: *В четверть одиннадцатого наконец сели в кареты и поехали. Но ещё нужно было заехать к Таврическому саду[P4_1]. Перонская была уже готова[P4_2]. Ростовы похвалили её вкус и туалет[P4].*

$$F_{/P4/} = (W_{i+1} \equiv \text{«.»}) \wedge (W_{i+2} \equiv \text{«.»}) \wedge (W_{i+2} \equiv \text{«↵«}) \wedge (T_{j-1} \not\equiv C7) \wedge$$
$$(T_{j-1} \not\equiv C8) \wedge (T_{j-1} \not\equiv C9) \wedge (T_{j-1} \not\equiv C10) \wedge (T_{j-1} \not\equiv C11) \wedge (T_{j-1} \not\equiv P4) \wedge (T_{j-1} \not\equiv P4\_1)$$

2.2.12. [P4_1] «устанавливается после слова, за которым встречен символ «.», за которым нет признака абзаца, и нет двух идущих подряд символов «.», и интонационный тип предыдущей синтагмы [P4].

[22] Пример: *В четверть одиннадцатого наконец сели в кареты и поехали. Но ещё нужно было заехать к Таврическому саду[P4_1]. Перонская была уже готова[P4_2]. Ростовы похвалили её вкус и туалет[P4].*

$$F/P4\_1/ = (W_{i+1} \equiv \text{«.»}) \wedge (W_{i+2} \not\equiv \text{«.»}) \wedge (W_{i+2} \not\equiv \text{«↵»}) \wedge (T_{j-1} \not\equiv P4)$$

2.2.13. [P4_2] «устанавливается после слова, за которым встречен символ «.», за которым нет признака абзаца, и нет двух идущих подряд символов «.», и интонационный тип предыдущей синтагмы [P4_1].

[23] Пример: *В четверть одиннадцатого наконец сели в кареты и поехали. Но ещё нужно было заехать к Таврическому саду[P4_1]. Перонская была уже готова[P4_2]. Ростовы похвалили её вкус и туалет[P4].*

$$F/P4\_2/ = (W_{i+1} \equiv \text{«.»}) \wedge (W_{i+2} \not\equiv \text{«.»}) \wedge (W_{i+2} \not\equiv \text{«↵»}) \wedge (T_{j-1} \not\equiv P4\_1)$$

2.2.14. [C3] «устанавливается после слова, за которым встречен символ «,», за которым нет символа «–», и нет слова, определяемого по спискам <M1>,<M2>, <M3>,<M4>,<M5>, и интонационный тип предыдущей синтагмы не [C3] и не [C3_1].

[24] Пример: *Видите[C3], погода мокрая[C3_1], говорил дядюшка[C3_2], отдохнули бы[C3], графинечку бы отвезли в дрожках[P4].*

$$F/C3/ = (W_{i+1} \equiv \text{«,»}) \wedge (W_{i+2} \not\equiv \text{«– »}) \wedge (W_{i+2} \not\in M1) \wedge (W_{i+2} \not\in M2) \wedge$$
$$(W_{i+2} \not\in M3) \wedge (W_{i+2} \not\in M4) \wedge (W_{i+2} \not\in M5) \wedge (T_{j-1} \not\equiv C3) \wedge (T_{j-1} \not\equiv C\_3)$$

2.2.15. [C3_1] «устанавливается после слова, за которым встречен символ «,», за которым нет символа «–», и нет слова, определяемого по спискам <M1>,<M2>,<M3>,<M4>,<M5>, и интонационный тип предыдущей синтагмы [C3].

[25] Пример: *Видите[C3], погода мокрая[C3_1], говорил дядюшка[C3_2], отдохнули бы[C3], графинечку бы отвезли в дрожках[P4].*

$$F/C3\_1/ = (W_{i+1} \equiv «,») \wedge (W_{i+2} \not\equiv «-») \wedge (W_{i+2} \notin M1) \wedge (W_{i+2} \notin M2) \wedge (W_{i+2} \notin M3) \wedge (W_{i+2} \notin M4) \wedge (W_{i+2} \notin M5) \wedge (T_{j-1} \equiv C3)$$

2.2.16. [C3_2] «устанавливается после слова, за которым встречен символ «,», за которым нет символа «–», и нет слова, определяемого по спискам <M1>,<M2>,<M3>,<M4>,<M5>, и интонационный тип предыдущей синтагмы [C3_1].

[26] Пример: *Видите[C3], погода мокрая[C3_1], говорил дядюшка[C3_2], отдохнули бы[C3], графинечку бы отвезли в дрожках[P4].*

$$F/C3\_1/ = (W_{i+1} \equiv «,») \wedge (W_{i+2} \not\equiv «-») \wedge (W_{i+2} \notin M1) \wedge (W_{i+2} \notin M2) \wedge (W_{i+2} \notin M3) \wedge (W_{i+2} \notin M4) \wedge (W_{i+2} \notin M5) \wedge (T_{j-1} \equiv C3\_1)$$

2.3. Логические функции определения пунктуационных синтагм с интонацией «вопрос и восклицание»

2.3.1. [Q1] «устанавливается после слова, если после него следует символ «?» и в текущей формируемой синтагме есть вопросительное слово, определяемое по списку вопросительных слов <M6>.

[27] Пример: *И как могла она допустить до этого Курагина[Q1]?*

$$F_{/Q1/} = (W_{i+1} \equiv «?») \wedge (\exists m, m \in M6 \vee m \in Sj)$$

2.3.2. [Q2] «устанавливается после слова, если после него следует символ «?» и в текущей формируемой синтагме отсутствуют вопросительные слова из списка вопросительных слов <M6>.

[28] Пример: *Прикажете наших из-под горы кликнуть[Q2]?*

$$F_{/Q2/} = (W_{i+1} \equiv «?») \wedge (\nexists m, m \in M6 \vee m \in Sj)$$

2.3.3. [Q1_1] «устанавливается в текущей формируемой синтагме, содержащей вопросительное слово из списка <M6> после слова, за которым следует один из символов «–», «(«, «)» или «,» за которым может быть определяемое по множествам <M1>, <M2>, <M3>, <M4>, <M5> слово и при этом интонационный тип следующей синтагмы один из следующих: [Q1], [Q2], [Q1_2], [Q2_2], а интонационный тип предыдущей синтагмы не [Q1_1] и не [Q2_1].

[29] Пример: *А что такое война[Q1_1], что нужно для успеха в военном деле[Q1_2], какие нравы военного общества[Q1]?*

$$F_{/Q1\_1/} = [[(W_{i+1} \equiv «,») \vee W_{i+1} \equiv «-» \; W_{i+1} \equiv «)«) \wedge (W_{i+2} \in M1) \wedge (W_{i+2} \in M2) \wedge (W_{i+2} \in M3) \wedge (W_{i+2} \in M4) \wedge (W_{i+2} \in M5))] \wedge (\exists m, m \in M6 \vee m \in Sj)] \wedge [(T_{j+1} \equiv Q1) \vee (T_{j+1} \equiv Q2) \vee (T_{j+1} \equiv Q1\_2) \vee (T_{j+1} \equiv Q2\_2) \wedge (T_{j+1} \not\equiv Q1\_1) \vee (T_{j+1} \not\equiv Q2\_1)]$$

2.3.4. [Q1_2] «устанавливается в текущей формируемой синтагме, содержащей вопросительное слово из списка <M6> после слова, за которым следует один из символов «–», «(«, «)» или «,» за которым может быть определяемое по множествам <M1>, <M2>, <M3>, <M4>, <M5> слово и при этом интонационный тип следующей синтагмы один из следующих: [Q1], [Q2], [Q1_1], [Q2_1], а интонационный тип предыдущей синтагмы «[Q1_2] или [Q2_2].

[30] Пример: *А что такое война[Q1_1], что нужно для успеха в военном деле[Q1_2], какие нравы военного общества[Q1]?*

$$F_{/Q1\_2/} = [[(W_{i+1} \equiv «,» \vee W_{i+1} \equiv «-» \; W_{i+1} \equiv «(« W_{i+1} \equiv «)«) \wedge ((W_{i+2} \in M1) \wedge (W_{i+2} \in M2) \wedge (W_{i+2} \in M3) \wedge (W_{i+2} \in M4) \wedge (W_{i+2} \in M5))] \wedge (\exists m, m \in M6 \vee m \in Sj)] \wedge [(T_{j+1} \equiv Q1) \vee (T_{j+1} \equiv Q2) \vee (T_{j+1} \equiv Q1\_1) \vee (T_{j+1} \equiv Q2\_1) \wedge (T_{j-1} \not\equiv Q1\_2) \vee (T_{j+1} \not\equiv Q2\_2)]$$

2.3.5. [Q2_1] «устанавливается в текущей формируемой синтагме, не содержащей вопросительного слова из списка <M6> после слова, за которым следует одно из сочетаний символов «и», «или», «да», «-», «(«, «)» или «,» за которым может быть определяемое по множествам <M1>, <M2>, <M3>, <M4>, <M5> слово и при этом интонационный тип следующей синтагмы один из следующих: [Q1], [Q2], [Q1_2], [Q2_2], а интонационный тип предыдущей синтагмы не [Q1_1] и не [Q2_1].

[31] Пример: *Любезности это[Q2_1], бабьи сказки[Q2_2], или она права[Q2]?*

$$F_{/Q2\_1/} = [[(W_{i+1} \equiv «,» \vee W_{i+1} \equiv «-» \vee W_{i+1} \equiv «(« \vee W_{i+1} \equiv «)«) \wedge ((W_{i+2} \in M1) \wedge (W_{i+2} \in M2) \wedge (W_{i+2} \in M3) \wedge (W_{i+2} \in M4) \wedge (W_{i+2} \in M5))] \wedge (\exists m, m \in M6 \vee m \in Sj)] \wedge [(T_{j+1} \equiv Q1) \vee (T_{j+1} \equiv Q2) \vee (T_{j+1} \equiv Q1\_2) \vee (T_{j+1} \equiv Q2\_2) \wedge (T_{j-1} \not\equiv Q1\_1) \vee (T_{j+1} \not\equiv Q2\_1)]$$

2.3.6. [Q2_2] «устанавливается в текущей формируемой синтагме, не содержащей вопросительного слова из списка <M6> после слова, за которым следует одно из сочетаний символов «и», «или», «да», «-», «(«, «)» или «,» за которым может быть определяемое по множествам <M1>, <M2>, <M3>, <M4>, <M5> слово и при этом интонационный тип следующей синтагмы один из следующих: [Q1], [Q2], [Q1_2], [Q2_2], а интонационный тип предыдущей синтагмы «[Q1_1] или [Q2_1].

[32] Пример: *Любезности это[Q2_1], бабьи сказки[Q2_2], или она права[Q2]?*

$$F/Q2\_2/ = [[(W_{i+1} \equiv «,» \vee W_{i+1} \equiv «-» \vee W_{i+1} \equiv «(« \vee W_{i+1} \equiv «)«) \wedge ((W_{i+2} \in M1) \wedge (W_{i+2} \in M2) \wedge (W_{i+2} \in M3) \wedge (W_{i+2} \in M4) \wedge (W_{i+2} \in M5))] \wedge (\exists m, m \in M6 \vee m \in Sj)] \wedge [(T_{j+1} \equiv Q1) \vee (T_{j+1} \equiv Q2) \vee (T_{j+1} \equiv Q1\_1) \vee (T_{j+1} \equiv Q2\_1) \wedge (T_{j-1} \not\equiv Q1\_2) \vee (T_{j+1} \not\equiv Q2\_2)]$$

2.3.7. [E1] «устанавливается после слова, если после него следует символ «!» и в текущей формируемой синтагме есть восклицательное слово, определяемое по списку восклицательных слов <M7>.

[33] Пример: *Ай да Данила Купор[E1]!*

$$F_{/E1/} = (W_{i+1} \equiv «!») \wedge (\exists m, m \in M7 \vee m \in Sj)$$

2.3.8. [E2] «устанавливается после слова, если после него следует символ «!» и в текущей формируемой синтагме отсутствуют восклицательные слова из списка восклицательных слов <M7>.

[34] Пример: *Хорошо он себя зарекомендовал в Букарещте[E2]!*

$$F_{/E2/} = (W_{i+1} \equiv «!») \wedge (\exists m, m \in M7 \vee m \in Sj)$$

2.3.9. [E1_1] «устанавливается в текущей формируемой синтагме, содержащей восклицательное слово из списка <M7> после слова, за которым следует один из символов «(«, «)» или «,» за которым может быть определяемое по множествам <M1>, <M2>, <M3>, <M4>, <M5> слово и при этом интонационный тип следующей синтагмы один из следующих: [E1], [E2], [E1_2], [E2_2], а интонационный тип предыдущей синтагмы не [E1_1] и не [E2_1].

[35] Пример: *Ах[E1_1], я право не думал оскорбить её[E1_2], я так понимаю[E2_1] и высоко ценю эти чувства[E2]!*

$$F_{/E1\_1/} = [[(W_{i+1} \equiv «,» \vee W_{i+1} \equiv «-» W_{i+1} \equiv «(«W_{i+1} \equiv «)«) \wedge ((W_{i+2} \in M1) \wedge (W_{i+2} \in M2) \wedge (W_{i+2} \in M3) \wedge (W_{i+2} \in M4) \wedge (W_{i+}2 \in M5))] \wedge (\exists m, m \in M7 \vee m \in Sj)] \wedge [(T_{j+1} \equiv E1) \vee (T_{j+1} \equiv E2) \vee (T_{j+1} \equiv E1\_2) \vee (T_{j+1} \equiv E2\_2) \wedge (T_{j-1} \not\equiv E1\_1) \vee (T_{j+1} \not\equiv E2\_1)]$$

2.3.10 [E1_2] «устанавливается в текущей формируемой синтагме, содержащей восклицательное слово из списка <M7> после слова, за которым следует одно из сочетаний символов «и», «или», «да», «–», «(«, «)» или «,» за которым может быть определяемое по множествам <M1>, <M2>, <M3>, <M4>, <M5> слово и при этом интонационный тип следующей синтагмы один из следующих: [E1], [E2], [E1_2], [E2_2], а интонационный тип предыдущей синтагмы «[E1_1] или [E2_1].

[36] Пример: *Ах[E1_1], я право не думал оскорбить её[E1_2], я так понимаю[E2_1] и высоко ценю эти чувства[E2]!*

$$F_{/E1\_2/} = [[(W_{i+1} \equiv «,» \vee W_{i+1} \equiv «-» W_{i+1} \equiv «(«W_{i+1} \equiv «)«) \wedge ((W_{i+2} \in M1) \wedge (W_{i+2} \in M2) \wedge (W_{i+2} \in M3) \wedge (W_{i+2} \in M4) \wedge (W_{i+2} \in M5))] \wedge (\exists m, m \in M7 \vee m \in Sj)] \wedge [(T_{j+1} \equiv E1) \vee (T_{j+1} \equiv E2) \vee (T_{j+1} \equiv E1\_1) \vee (T_{j+1} \equiv E2\_1) \wedge (T_{j-1} \not\equiv E1\_2) \vee (T_{j+1} \not\equiv E2\_2)]$$

2.3.11. [E2_1] «устанавливается в текущей формируемой синтагме, не содержащей восклицательного слова из списка <M7> после слова, за которым следует одно из сочетаний символов «и», «или», «да», «–», «(«, «)» или «,» за которым может быть определяемое по множествам <M1>, <M2>, <M3>, <M4>, <M5> слово и при этом интонационный тип следующей синтагмы один из следующих: [E1], [E2], [E1_2], [E2_2], а интонационный тип предыдущей синтагмы не [E1_1] и не [E2_1].

[37] Пример: *Голубушка[E2_1], мамаша[E2_2], как я вас люблю[E2_1], как мне хорошо[E2]!*

$$F_{/E2\_1/} = [[(W_{i+1} \equiv «,» \vee W_{i+1} \equiv «-» W_{i+1} \equiv «(«W_{i+1} \equiv «)«) \wedge ((W_{i+2} \in M1) \wedge (W_{i+2} \in M2) \wedge (W_{i+2} \in M3) \wedge (W_{i+2} \in M4) \wedge (W_{i+2} \in M5))] \wedge (\exists m, m \in M7 \vee m \in Sj)] \wedge [(T_{j+1} \equiv E1) \vee (T_{j+1} \equiv E2) \vee (T_{j+1} \equiv E1\_2) \vee (T_{j+1} \equiv E2\_2) \wedge (T_{j-1} \not\equiv E1\_1) \vee (T_{j+1} \not\equiv E2\_1)]$$

2.3.12. [E2_2] «устанавливается в текущей формируемой синтагме, не содержащей восклицательного слова из списка <M7> после слова, за которым следует одно из сочетаний символов «и», «или», «да», «–», «(«, «)» или «,» за которым может быть определяемое по множествам <M1>, <M2>, <M3>, <M4>, <M5> слово и при этом интонационный тип следующей синтагмы один из следующих: [E1], [E2], [E1_2], [E2_2], а интонационный тип предыдущей синтагмы «[E1_1] или [E2_1].

[38]Пример: *Голубушка[E2_1], мамаша[E2_2], как я вас люблю[E2_1], как мне хорошо[E2]!*

$$F_{/E2\_2/} = [[(W_{i+1} \equiv «,» \vee W_{i+1} \equiv «-» \; W_{i+1} \equiv «(«W_{i+1} \equiv «)«) \wedge ((W_{i+2} \in M1) \wedge (W_{i+2} \in M2) \wedge (W_{i+2} \in M3) \wedge (W_{i+2} \in M4) \wedge (W_{i+2} \in M5))] \wedge (\nexists m, m \in M7 \vee m \in Sj)] \wedge [(T_{j+1} \equiv E1) \vee (T_{j+1} \equiv E2) \vee (T_{j+1} \equiv E1\_1) \vee (T_{j+1} \equiv E2\_1) \wedge (T_{j-1} \not\equiv E2\_1) \vee (T_{j+1} \not\equiv E2\_2)]$$

Заключение

Представленные в данной работе результаты исследований позволяют программно реализовать однопроходный алгоритм блока разбиения на пунктуационные синтагмы, выполняемый просодическим процессором. Приведённые правила интонационного анализа текста позволяют реализовать базовую часть функций просодического процессора. На основании результатов работы этапа разбиения на пунктуационные синтагмы производится членение их на синтаксические синтагмы и формирование акцентных единиц. Алгоритмы и правила такого анализа будут представлены в следующей публикации авторов по теме «Правила просодического анализа текста для синтеза речи».

Литература

1. *Лобанов Б.М., Цирульник Л.И.* Компьютерный синтез и клонирование речи. Мн.: Белорусская наука, 2008.

2. Алгоритм интонационной разметки повествовательных предложений для синтеза речи по тексту / Л.И. Цирульник, Б.М. Лобанов, О.Г. Сизонов // Труды Международной конференции «Компьютерная лингвистика и интеллектуальные технологии» (Диалог'2008). М.: Наука, 2008. С. 563–568.

Сизонов Олег Геннадьевич —

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.

Отображение и оценка формантных свойств артикуляции речи интегральными AFB-параметрами динамических спектров речевых сигналов

Н.П. Дегтярёв

Анализируются причины ненадёжного выделения и оценки формантных параметров речевых сигналов известными методами. Предложено измерять не сами формантные параметры речевых сигналов, а *AFB*-параметры обобщённых формант динамических спектров речи. Рассматривается параметрическая модель описания обобщённых формант спектров речевых сигналов, отличающаяся повышенной надёжностью получаемых оценок *AFB*-параметров относительно вариативности исходных характеристик каналов передачи и голосов дикторов. Исследуется фонетическая метрика оценки артикуляции с помощью полученных *AFB*-параметров обобщённых формант спектров речи.

Abstract

The causes of insecure allocation and estimation of the formant parameters of the speech signals by the known methods are analyzed. It is proposed to measure not oneself formants of the speech signals, but the *AFB*-parameters of generalized formants of dynamic speech spectrum. The parametric model of a description of generalized formants of dynamic speech spectrum and its dignity is considered. Phonetic metrics of estimation of articulation with the help of the obtained *AFB*-parameters of generalized formants is investigated.

Введение

Проблема надёжности автоматического выделения и оценки формантных параметров речевого сигнала, несмотря на все как давние [1-3], так и последние [4] предпринятые усилия, не находит своего удовлетворительного решения. Такое уже давно

сложившееся положение заставляет критически переосмыслить подходы к решению проблемы формантного анализа речевых сигналов применительно к задаче распознавания речи. Предложенные методы формантного анализа опираются, как правило, на некоторую универсальную модель анализа речевого сигнала, призванную выделять амплитуды и частоты первых трёх-четырёх формант речи безотносительно к характеру (голосовые или фрикативные) анализируемых сегментов речевого сигнала. Поэтому наиболее устойчивые результаты анализа наблюдаются только на сегментах, согласующихся с заложенной моделью [3]. Характер и причины проявляющихся ошибок (потеря третьей и четвёртой формант из-за низкого уровня относительно шумов, а также второй форманты из-за её шунтирования при назализации или маскировки первой формантой при их сближении и др. [1]) как раз демонстрируют структурную ограниченность используемых моделей формантного анализа, в результате чего не учитываются существенно различные свойства речевого сигнала для разных по способу образования звуков (сегментов) речи. Всё это заставляет обратиться к проблеме акустического описания артикуляции речи с позиции учёта структурных свойств формантной модели образования речевого сигнала.

1. Структурные свойства артикуляции речи

Теория речеобразования [5,6] для каждого из способов образования звуков речи (*рис. 1*) предлагает присущую ему акустическую или эквивалентную электрическую подмодель. Структура каждой такой подмодели специфична, ибо отражает артикуляцию (конфигурацию речевого тракта, место и тип источника возбуждения), свойственную только данному способу. В силу этого каждая подмодель артикуляции способа образования описывается своим, отличным от других подмоделей, набором значащих формантных параметров. Необходимость учёта в процессе анализа порождаемых таким образом структурных свойств речевого сигнала приводит нас к следующим принципиально важным выводам. Во-первых, полная модель формантного анализа речевого сигнала должна включать хотя бы основные подмодели артикуляции звуков речи (*рис. 1*), различающиеся способом их образования.

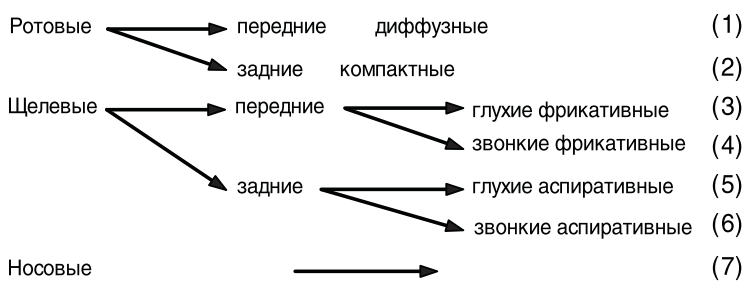


Рис. 1. Структура основных способов образования звуков речи, которые должны учитываться полной моделью формантного анализа речевых сигналов

И главный вывод, к которому мы вынуждены прийти, состоит в том, что решение задачи формантного анализа речевого сигнала возможно только в рамках полной структурной модели формантного анализа с использованием фонетического контекста [3]. Заметим, что попытки построить полную

математическую модель речевого сигнала также приводят к многофакторной структуре такой модели [7].

Следуя далее принятой нами концепции, мы должны признать, что понятие фор-манты является обусловленным определённым контекстом способ образования звуков речи, а формантное описание артикуляции речи по своей природе является контекстно-зависимым. Понятно, что результативность решения задачи формантного анализа в таком случае напрямую связана с корректностью моделирования верхних языковых уровней речевого процесса [3]. Вышесказанное ещё раз указывает на концептуальную и методологическую сложность корректного решения задачи анализа формантных параметров речевых сигналов.

Поэтому разработка методов внеконтекстного анализа и оценивания параметров описания формантных свойств речевых сигналов, на наш взгляд, становится весьма актуальной задачей, поскольку такие параметры описания, с одной стороны, могут составить хорошую основу для первичного (прямого) гипотезирования и распознавания смысловых элементов речи, а с другой могут стать важной составной частью полного структурно-параметрического формантного описания артикуляции речи. Исходя из сказанного, в данной работе рассматривается возможность описания и оценки артикуляции речи с помощью системы *AFB*-параметров обобщённых формант динамических спектров речевых сигналов, вытекающей из предложенной нами ранее двухформантной модели описания артикуляции речи [8–10].

2. Связь артикуляции с формантными свойствами речевого сигнала

Известно [11], что минимальным речеобразующим жестом является слог, минимальным смыслообразующим элементом речи — слово, а минимальным смысловым элементом речевого сообщения — предложение. В общем случае речевое сообщение состоит из последовательности слов и образуется путём артикуляции последовательности составляющих их слогов. В предельном случае речевое сообщение может состоять из одного односложного слова и образовываться минимальным артикуляционным жестом — одним слогом. Известно также, что в процессе речеобразования вследствие инерционности органов артикуляции имеет место явление коартикуляции соседних звуков речи, т.е. взаимовлияние артикуляции соседних звуков, приводящее к взаимозависимости их артикуляторных параметров. Явление коартикуляции звуков речи также имеет закономерную природу, описываемую так называемыми «звуковыми законами» слитной и разговорной речи [12, 13].

Отмеченные свойства речеобразования указывают на то, что основа образования речевых сообщений — закономерные и взаимозависимые движения органов артикуляции, задаваемые программой реализации артикуляторного жеста как минимального смыслообразущего элемента речи. А поскольку процессы артикуляции речи на акустическом уровне отображаются в закономерные изменения во времени структуры и значений формантных параметров речевого сигнала, то именно в закономерностях изменений формантных параметров речевого сигнала и нужно искать акустические инварианты описания минимальных элементов (артикулем) речи. При этом нужно помнить, что искомые инварианты имеют смысл искать в различных реализациях только одного и того же артикуляторного жеста (слова).

Тогда в качестве инвариантной по дикторам функции $\boldsymbol{P^*(t)}$ описания смысловых модуляций параметра $\boldsymbol{P^*(t)}$, адекватной модуляционным свойствам процесса речеобразования, физиологическому закону восприятия раздражений (закон Вебера-Фехнера) и оценке «количества информации» может служить функция,

$$P^*(t) = \ln P(t) - \ln P(t-\tau) = \ln \frac{P(t)}{P(t-\tau)} \qquad (1)$$

где τ — интервал времени, соответствующий разрешающей способности слуха во времени. Нетрудно видеть, что функция (1) не претерпевает существенных изменений при медленных по сравнению с τ изменениях параметра $\boldsymbol{P(t)}$, когда $\boldsymbol{P(t) \cong P(t-\tau)}$.

При условии, что параметр $\boldsymbol{P(t)}$ явным образом отображает движения артикулятора или изменения во времени связанного с ним формантного параметра, функция $\boldsymbol{P^*(t)}$ приобретает смысл фонетической функции, поскольку она инвариантна относительно средних значений $\boldsymbol{\overline{P(t)}}$, которые характеризуют индивидуальные свойства параметра $\boldsymbol{P(t)}$ артикуляционного аппарата каждого конкретного говорящего. Таким образом, суть принципиально важного требования к параметрическому описанию речевого сигнала в задаче дикторонезависимого распознавания речи заключается в том, что каждый из акустических параметров описания речевого сигнала в одном и том же смысловом (фонетическом) контексте должен обладать инвариантностью характера его изменений во времени относительно вариаций (смещений), связанных с индивидуальностью голоса говорящего. Этому требованию отвечают только параметры, отображающие формантные свойства динамических спектров речевых сигналов [7, 8–10].

И, напротив, широко используемые в современных системах распознавания речи спектральные описания речевого сигнала (Фурье, LPC-параметры, кепстральные параметры [14]) не отвечают этому требованию и поэтому не могут служить основой для построения систем дикторонезависимого распознавания речи. Объясняется это тем, что преобразование названных параметров по (1) не устраняет дикторской вариативности спектров, связанной с индивидуальным для каждого диктора диапазоном изменений частот формантных максимумов спектров, в то время как преобразование формантных параметров по (1) отфильтровывает их смещения, связанные с дикторскими вариативностями спектров речи. В этом состоит принципиальное отличие изложенного здесь определения понятия фонетической функции по (1) от предложенного ранее в работе [15]. Модуляционная природа артикуляции речи порождает не только инварианты (в заданном контексте) описания смысловых элементов, но и вариативности, связанные с явлениями коартикуляции звуков речи, закономерно проявляющиеся для различных норм произношения (полный стиль, разговорная речь), а также акцента и темпа произношения.

Таким образом, модель дикторонезависимого распознавания речи должна учитывать не только вариативности, связанные с источниками и переносчиками речевого сигнала (индивидуальные параметры речевого тракта, индивидуальность голоса, параметры среды и канала передачи), но и вариативности, проявляющиеся как в пределах действия звуковых законов слитной речи (стиль, темп), так и в форме индивидуальных особенностей (акцента) произношения. Вариативности второго (модуляционного) плана уходят в область лингвистических закономерностей артикуляции слитной речи. Поэтому их задание и описание в модели дикторонезависимого распознавания слитной речи возможны только через задание и описание базовых (фонетических) элементов артикуляции слитной речи [16].

3. Описание артикуляции интегральными *AFB*-параметрами обобщённых формант спектров речи

Процедура получения и оценки *AFB*-параметров исходя из требований двухформантной модели описания артикуляции речи [8–10, 17, 18] включает ряд последовательных этапов: предварительную обработку речевого сигнала [17], согласованный спектральный анализ [18], локализацию обобщённых формант спектров речи, и, наконец, оценку *AFB*-параметров обобщённых формант [8–10, 17, 18].

3.1. Модель предобработки речевых сигналов в задаче получения интегральных AFB-параметров описания артикуляции речи

Преобразование речевого сигнала на первом этапе основывается на отличительных свойствах первой и высших формант речи. При этом учитываются следующие два основных свойства.

1. Первый формантный максимум на огибающей спектра для большинства звуков речи является глобальным (наибольшим) при условии соответствующей коррекции спектра голосового источника по методике, предложенной в работе [19].

2. Динамические диапазоны высших формант, связанные с их частотной перестройкой, существенно больше динамического диапазона первой форманты.

Остановимся более подробно на втором из названных свойств. Динамический диапазон речевого сигнала на входе приёмника (ухо, устройство оценки артикуляции, система распознавания речи) складывается из диапазонов изменений [20]: уровней звуков речи — до 45 дБ; громкости речи дикторов — до 15 дБ; затухания телефонных каналов передачи речи — до 10 дБ; расстояния до микрофона (телефонной трубки) от губ говорящего — до 15 дБ. В итоге средний динамический диапазон изменений уровней речевого сигнала на входе анализирующей (распознающей) системы равен 85 дБ. Однако для построения алгоритма оценивания формантных параметров, инвариантного от изменений уровней речевого сигнала, необходимо учесть также диапазоны изменений уровней формантных составляющих речевого сигнала. Это можно сделать, опираясь на закономерности связей уровней формант с их перестройками в соответствующих частотных диапазонах [5]. Оказывается, что изменения значащих уровней первой форманты укладываются в диапазон 6 дБ, а второй и высших формант — в диапазон 30 дБ относительно уровня первой форманты. Тогда общий (с учётом диапазона изменений уровней речевого сигнала — 85 дБ) диапазон изменений уровней первой форманты составит в среднем 6 + 85 = 91 дБ, а общий диапазон изменений уровней второй и высших формант — соответственно 6 + 30 + 85 = 121 дБ.

Таким образом, диапазоны изменений нужных нам компонент речевого сигнала различны и существенно больше диапазона уровней собственно речевого сигнала. Это свойство формантных компонент речевого сигнала указывает на целесообразность их предварительной отфильтровки и сжатия (компрессии) их динамических диапазонов. Кроме того, и это принципиально важно, раздельная компрессия отфильтрованных сигналов позволяет нормировать (уменьшить) вариативности уровней анализируемых сигналов от действия следующих факторов:

— разброса динамических диапазонов первой и высших формант по множеству дикторов;
— перекосов (изменений) частотных характеристик микрофонов и трактов передачи речевого сигнала.

Предварительная обработка речевого сигнала, заключающаяся в разделении его на две названные выше компоненты с последующей компрессией получаемых сигналов, с одной стороны, реализует часть методики [8 –10, 17, 18] получения *AFB*-параметров, а с другой — позволяет существенно сократить информационный поток оцифрованных сигналов и минимизировать тем самым последующие вычислительные затраты. Разработанный нами программный модуль SoftBoard реализует предварительную двухполосную обработку речевых сигналов [17]. Заметим также, что одновременно при этом решается задача повышения и стабилизации (нормализации) разборчивости речевых сигналов в условиях изменчивости параметров голосов дикторов и телефонных каналов передачи путём усиления и нормирования уровней информативных компонент (верхних формант) речевых сигналов.

Отметим, что проблема повышения качества и разборчивости речевых сигналов ранее активно исследовалась только применительно к системам связи [21–24], а позднее и применительно к прикладным системам распознавания речи [25] в связи с тем, что для современных систем распознавания речи речевой сигнал оказывается «недостаточно разборчивым». По этой причине для повышения «разборчивости речевых сигналов» для систем распознавания речи разрабатываются различные методы предобработки речевых сигналов, которые дают ощутимые положительные результаты [25], что и подтверждает актуальность реализуемой с помощью программного модуля SoftBoard предварительной обработки речевых сигналов.

3.2. Модель согласованного спектрального анализа речевых сигналов в задаче получения интегральных AFB-параметров описания артикуляции речи

Спектральный анализатор речевого сигнала должен адекватно отображать формантные свойства как голосовых, так и шумовых сигналов речи. Для удовлетворения этого требования модель разложения речевого сигнала в ряд Фурье должна быть согласована с моделью образования речевого сигнала. Если мы обратимся к формантной модели речевого сигнала [5]

$$f(t) \approx \sum_{n=1}^{r} A_n e^{\sigma n (t - m T_0)} \, sin\left[2\pi F_n (t - m T_0) + \theta_n\right]$$

и сравним её с моделью обобщённого Фурье-разложения

$$g(t) \approx \sum_{k=1}^{\nu} A_k \varphi_k(t),$$

то становится очевидным, что речевой сигнал аппроксимируется последовательностью затухающих синусоид частотой F_n с периодом основного тона F_0. Каждое отдельно взятое формантное колебание

$$\varphi_n(t) \approx A_n e^{\sigma n (t - m T_0)} \, sin\left[2\pi F_n (t - m T_0) + \theta_n\right] \qquad (2)$$

представляет собой синусоидальное колебание с частотой F_n, модулированное по амплитуде частотой F_0. Следовательно, каждая формантная компонента речевого сигнала есть сложный широкополосный сигнал. Для согласованной фильтрации таких сигналов требуется гребёнка полосовых фильтров с полосами пропускания не менее 2 F_0. Модель широкополосного спектрального

анализатора согласуется с моделями спектрального анализа случайных сигналов [26], к которым относятся шумовые сигналы речи, и не противоречит свойствам слухового анализа акустических сигналов [27].

Для построения спектрального анализатора использованы цифровые фильтры 2-го порядка, описываемые уравнением:

$$Y(t,i) = \{2*Y(t-1,i) - X(t-1)\}*k_1(i) - Y(t-2,i)*k_2(i) + X(t), \qquad (3)$$

где $Y(t, i)$, $Y(t-1, i)$, $Y(t-2, i)$ — значения выходных сигналов i-го фильтра в t-й, $t-1$, $t-2$-й от-счёты времени; $X(t)$, $X(t-1)$ — значения РС в t-й и $t-1$-й отсчёты времени; $k1(i)$, $k2(i)$ — коэффициенты, определяющие частоту настройки и полосу пропускания i-го фильтра.

Коэффициенты фильтра связаны с центральной частотой $f(i)$ и полосой пропускания $B(i)$ фильтров следующими соотношениями:

$$k_1(i) = \left(1 - \pi * B(i)/f_d\right) * \cos\left(2*\pi * f(i)/f_d\right),$$

$$k_2(i) = \left(1 - \pi * B(i)/f_d\right)^2,$$

где π = 3,14 — константа; fd — частота дискретизации речевого сигнала.

Для лучшего согласования фильтров анализатора с формантными компонентами (2) речевых сигналов процедура (3) фильтрации реализуется последовательно дважды.

4. Сравнение формантных свойств исходных и преобразованных речевых сигналов

Из предыдущего мы знаем, что первой в ряду ступеней предобработки речевых сигналов является раздельная фильтрация первых двух формантных сигналов речи с последующим усилением и нормированием их уровней. На рисунках 2–5 приведены сравнительные изображения осциллограмм входных сигналов (верхние осциллограммы) и откорректированных в первых двух формантных областях (нижние осциллограммы) сигналов для звуков ***а*** и ***i***.

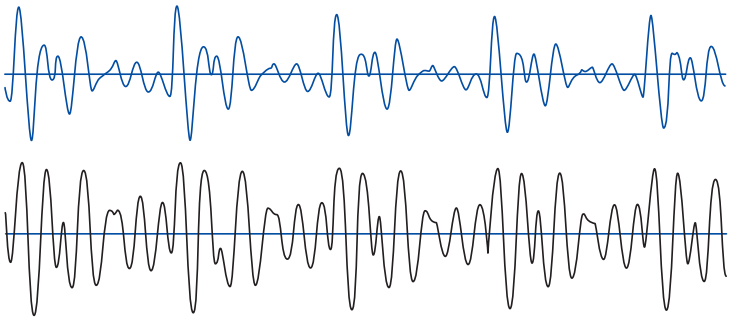


Рис. 2. Результат формирования сигнала ww1 первой форманты (нижний рисунок) для реализации звука «а» (верхний рисунок)

Верхние осциллограммы на рисунках 2 и 3 иллюстрируют сумму двух первых формант, близких по частоте (для звука а). Нижние осциллограммы демонстрируют эффективное подчёркивание (локализацию) и равнозначное с первым сигналом взвешивание (нормирование) сигнала второй форманты.

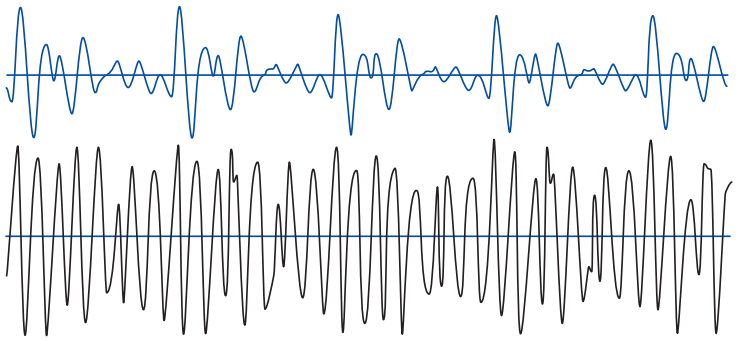


Рис. 3. Результат формирования сигнала ww2 второй форманты (нижний рисунок) для реализации звука «a» (верхний рисунок)

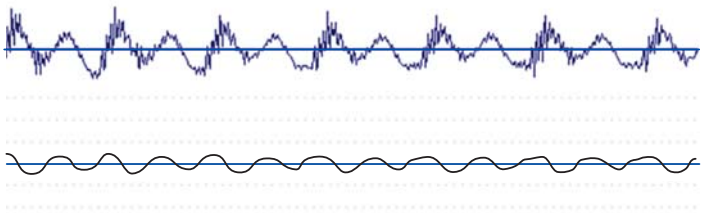


Рис. 4. Результат формирования сигнала ww1 первой форманты (нижний рисунок) для реализации звука «i» (верхний рисунок)

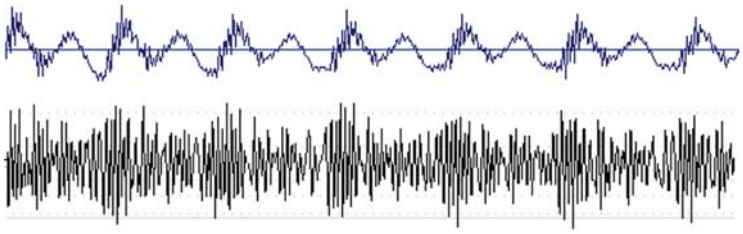


Рис. 5. Результат формирования сигнала ww2 второй форманты (нижний рисунок) для реализации звука «i» (верхний рисунок)

Особенно ярко эти свойства проявляются на иллюстрациях (*см. рис. 4 и 5*) для звука i ввиду большой разницы частот первых двух формант.

На верхних осциллограммах хорошо видно, что сигнал второй форманты (высокий по частоте) на периоде основного тона быстро затухает, в результате чего он проявляется лишь в виде «вспышек» на коротких интервалах времени на периоде основного тона речи.

На нижней осциллограмме *рис. 5* мы наблюдаем эффективное подчёркивание сигнала второй форманты на всём протяжении периода основного тона, и, следовательно, на протяжении всего времени реализаций соответствующих звуков речи.

Сопоставление изображений осциллограмм входных сигналов (верхние осциллограммы) и первых двух формант (нижние осциллограммы) для звуков a и i, представленных на *рис. 2–5*, показывает эффективность локализации и подчёркивания формантных компонент речевого сигнала, которая иллюстрируется в виде чёткой локализации интенсивности «формантных треков» на широкополосных динамических спектрах преобразованных сигналов на *рис. 6a*. На *рис. 6* представлены динамические спектры реализации слитно произнесённой женским голосом (диктором) фразы «ноль два».

Сопоставление изображений осциллограмм входных сигналов (верхние осциллограммы) и откорректированных в первых двух частотных формантных областях (нижние осциллограммы) для звуков a и i, представленных на *рис. 2–5*, показывает эффективность локализации и подчёркивания формантных компонент речевого сигнала, которая иллюстрируется в виде чёткой локализации «формантных треков» на широкополосных динамических спектрах преобразованных сигналов на *рис. 6a*.

Необходимо отметить также принципиально важное свойство получаемого спектрального отображения «формантных треков» на *рис. 6а*, на которых отсутствуют проявления гармонической структуры, характерные для динамических спектров речи с высокой частотой основного тона голоса, и которые наблюдаются на стандартной спектрограмме этой фразы (см. *рис. 6б*).

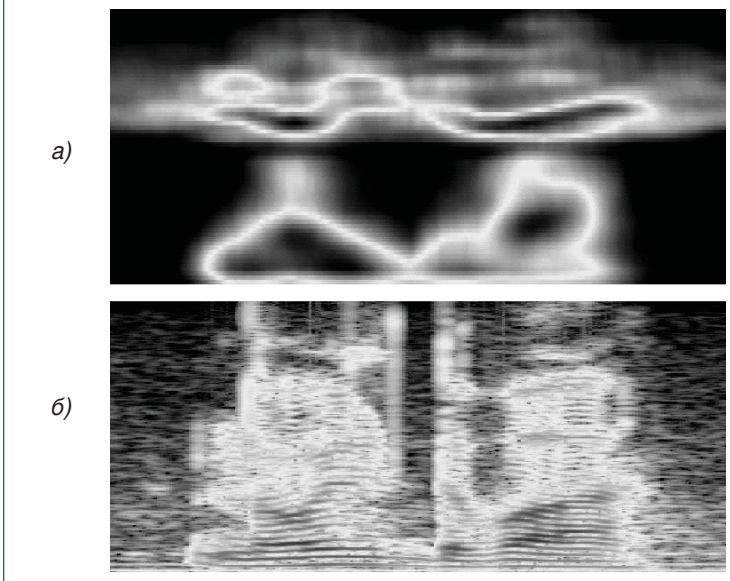


Рис. 6. Динамические спектры реализации слитно произнесённой женским голосом фразы «нольдва»:
а) оригинальное отображение программой VideoSpeech широкополосных динамических спектров преобразованных компонент (см. рисунки 2–5) речевого сигнала;
б) стандартное отображение программой Sound Forge исходного речевого сигнала.

5. Алгоритм получения интегральных *AFB*-параметров описания артикуляции речи

Для построения системы параметров описания артикуляции проанализируем общие свойства формант речи, устойчиво проявляющиеся на огибающих спектрах речевых сигналов. Можно выделить четыре типичных вида огибающих спектров, отображающих основные свойства артикуляции способов образования звуков речи (*рис. 7*).

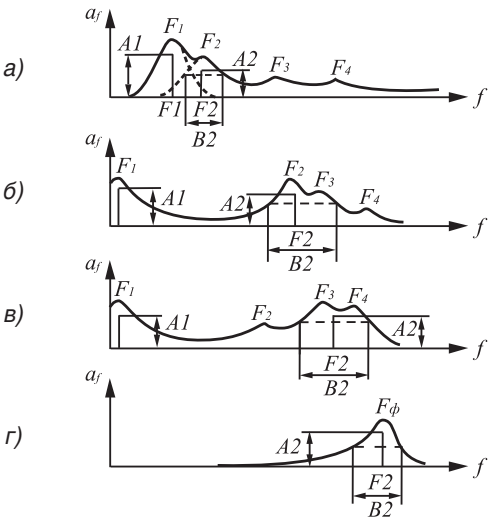


Рис. 7. Типичные формантные группы и их обобщённые параметры для звуков речи:
а) компактных и аспиративных;
б) диффузных;
в) носовых и звонких фрикативных;
г) глухих фрикативных

В работах [8–10] мы предложили систему интегральных параметров ***A1 , F1, A2, F2, B2*** хорошо отображающих формантные свойства огибающих спектров речи (*рис. 7*). Основная идея алгоритма оценивания названных параметров состоит в том, что в двух областях спектра речи с адаптивной границей их разделения формантные группы описываются моментами от отсчётов спектра, группирующихся около максимального из них:

$$A = \frac{\sum_{1}^{m} a_y}{m}; \qquad F = \frac{\sum_{1}^{m} F_y a_y}{\sum_{1}^{m} a_y};$$

$$B = \frac{\sum_{1}^{m} (F_y - F_{y-1}) a_y}{\max\limits_{j} a_j}, \qquad (4)$$

где a_j — отсчёты мгновенного спектра мощности на частотах F_j; $j = \overline{1,n}$; $y = argz_y$; $z_y = sign(a_j - \frac{h max a_j}{j})$; $0 < h < 1$; m — число отсчётов, превысивших порог $h \max\limits_{j} a_j$.

Физический смысл оцениваемых согласно соотношениям (4) параметров состоит в том, что они выражают средневзвешенные (интегральные) значения амплитуды A, частоты F и ширины B выделенных отсчётов спектра a^y, представляющих данную формантную группу.

Свойства параметров (4) существенно зависят от значения коэффициента h. При $h \to 1$ параметры A и F представляются амплитудой и частотой максимального отсчёта спектра, а при $h \to 0$ — соответственно интенсивностью и средней частотой, выраженными через момент нулевого и первого порядка от выделенных отсчётов спектра. Параметр B связан с эффективной шириной спектра, а значение коэффициента h определяет степень его чувствительности к модуляциям ширины спектра. Кроме того, коэффициент h влияет на число выделенных отчётов m, группирующихся около максимального отсчёта и определяющих значения и интегральные свойства параметров (4). Поэтому значения коэффициента h оптимизируются для каждой из двух аппроксимирующих формантных групп в зависимости от амплитудных отношений составляющих их формант. Тем самым обеспечивается свойство несмещённости получаемых согласно алгоритму (4) оценок интегральных формантных параметров. Выбор формантных групп и основные принципы (алгоритм) обработки их спектрального представления сводятся к следующему.

1. В частотных границах первой форманты звонкие звуки речи образуют формантные группы, состоящие не более чем из двух первых формант. При этом амплитуда первой форманты в таких группах, как правило, является большей. Поэтому выбор значения h в пределах 0,5–0,8 обеспечивает хорошую корреляцию параметров A^1 и F^1, определяемых по выражениям (4), с амплитудой и частотой первой форманты. Названное выше свойство первой форманты позволяет также обнаруживать первую формантную группу по максимальному отсчёту спектра в диапазоне её существования.

2. По найденному значению F^1 производится инверсная фильтрация отсчётов спектра первой форманты (см. *рис. 2а*), что обеспечивает эффективное разделение первой и второй формант в случаях, когда они сближаются так, что составляют одну формантную группу.

3. Полученные указанным способом спектральные отсчёты второй обобщённой формантной группы описываются параметрами A^2, F^2 и B^2, определяемыми соответственно (4). При выборе значения коэффициента h в пределах 0,3–0,6 названые параметры хорошо отображают амплитудно-частотные отношения второй и более высоких голосовых формант, частотное положение и эффективную ширину фрикативных и аспиративных формант.

Таким образом, двухформантная модель описания артикуляции интегральными *AFB*-параметрами обобщённых формант спектра речи не требует использования

фонетического контекста и в то же время хорошо отображает связанные с ним формантные свойства спектрально-временного описания речевых сигналов (*рис. 8*).

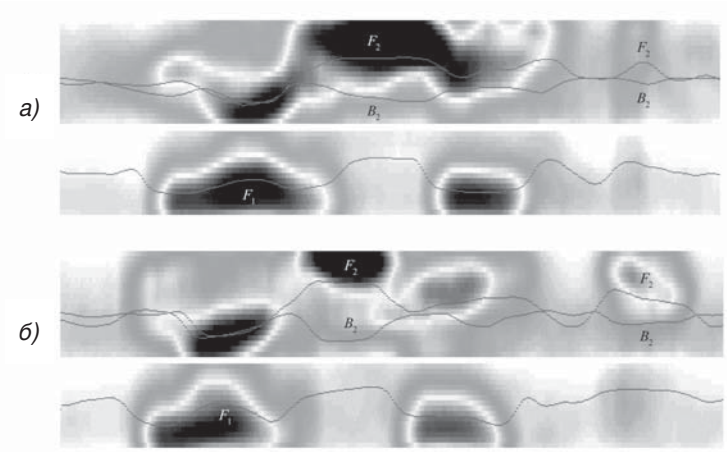


Рис. 8. Исходные спектрограммы и соответствующие им базовые F1 F2 B2-параметры описания слитно произнесённой фразы ***«нольшесть»****:*
а) мужской голос;
б) женский голос

Основными достоинствами предложенной системы параметров и алгоритма их выделения являются:

— возможность разделения первых двух формант (*рис. 8, а*) даже в случае их взаимной маскировки;

— возможность отображения формантных свойств (*рис. 8, б* и *в*) без разделения верхних формант, представляющееся наиболее сложной задачей;

— универсальность параметров описания, выражающаяся в равной эффективности отображения формантных свойств спектров, различных по способу (звонкий, фрикативный) образования звуков речи;

— инвариантность (независимость) получаемых спектрограмм и соответствующих им *AFB*-параметров описания артикуляции речевых сигналов от типа голоса: мужской — женский, перекосов частотных характеристик микрофонов и каналов передачи;

— повышенная надёжность и помехоустойчивость интегрального принципа (4) оценивания *AFB*-параметров. Последнее свойство интегральных методов оценки формантных свойств спектров речи позднее было замечено и другими исследователями [28].

В связи с контекстуальной зависимостью формантное описание речи имеет изменчивую структуру значащих параметров, связанную со способом образования звуков речи (*рис. 3*). Это принципиально важное свойство формантного описания в полной мере может быть учтено при построении оценок формантных параметров методами анализа через синтез. И хотя алгоритм (4) получения *AFB*-параметров контекстуально независим, тем не менее само *AFB*-описание, будучи связанным с формантным, естественным образом отображает структурно-параметрические свойства различных по способу образования звуков речи (см., например, сегменты S_2, S_4, S_7 на *рис. 9*). Задача состоит в том, чтобы найти способ моделирования этого явления.

Рассмотрим 8-параметрическое *AFB*-описание слитно произнесённой фразы «нольшесть» (*рис. 9*), где K_vF_1, K_vF_2 и K_vA_1 — нормированные производные от соответствующих параметров. Опишем анализируемую реализацию S-последовательностью сегментов $S_1 = S_1, S_2, \ldots, S_9$, каждый из которых отображается значащими параметрами, характеризующими их физическую принадлежность данному сегменту (на *рис. 9* значащие параметры

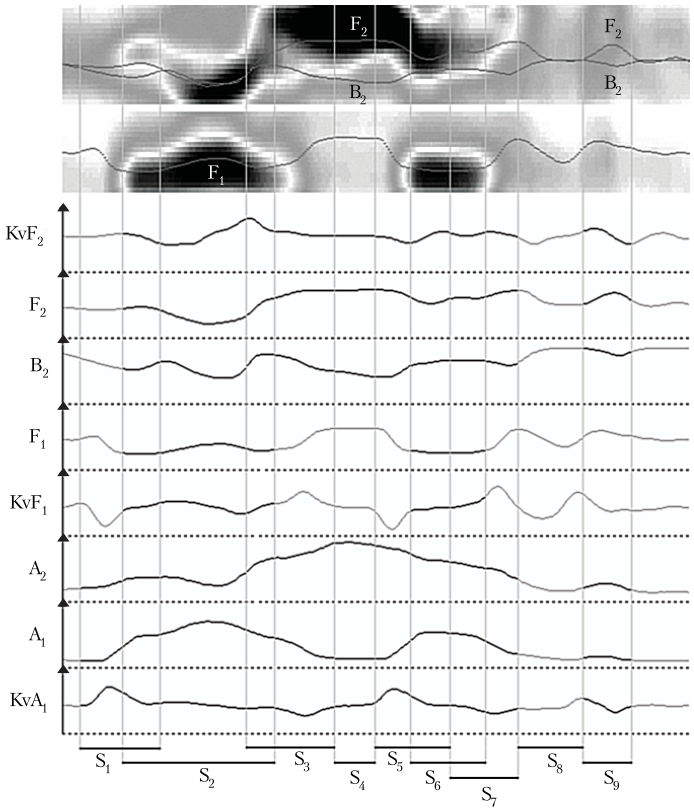


Рис. 9. Пример структурно-параметрического AFB-описания слитно произнесённой фразы «нольшесть»

контрастно выделены). Если теперь построить алгоритмы создания эталонных описаний элементов S_i их значащими параметрами, учитывающими все закономерности артикуляции слитной речи, и автоматической сегментации (описания описания) реализаций речевых сигналов в алфавите S_i эталонных элементов, то поставленная задача будет решена.

6. Оценка артикуляции элементов речи

Введём понятие артикулемы как интересующего нас (значащего) элемента (S_i сегмента) артикуляции речи. Пусть нами создана некоторая база данных N словаря артикулем. Для оценки [29] меры сходства D^n реализации n-го эталона из N словаря артикулем используем ДП-алгоритм [30] с оптимальным для речевых сигналов коэффициентом ограничения деформации времени сравниваемых сигналов, равном двум:

$$D(i,j) = min \begin{cases} D(i-1,j-2) + 2d(i,j-1) + d(i,j) \\ D(i-1,j-1) + 2d(i,j) \\ D(i-2,j-1) + 2d(i-1,j) + d(i,j), \end{cases} \quad (5)$$

где $i = \overline{0,I}$ — отсчёты параметрического описания реализации; $i = \overline{0,J}$ — отсчёты параметрического описания эталона.

Здесь функция локального расстояния $d^n_{i,j}$ между отсчётами i реализации и отсчётами j^n -го эталона определяется в соответствии с выражением

$$d^n_{i,j} = \frac{1}{k}\sum_{1}^{k} \beta_a \frac{\left|P_{a,i} - P^n_{a,j}\right|}{P_{a,i} + P^n_{a,j}}, \quad (6)$$

где $0 \le \beta \le 1$ — коэффициент «взвешивания» параметра P_a; $a = \overline{1,k}$ — индексы активных параметров n-го эталона.

Оценка расстояния D^n и обнаружение n-го эталона, удовлетворяющего условию достаточного правдоподобия, определяются из соотношений

$$D^n = \min_i D^n(i, J^n), \quad D^n \le 0{,}5P_M^n, \qquad (7)$$

где P_m^n — масштабирующий параметр, характеризующий «фонетический» вес n-го эталона.

Окончание подобия n-го эталона находится в соответствии с правилом

$$i_n^K = arg \min_i D^n(i, J^n). \qquad (8)$$

Функция $D^n(i,J^n)$ текущего расстояния n-го эталона к реализации речевого сигнала, используемая при решении задач (5) и (6), вычисляется с помощью рекуррентного ДП-уравнения (5), модифицированного для случая свободного (незакреплённого) начала реализации. Для n-го эталона, удовлетворяющего условию достаточного правдоподобия (7), относительно найденного i_n^K окончания подобия на обратном временном окне длиною $2J^n$ решается задача (5) определения его начала i_n^H с помощью модификации алгоритма (5) для закреплённого начала анализируемого процесса.

Нормированная мера подобия R^n найденного n-го эталона определяется следующим образом:

$$R^n = \frac{(P_M^n - D^n)\cdot 100}{P_M^n},$$

$$0 \le D^n \le P_M^n. \qquad (9)$$

Алгоритм (5) – (9) обнаружения и оценки артикуляции заданных элементов речи может быть использован как в задаче распознавания, так и в задаче обучения (оценивания) артикуляции слитной речи.

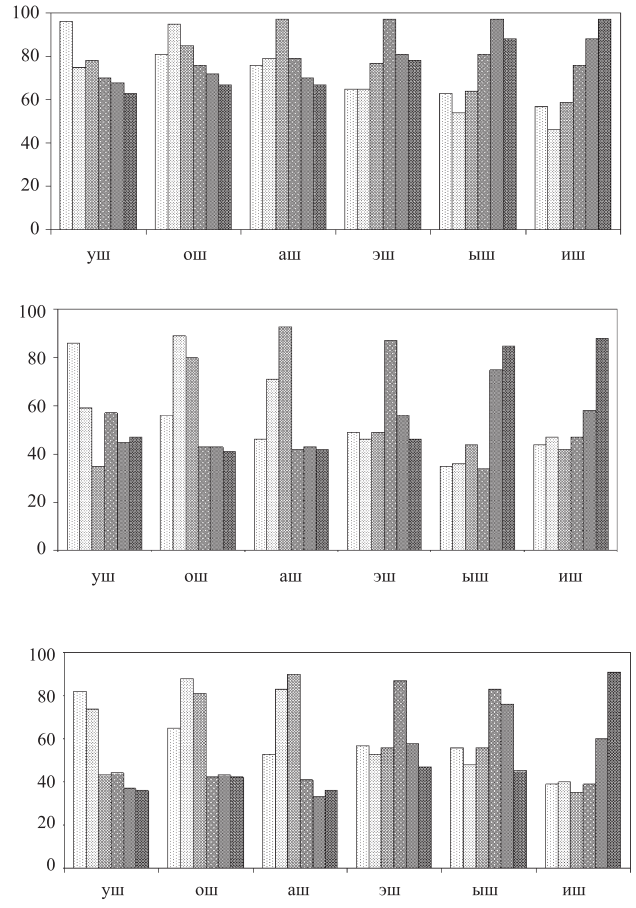


Рис. 10. Гистограммы корреляционных оценок артикуляции слогов

На *рис. 10* приведены гистограммы корреляционных оценок (9) артикуляии выделенных слогов при произнесении слов суша, ноша, саша, кэш, крыша, ниша в указанной последовательности.

Гистограммы на *рис. 10 а* отображают корреляционные оценки артикуляции выделенных слогов для одного диктора, полученные с помощью алгоритма (5)–(9) для случая описания артикуляции совместно параметрами способа и места образования. Гистограммы на *рис. 10 б* и *рис. 10 в* отображают соответствующие оценки для двух разных дикторов при описании артикуляции только параметрам места образования. Причём необходимые оценки в этом случае получены в два этапа: на первом этапе с помощью алгоритма (5)–(9) по параметрам способа образования найдены границы слогов, а на втором — искомые оценки при описании артикуляции только параметрами места образования, что указывает на более широкие возможности построения акустических оценок атикуляции элементов речи с помощью *AFB*-параметров первичного описания речевых сигналов.

Из сравнения данных *рис. 10 а* и *рис. 10 б, в* видно, что во втором случае полученные оценки, сохраняя свойство выраженной корреляции с адекватными типами артикуляции для разных дикторов, одновременно обладают большей различающей способностью (информативностью). Отмеченные свойства указывают на преимущества и актуальность построения локальных фонетических метрик оценивания артикуляции элементов слитной речи.

Выводы

1. Задача оценки формантных параметров речевых сигналов не имеет корректного решения вне контекста способа образования анализируемых сегментов.

2. Предложено измерять не сами формантные параметры речевых сигналов, а *AFB*-параметры обобщённых формант динамических спектров речи.

3. Предложенные система обобщённых *AFB*-параметров первичного описания артикуляции речи и метод обнаружения и оценки артикуляции заданных элементов речи могут составить хорошую основу для решения задач распознавания слитной речи.

4. Предложенные *AFB*-параметры описания речевого сигнала хорошо отображают формантные свойства спектров речи и в связи с этим удовлетворяют требованиям линейной модели [7] аппроксимации параметров описания речевых сигналов, в рамках которой возможно создание топологических инвариантов (относительно характеристик каналов передачи и голосов дикторов) описания артикуляции (артикулем) слитной речи [31].

5. Использование *AFB*-параметров описания речевого сигнала в сочетании с современными моделями описания фонетического уровня [16] в системах автоматического распознавания речи даёт надежду на определённое продвижение в решении проблемы многодикторного распознавания слитной речи.

6. И наконец, названные выше инвариантные свойства *AFB*-параметров описания позволяют выйти на решение прикладных задач многодикторного распознавания слитной речи в условиях изменчивости АЧХ среды и каналов передачи речевых сигналов.

Литература

1. *Бухтилов Л.Д., Лобанов Б.М.* Алгоритм оценки формантных частот // Автоматическое распознавание слуховых образов (АРСО-14). Каунас, 1986. Ч. 1. С. 10–11.
2. *Deng L., Ma J.* Spontaneous Speech Recognition Using a Statistical Coarticulatory Model for the Vocal-Tract-Resonance Dynamics // J. of American Society of Acoustics. Vol. 108. № 6. 2000. P. 3036–3048.
3. *Lee M., Santen J., at all.* Formant tracking using segmental phonemic information // Proc. of the European Conf. on Speech Commun. and Techn. Eurospeech '99. Budapest, Sept. 5-9, 1999. Vol. 6. P. 2789–2792.
4. *Lobanov B.* On the Way to Precise and Robust Formant Frequencies Tracking / B. Lobanov, A. Davydau // Speech and Computer: proceedings of the 13th International conference SPECOM'2009, St. Petersburg, Russia, 21–25 June, 2009 / St. Petersburg Institute for Informatics and Automation of RAS (SPIIRAS). St. Petersburg: Anatolia, 2009. P. 340–344.
5. *Фант Г.* Акустическая теория речеобразования. М.: Наука, 1964.
6. *Фланаган Д.Л.* Анализ, синтез и восприятие речи. М.: Связь, 1968.
7. *Винцюк Т.К.* О математических моделях речевого сигнала, используемых в распознавании речи // Автоматическое распознавание слуховых образов (АРСО-12). Киев, 1982. С. 34–37.
8. *Дегтярев Н.П.* Двухформантная аппроксимация спектров речи // Автоматическое распознавание слуховых образов (АРСО-14). Каунас, 1986. Ч. 1. С. 12–13.
9. *Дегтярев Н.П.* Акустическое описание артикуляции параметрами обобщённых формант спектра речи // Автоматическое распознавание слуховых образов (АРСО-15). Таллинн, 1989. С. 145–149.
10. *Degtjarev N.P.* Two-Formant Model of the Acoustic Description of Speech Articulation // Proceedings of the XII-th International Congress of Phonetic science. France, Aix-en-Provence, 1991. Vol. 2. P. 410–413.
11. *Чистович Л.А., Кожевников В.А. и др.* Речь. Артикуляция и восприятие. М.-Л.: Наука, 1965.
12. *Гвоздев А.Н.* Современный русский литературный язык. Ч. 1.Фонетика и морфология. М.: Просвещение, 1973.
13. Русская разговорная речь / Под ред. Е.А. Земской. М.: Наука, 1973.
14. *Рабинер Л.Р., Шафер Р.В.* Цифровая обработка речевых сигналов / Пер. с англ. М.: Связь, 1981.
15. *Пирогов А.А.* К вопросу о фонетическом кодировании речи // Электросвязь. 1967. № 5. С. 24–31.
16. *Коваль С.Л., Смирнова Н.С., Хитров М.В.* К проблеме разработки фонетического уровня в системах автоматического распознавания речи // Труды Междунар. семинара Диалог '2002 по компьютерной лингвистике и её приложениям, Т.2. М., 2002. С. 197–206.
17. *Дегтярев Н.П.* Выбор модели предобработки и спектрального анализа речевых сигналов в задаче получения AFB-параметров описания артикуляции речи // Автоматическое распознавание слуховых образов (АРСО-17). Ижевск, 1992. С. 176–180.
18. *Дегтярев Н.П.* Выбор модели спектрального анализатора речевых сигналов // Вычислительная техника и краевые задачи. Процессоры цифровой обработки сигналов. Рига, 1992. С. 89–96.
19. *Дегтярев Н.П.* Погрешности анализа формантных частот методом неадаптивной фильтрации // Автоматическое распознавание слуховых образов (АРСО-11). Ереван, 1980. С. 58–61.
20. *Вемян Г.В.* Передача речи по системам электросвязи. М.: Радио и связь, 1985.
21. *Сапожков М.А.* Защита трактов радио и проводной телефонной связи от помех и шумов. М.: Связьиздат, 1959.
22. *Бандура Н.В., Бухвинер В.Е., Добровольский Е.Е.* Управляемые компандеры в радиосвязи и радиовещании. Оценка эффективности // Электросвязь. 1974. № 12. С. 36–40.
23. *Рыффа В.Н.* Повышение разборчивости речи путём сжатия динамического диапазона // Электросвязь и передача данных. Киев, 1969.
24. Optimum Lianear Filter for Speech Transmission // The Journ. of the Acoust. Soc. of Amer. Vol. 43. № 1. 1968. P. 81–86.
25. *Sadaoki Furui.* Perspectives of Speech Processing Technologies. International Workshop «Speech and Computer», Specom'98 St.-Petersburg, October 26–29 1998. P. 1–6.
26. *Харкевич А.А.* Спектры и анализ. М.: Гостехиздат, 1963.

27. *Молчанов А.П., Лабутин В.К.* Механизмы анализа сигналов в органе слуха и проблемы их моделирования // Распознавание слуховых образов / Под ред. Н.Г. Загоруйко и Г.Я. Волошина. Новосибирск: Наука СО, 1970. С. 142–204
28. *Gajirc B., Paliwal K.* Robust Parameters for Speech Recognition Based on Subband Spectral Centroid Histograms // Proc. of the European Conf. on Speech Commun. and Techn. Eurospeech '01. Scandinavia, 2001. Vol. 1. P. 591–594.
29. *Дегтярев Н.П., Черников Д.А.* Формантное отображение и оценка артикуляции речи // Анализ цифровых изображений. Минск: ОИПИ НАН Беларуси, 2003. Вып. 2. С. 174–185.
30. *Sakoe H.* Two-level DP — Matching — A Dynamic Programming Based on Pattern matching Algorithm For Connect Word Recognition // IEEE Trans on ASSP. Vol. 27. № 6. 1979. P. 588–595.
31. *Дегтярев Н.П.* Модуляционная основа инвариантов акустического описания артикуляции речи // Автоматическое распознавание слуховых образов (АРСО-16). Суздаль, 1991. С. 106–107.
32. *Дегтярев Н.П.* Параметрическое и информационное описание речевых сигналов. Минск: Объединенный институт проблем информатики Национальной академии наук Беларуси, 2003.

Дегтярев Николай Петрович —

главный конструктор проектов отдела совместных программ космических и информационных технологий Объединенного института проблем информатики Национальной академии наук Беларуси, автор монографии и более 50-ти научных публикаций в области речевых технологий, 8-ми авторских свидетельств на изобретения, награжден знаком «Изобретатель СССР» и серебряной медалью ВДНХ СССР.

Система синтеза речи по тексту для мобильных телефонов

Л.И. Цирульник,
кандидат технических наук, доцент

Д.А. Покладок,
аспирант

Системы синтеза речи по тексту широко применяются на персональных компьютерах. Использование синтезаторов речи по тексту на мобильных телефонах очень ограничено, поскольку последние характеризуются низким быстродействием и малым объёмом памяти, что не позволяет напрямую «перенести» на них уже существующие синтезаторы.

В статье предлагается новая архитектура системы синтеза речи по тексту, в которой текст обрабатывается на сервере, а речевой сигнал — на телефоне. Описываемые алгоритмы обработки речевого сигнала имеют линейную вычислительную сложность и позволяют синтезировать речевой сигнал в реальном масштабе времени.

Abstract

Nowadays Text-To-Speech synthesis systems are widely used on personal computers. But usage the TTS-systems on mobile phones is restricted, because they are characterized by small memory and low operating speed. This fact do not allow transferring directly the existing TTS-systems to the mobile phones.

In this paper a new architecture of the TTS-synthesis system is proposed. It assumes that the process is distributed between the server, that does text processing, and the phone, that per-forms speech signal processing. The described algorithms have linear computational complexity and allow synthesizing the speech signal in real time.

Введение

Системы синтеза речи по тексту к настоящему моменту достигли высокого качества как по критериям разборчивости и естественности синтезируемого голоса, так и по техническим характеристикам, что способствует их широкому применению

в практических приложениях, например, в центрах обработки вызовов, при управлении сложными объектами, для создания аудиокниг и т.д. Всё более широкое распространение получает использование систем синтеза речи на мобильных устройствах, таких как карманные персональные компьютеры или смартфоны. Это и озвучивание SMS-сообщений, и чтение писем электронной почты, и озвучивание указаний автомобильной навигационной системы. Использование систем синтеза речи на мобильных телефонах, как отмечалось, ограничено из-за низкого быстродействия и относительно небольшого объёма памяти. Современные системы синтеза речи требуют большого объёма памяти для хранения лингвистических и акустических ресурсов, что не позволяет напрямую «перенести» существующие системы на мобильные платформы.

Существуют три возможные схемы работы системы синтеза речи по тексту на мобильных телефонах:

1. Серверная, при которой система синтеза речи полностью расположена на сервере. Абоненту на мобильный телефон передаётся синтезированный речевой сигнал.

2. Клиентская, при которой система синтеза речи расположена полностью на мобильном телефоне.

3. Распределённая, при которой система синтеза речи частично расположена на сервере, частично — на мобильном телефоне.

Первая схема реализована, в частности, компанией MATERNA Information & Communications GmbH для предоставления услуги SMS2Voice (SMS2Fix) пользователям некоторых мобильных операторов в России и Украине [1]. Услуга позволяет отправлять текстовые сообщения, которые передаются синтезированным голосом на мобильные и стационарные номера.

Достоинствами данной схемы являются: возможность выбора метода синтеза, обеспечивающего наилучшее качество синтезируемой речи (поскольку нет ограничений на объём памяти и быстродействие); возможность воспользоваться данной услугой любому пользователю вне зависимости от технических характеристик его телефона; возможность модификации и обновления системы синтеза речи независимо от пользователей; высокая степень защиты системы от нелегального использования.

Очевидно, что подобным образом можно было бы передавать на мобильный телефон не только SMS-сообщения, но и любую информацию, озвученную на сервере, с использованием системы синтеза речи.

Однако эта схема имеет следующие недостатки: абонент услышит сообщение только один раз; передаваемая на мобильный телефон речевая информация имеет в несколько раз больший объём, чем исходная текстовая информация, что влечёт дополнительную существенную нагрузку на канал связи; прекращение функционирования хотя бы одного узла вызывает остановку всей службы.

Вторая из перечисленных схем получила достаточно широкое распространение. Именно по такой схеме работают программы Acapela TTS for Windows

Mobile [2], Nuance TALKS [3], Mobile Speak [4] и др. В этих продуктах синтез речи по тексту полностью осуществляется на смартфонах под управлением операционных систем Windows Mobile или Symbian.

При всех очевидных преимуществах данной схемы она имеет существенный практический недостаток: смартфоны, на которых возможна работа этих систем синтеза речи, составляют только 7% рынка мобильных телефонов [5].

Третья схема до настоящего времени не была реализована, хотя она имеет высокий потенциал. Принцип работы в этой схеме основан на разделении операций между сервером и мобильным телефоном: обработка текста выполняется на сервере, в то время как работа с речевым сигналом осуществляется на мобильном телефоне. Преимущества данной схемы: возможность сохранять озвученные сообщения на телефоне; возможность выбирать просодические стили и различные голоса для синтеза; возможность использования на большинстве мобильных телефонов.

В данной работе описывается система синтеза речи по тексту для мобильных телефонов, для реализации которой выбрана последняя из описанных схем. В первом разделе представлена архитектура разработанной системы; блок обработки речевого сигнала, который работает на мобильном телефоне, описан в разделе 2; раздел 3 посвящён описанию особенностей программной реализации блока обработки речевого сигнала на языке программирования Java. Раздел 4 — заключение — суммирует основные положения данной статьи.

1. Общая структура системы синтеза речи по тексту

Система синтеза речи по тексту (*рис. 1*) содержит два основных блока: блок преобразования текста и блок работы с речевым сигналом [6]. На первом этапе входной орфографический текст преобразуется в последовательность просодических синтагм с указанием интонационного типа каждой синтагмы, причём синтагма представлена последовательностью аллофонов (оттенков фонем в речевом потоке). На втором этапе из базы данных (БД) звуковых волн аллофонов извлекаются требуемые аллофоны, вычисляются целевые значения частоты основного тона (F_0), амплитуды (A) и длительности (T) для каждого аллофона, звуковые волны аллофонов модифицируются в соответствии с целевыми просодическими значениями и соединяются в непрерывный речевой сигнал.

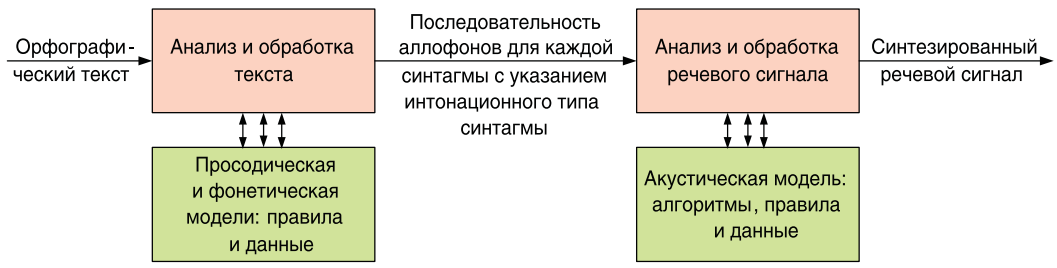


Рис. 1. Общая структурная схема системы синтеза речи по тексту

Блок анализа и преобразования текста (*рис. 2*) содержит модули лингвистической, просодической и фонетической обработки. Лингвистическая и просодическая обработка включают деление орфографического текста на фразы; преобразование чисел, аббревиатур, сокращений; деление фраз на просодические синтагмы; расстановку словесных ударений; деление синтагм на акцентные единицы (где под акцентной единицей понимается слово или группа слов с одним сильным ударением); маркировку интонационного типа синтагмы. Основными ресурсами лингвистического и просодического блоков являются грамматический словарь, а также правила морфологии и синтаксиса. Словарь используется для определения словесного ударения и лексико-грамматических характеристик каждого слова текста. Правила морфологии и синтаксиса используются для деления текста на фразы, фраз — на синтагмы, синтагм — на акцентные единицы, а также для определения интонационного типа синтагм.

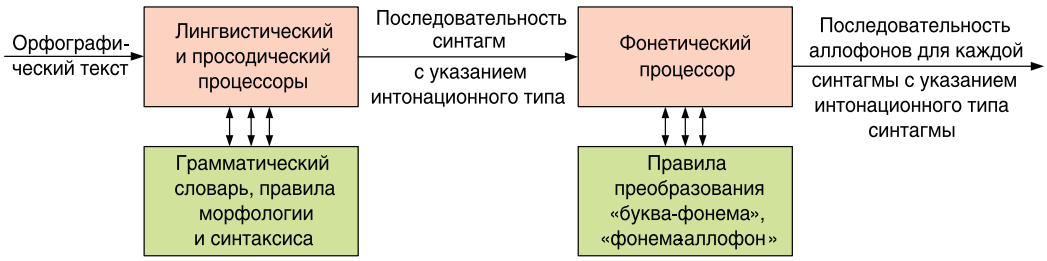


Рис. 2. Структура модуля обработки текста

Затем каждая интонационно размеченная синтагма поступает на фонетический процессор, который выполняет следующие задачи: фонетическое транскрибирование орфографического текста; определение позиционных и комбинаторных аллофонов; генерация аллофонных и мульти-аллофонных последовательностей, которые необходимо синтезировать.

Результат работы модуля обработки текста — последовательность синтагм с указанием интонационного типа каждой синтагмы, где каждая синтагма представлена последовательностью аллофонов — поступает в модуль обработки речевого сигнала.

В модуле обработки речевого сигнала (*рис. 3*) на первом этапе из речевой БД извлекаются речевые реализации аллофонов, соответствующие именам аллофонов во входной последовательности. Затем из БД просодических элементов извлекается просодический контур для соответствующего стиля и соответствующего типа синтагмы. После этого вычисляются целевые значения F_0, A, T. Такая последовательность шагов алгоритма обусловлена тем, что вычисление целевых значений F_0 должно осуществляться для каждого периода основного тона каждого вокализованного аллофона, а число периодов основного тона в аллофонах определяется после их извлечения из речевой БД.

Модуль обработки текста требует гораздо большего объёма памяти для хранения и использования ресурсов, чем модуль обработки речевого сигнала, а также характеризуется большей вычислительной сложностью. Действительно, один из основных лингвистических ресурсов — грамматический

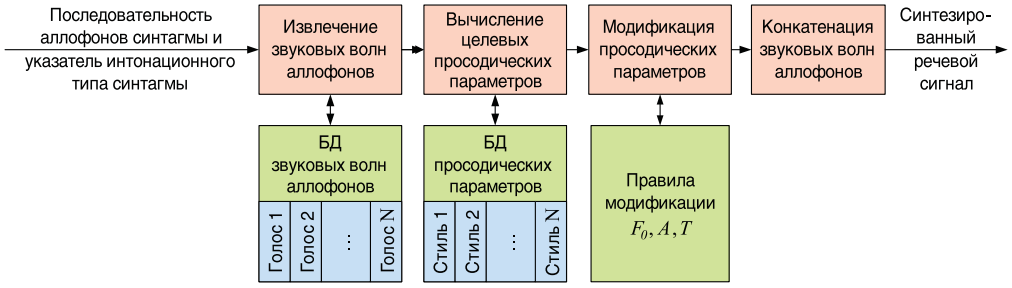


Рис. 3. Структура модуля обработки речевого сигнала

словарь русского языка — содержит более 3,5 миллиона словоформ [7]. Учитывая, что русский является флективным языком, целесообразно хранить словарь в виде компактной двухуровневой структуры, в которой первый уровень содержит неизменные части слов, а второй уровень — флексии. Для хранения в таком виде словаря объёмом 3,5 миллиона словоформ требуется порядка 50 мБ дискового пространства. Вычислительная сложность операций поиска слова в словаре равна $O(log_2 n)$, где n — количество слов в словаре. Вычислительная сложность всех операций, выполняемых лингвистическим, просодическим и фонетическим процессорами текста, включая операции поиска слова в словаре, равна $O(m)*O(n)$, где m — число слов входного текста.

Ресурсы блока обработки речевого сигнала — БД звуковых волн аллофонов и БД просодических параметров — требуют соответственно 750 кБ для одного голоса и 11 кБ для одного интонационного стиля. Вычислительная сложность алгоритмов обработки речевого сигнала равна $O(k)$, где k — количество аллофонов во входной последовательности.

При оценке алгоритмов синтеза речи по тексту важно учитывать тактовую частоту устройства, на котором должна быть реализована система, поскольку среднее время обработки одной синтагмы должно быть намного меньше, чем время воспроизведения синтезированной синтагмы, которое составляет в среднем от 1 до 10 секунд. Время обработки одной синтагмы на персональном компьютере с тактовой частотой 1,3 ГГц составляет 0,4–0,5 секунды.

Большинство современных мобильных телефонов обладает следующими характеристиками: доступная память от 128 КБ до 4 МБ, 32-битный RISC-процессор с тактовой частотой от 50МГц и выше, поддержка языка программирования Java ME и конфигурации CLDC. Такие характеристики не могут обеспечить достаточно быструю работу блока обработки текста, но удовлетворительны для быстрой работы блока обработки речевого сигнала.

Таким образом, оптимальна для реализации на большинстве современных мобильных телефонов архитектура, при которой блок обработки текста расположен на сервере, в то время как блок обработки речевого сигнала находится на мобильном телефоне. Дополнительным достоинством такой архитектуры является возможность синтеза речи по одному и тому же размеченному тексту, поступающему с сервера, с использованием различных голосов и различных просодических стилей, находящихся на мобильном телефоне.

2. Обработка речевого сигнала на мобильном телефоне

Из четырёх блоков обработки речевого сигнала, представленных на *рис. 3*, наибольший интерес представляют блоки вычисления целевых просодических параметров и модификации

просодических параметров в речевом сигнале. Особенности работы этих блоков описаны в данном разделе.

2.1. Блок вычисления целевых просодических параметров

Для вычисления целевых просодических параметров используется просодическая модель Портретов Акцентных Единиц (ПАЕ-модель) [8]. Согласно ПАЕ-модели, каждое предложение состоит из последовательности синтагм, где под синтагмой понимается самостоятельная в интонационном смысле часть предложения. Каждая синтагма, в свою очередь, состоит из одной или более акцентных единиц. Акцентная единица (АЕ) является минимальной просодической единицей и состоит из одного или более слов, имеющих лишь один полноударный гласный. Интонационно значимыми элементами АЕ являются ядро (полноударный гласный), предъядро (все фонемы, предшествующие полноударному гласному) и заядро (все фонемы, следующие за полноударным гласным).

Основное предположение ПАЕ-модели в том, что топологические свойства просодических параметров не зависят от конкретного фонетического контекста и количества слогов в предъядре и заядре для конкретного типа интонации. Таким образом, просодические характеристики могут задаваться «портретами» акцентных единиц, которые указывают нормированные значения F_0, A, и T на участках предъядра, ядра и заядра.

Полный набор таких «портретов», содержащий интонационные характеристики для разных типов синтагм, составляет просодический стиль. БД просодических параметров, используемая на данном этапе, может содержать несколько различных просодических стилей.

В блок вычисления целевых просодических параметров информация подаётся по синтагмам. На первом этапе определяется просодический тип синтагмы и количество АЕ в ней, после чего из БД просодических параметров извлекается соответствующий просодический «портрет». Затем в каждой АЕ выделяются аллофоны, составляющие предъядро, ядро и заядро.

Для каждого аллофона на основе ритмического «портрета», а также на основе положения аллофона в предъядре, ядре или заядре вычисляется коэффициент изменения длительности аллофона (в процентах) k_a. Затем вычисляется целевое значение длительности каждого i-того аллофона T'_{ai}:

$$T'_{ai} = \frac{T_{ai} \cdot k_a}{100}, \qquad (1)$$

где T_{ai} — исходная длительность аллофона.

Целевые интонационные значения вычисляются только для вокализованных аллофонов, при этом интонационные характеристики вычисляются (в отличие от ритмических характеристик) не для всего аллофона, а для каждого периода основного тона аллофона. На основе интонационного «портрета», а также на основе положения аллофона в предъядре, ядре или заядре вычисляются нормализованные целевые значения F_0. Затем с учётом диапазона

частоты основного тона используемой речевой БД вычисляются целевые значения длительностей периодов:

$$T'_{0i} = \frac{f_{дискр} \cdot 100}{F_{0norm\,i} \cdot (F_{0\max} - F_{0\min}) + F_{0\min} \cdot 100}, \qquad (2)$$

где T'_{ai} — целевое значение i-ого периода основного тона (количество отсчётов сигнала);

$f_{дискр}$ — частота дискретизации сигнала;

$F_{0norm\,i}$ — нормализованное (в диапазоне [0..100]) значение частоты основного тона i-ого периода;

F_{0max}, F_{0min} — максимальное и минимальное значение частоты основного тона для речевой базы.

Полученные целевые значения передаются в блок модификации просодических параметров в речевом сигнале.

2.2. Блок модификации просодических параметров в речевом сигнале

Модификация просодических параметров в речевом сигнале осуществляется с использованием метода «плавной сшивки» периодов основного тона [6]. Основное достоинство данного метода — неизменность речевого сигнала на участке периода основного тона, который соответствует моменту схлопывания голосовых связок, что позволяет сохранить индивидуальные тембральные характеристики обрабатываемого голоса. Несомненное достоинство алгоритма «плавной сшивки», важное при его реализации на мобильных телефонах, — линейная вычислительная сложность.

Процесс уменьшения периода показан на *рис. 4* и *5*. Удаляется часть периода длиной N, где

$$N = T_o - T'_o , \qquad (3)$$

где T_o — текущая длина i-того периода;

T'_o — целевая длина периода основного тона.

Удаляемая часть смещается и накладывается на предшествующую часть периода (*рис.4*). Накладывание двух участков сигнала происходит путём плавного уменьшения первого сигнала и увеличения второго сигнала (*рис. 5*).

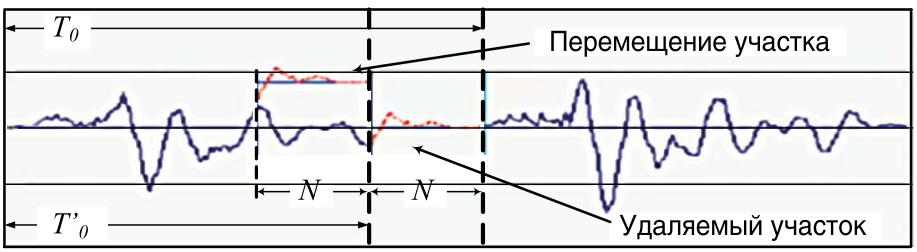


Рис. 4. Перемещение удаляемого участка сигнала

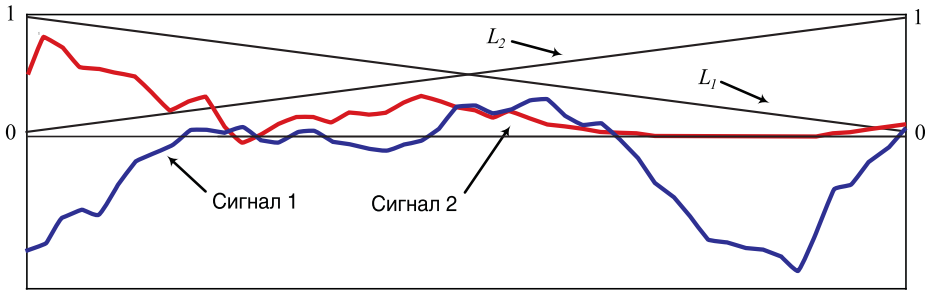


Рис. 5. Формирование переходного участка путём «плавной сшивки» двух сигналов

Модификация сигнала при уменьшении длительности периода основного тона осуществляется в соответствии с формулой:

$$\tilde{s}(n) = \frac{(N-n)\cdot s(n) + n\cdot s(n+N)}{N}, \quad (T_0^{'} - N) \le n \le T_0^{'}, \qquad (4)$$

где $\tilde{s}(n)$ — результирующий речевой сигнал; $s(n)$ — исходный сигнал.

Аналогичная процедура осуществляется при увеличении периода основного тона [6]. При этом результирующий речевой сигнал $\tilde{s}(n)$ вычисляется в соответствии с формулой:

$$\tilde{s}(n) = \frac{(T_0^{'} - n)\cdot s(n) + n\cdot s(n-N)}{T_0^{'}}, \quad N \le n \le T_0^{'} \qquad (5)$$

3. Программная реализация системы на мобильном телефоне

Блок обработки речевого сигнала реализован на языке Java Mobile Edition [9] для минимальной конфигурации CLDC 1.0 [10] и профиля MIDP 2.0 [11], что позволяет использовать его практически на любом современном мобильном телефоне. В следующих разделах описывается пользовательский интерфейс созданной системы и особенности её программной реализации.

3.1. Пользовательский интерфейс системы

Главное меню системы (рис. 6а) включает выбор текстового файла для воспроизведения, просмотр/изменение настроек, непосредственно воспроизведение и справочную информацию, которая содержится в элементе меню «О программе». Настройки включают выбор голосовой базы и выбор просодического стиля для синтеза речи (рис. 6 б). Для синтеза речи пользователь должен сначала указать текстовый файл, содержащий размеченный текст, затем выбрать элемент меню «воспроизведение». При воспроизведении (рис. 6в) в системе реализованы функции паузы/возобновления, а также остановки воспроизведения.

а) б) в)

Рис. 6. Интерфейс системы: а) главное меню; б) выбор голосовой БД и просодического стиля для синтеза речи; в) воспроизведение речевого сигнала

3.2. Особенности программной реализации системы

Звуковые волны аллофонов, содержащиеся в речевой БД, хранятся в формате WAVE PCM. В процессе генерации речевого сигнала они извлекаются из БД, модифицируются в соответствии с описанными выше алгоритмами и помещаются в буфер для воспроизведения. После того, как очередная речевая синтагма подготовлена, она воспроизводится с использованием стандартного класса J2ME Player.

Поскольку процесс генерации речевого сигнала должен происходить практически одновременно с процессом воспроизведения синтезированной речи, в системе реализована многопоточность. При этом главный поток управляет действиями двух дочерних, один из которых генерирует очередную речевую синтагму, другой — воспроизводит. Первый из потоков характеризуется высокой трудоёмкостью выполнения и поэтому имеет больший приоритет, чем второй поток. В то же время выходные данные первого потока являются входными данными второго, поэтому потоки синхронизированы с тем, чтобы работа второго потока всегда начиналась после завершения первого. Такая синхронизация осуществляется главным потоком.

4. Заключение

Разработанная система была успешно протестирована на мобильных телефонах Motorola, Sony-Ericsson, LG, которые характеризуются тактовой частотой ARM-процессора от 68 до 115 МГц, объёмом памяти от 3 500 до 4 200 КБ, поддержкой конфигурации CLDC 1.0 и профиля MIDP 2.0. Система позволяет синтезировать речевой сигнал в реальном времени на мобильных телефонах с ARM-процессорами седьмого поколения.

Экспертная оценка качества синтезированной речи показала, что оно не уступает качеству синтезированной речи, получаемому на персональных компьютерах с использованием тех же методов обработки текстовой и речевой информации.

Созданная система универсальна в том смысле, что замена используемой голосовой базы (например, мужской на женскую) не требует дополнительной обработки входного текста. Созданная система может быть модернизирована с целью озвучивания входящих SMS-сообщений и электронных текстов, полученных через Сеть Internet.

Литература

1. SMS2Voice. Сервис голосовых сообщений [Электронный ресурс]. Электронные данные. Режим доступа: http://voice.s-soft.org. Дата доступа: 10.12.09.
2. Acapela TTS for Windows Mobile [Электронный ресурс]. Электронные данные. Режим доступа: http://www.acapela-group.com/acapela-tts-for-windows-mobile-2-2-speech-solutions-tts.html. Дата доступа: 01.06.10.
3. Nuance TALKS [Электронный ресурс]. Электронные данные. Режим доступа: http://www.nuance.com/talks/. Дата доступа: 10.12.09.
4. Mobile Speak [Электронный ресурс]. Электронные данные. Режим доступа: http://www.codefactory.es/en/products.asp?id=316. Дата доступа: 10.12.09.
5. Gartner Says Worldwide Mobile Phone Sales Grew 17 Per Cent in First Quarter 2010. Press Release. Электронный ресурс. Режим доступа: http://www.gartner.com/it/page.jsp?id=1372013. Дата доступа: 01.07.10.
6. *Лобанов Б.М., Цирульник Л.И.* Компьютерный синтез и клонирование речи. Мн.: Белорусская наука, 2008.
7. *Жадинец Д.В., Сизонов О.Г., Цирульник Л.И.* Электронные словари русского и белорусского языков для двуязычной системы синтеза речи по тексту // Танаевские чтения: Доклады межд. конф., Минск, 28 марта 2007 г. М.: Объединённый институт проблем информатики, 2007. С. 65–69.
8. *Lobanov B., Karnevskaya E.* Auditory Estimation of Effectiveness of the AUP-Stylization Model of the Melodic Contour TTS-synthesis and Voice Cloning. Proc. 13-th Int. Conf. SPECOM'2009, June 21–25, 2009, St.-Pet. P. 130–135.
9. Java ME at a Glance [Электронный ресурс]. Электронные данные. Режим доступа: http://www.oracle.com/technetwork/java/javame/overview/index.html. Дата доступа: 1.08.10.
10. Connected Limited Device Configuration (CLDC); JSR 30, JSR 139 Overview [Электронный ресурс]. Электронные данные. Режим доступа: http://www.oracle.com/technetwork/java/overview-142076.html. Дата доступа: 1.08.10.
11. Mobile Information Device Profile (MIDP); JSR 37, JSR 118 Overview [Электронный ресурс]. Электронные данные. Режим доступа: http://www.oracle.com/technetwork/java/overview-140208.html. Дата доступа: 1.08.10.

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 50 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. E-mail: liliya.tsirulnik@gmail.com

Покладок Дмитрий Александрович —

окончил факультет компьютерного проектирования Белорусского государственного университета информатики и радиоэлектроники. Магистр физико-математических наук. Аспирант, младший научный сотрудник лаборатории распознавания и синтеза речи Объединённого института проблем информатики Национальной академии наук Беларуси. Область научных интересов — системы синтеза речи по тексту для мобильных устройств. E-mail: dima.pokladok@gmail.com

Система синтеза белорусской речи по тексту

Ю.С. Гецевич,
аспирант

Б.М. Лобанов,
доктор технических наук

Рассмотрены особенности фонетических систем белорусского и русского языков и структура белорусскоязычного синтезатора речи по тексту. Описаны структура и состав белорусского электронного словаря, а также программная реализация синтезатора белорусской речи.

Abstract

Peculiarities of the phonetic systems of the Belarusian and Russian languages and a structure of the Belarusian text-to-speech synthesizer are considered. A structure and composition of the Belarusian electronic dictionary, as well as the programme realization of a Belarusian speech synthesizer of are described.

Основные методы верификации речевых данных

Для некоторых славянских языков, таких как русский, чешский, польский, украинский, уже существуют практически используемые или экспериментальные образцы синтезаторов речи по тексту (СРТ) [1]. В литературе нет, однако, никаких сведений о создании синтезаторов речи по тексту для белорусского языка. Работа является продолжением проводимых ранее исследований, базирующихся на разработке русскоязычного синтезатора речи [2]. Результатом этой работы стало создание двуязычной системы синтеза речи по тексту. Синтез речи по тексту на двух славянских языках — белорусском и русском — предполагает создание фонетико-акустической базы данных, построенной на единых принципах, отражающих внутри- и межъязыковую специфику фонетических систем и позиционно-комбинаторных явлений, порождающих аллофонию речи.

1. Особенности фонетических систем белорусского и русского языков

Фонетические системы языков, относящихся к группе славянских, имеют между собой значительное сходство, однако каждый из них обладает также специфическими

особенностями, иногда значительными. Исследуемые фонетические системы белорусского и русского языков относительно близки. В белорусском языке насчитывается 41 фонема, из них 6 гласных и 35 согласных, а в русском — 42 фонемы: гласных — 6 и согласных — 36. В *табл. 1* представлена обобщённая информация о фонемном составе двух языков и об их различии по способу и месту образования. В каждой ячейке таблицы представлены имена фонем, характеризующихся определённым способом и местом образования, для белорусского и русского языков порядке «сверху — вниз». Для обозначения фонем используются традиционные для каждого языка буквы алфавита.

Таблица 1

Фонетические системы белорусского и русского языков

Место образования \ Способ образования		Согласные: Глухие: Взрывные	Глухие: Аффрикаты	Глухие: Щелевые	Звонкие: Взрывные	Звонкие: Аффрикаты	Звонкие: Щелевые	Сонорные: Дрожащие	Сонорные: Носовые	Сонорные: Боковые	Сонорные: Плавные	Гласные	Передняя	Высокая	Огубленная
Задне-язычные	Мягкие	к' к'		х' х'	~ г'	~	гх' ~				й й	у у	0	1	1
	Твёрдые	к к		х х	~ г		гх ~					о о	0	0	1
Средне-язычные	Мягкие		~ ч'	~ ш'				~ р'				а а	0	0	0
	Твёрдые		ч ~	ш ш		д ж ~	ж ж	р р				э э	1	0	0
Передне-язычные	Мягкие	~ т'	ц' ~	с' с'	~ д'	дз' ~	з' з'		н' н'	л' л'		ы ы	0	1	0
	Твёрдые	т т	ц ц	с с	д д		з з		н н	л л		і и	1	1	0
Губные	Мягкие	п' п'		ф' ф'	б' б'		в' в'		м' м'						
	Твёрдые	п п		ф ф	б б		в в		м м		ў ~				

Таблица 1 затемнены ячейки, фонетическое качество звуков которых имеет практически полное сходство для каждого из языков. Как видно из таблицы, количество таких ячеек в процентном отношении ко всем использующимся ячейкам довольно значительно — 71%. Отличительные особенности фонетических систем белорусского и русского языков заключаются в следующем.

В белорусском языке отсутствуют следующие фонемы:

- мягкие согласные ***Т, Д, Ш, Ч, Р***;
- мягкая и твёрдая ***Г****

В белорусском языке есть ряд специфических фонем, отсутствующих в русском:

- плавная ***Ў***;
- мягкая ***Ц*** и твёрдая ***Ч***;
- мягкая аффриката ***Дз*** и твёрдая ***Дж***;
- мягкая и твёрдая щелевая ***Гх***.

* произношение взрывного звука Г возможно в некоторых заимствованных словах (гандаль, ганак, гузік) и в коренных буквосочетаниях ЗГ, ДЗГ, ДЖГ (мазгі, бразгаць, джгаць), в остальных случаях произносится звонкий фрикативный звук «Гх».

2. Структура белорусскоязычного синтезатора речи по тексту

Синтез устной речи по тексту осуществляется на основе лексико-грамматического анализа входного текста путём моделирования процессов речеобразования с учётом правил произношения звуков и интонирования, свойственных белорусскому языку. Орфографический текст документа (книги, статьи, веб-страницы и т.п.) поступает на вход синтезатора и далее подвергается последовательной обработке рядом специализированных процессоров в соответствии с общей структурой синтезатора речи по тексту, представленной на *рис. 1*. Синтезатор включает четыре основных модуля: текстовый процессор, просодический процессор, фонетический процессор и акустический процессор. Каждый из этих модулей поддерживается наборами соответствующих БД и правил.

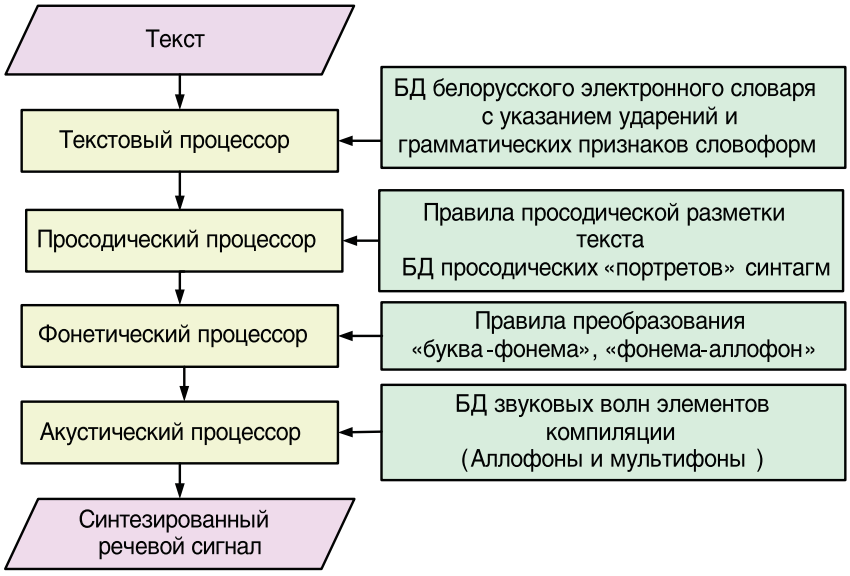


Рис. 1. Общая структура синтезатора речи по тексту

Входной орфографический текст подвергается ряду последовательных обработок в каждом из процессоров. Текстовый процессор обрабатывает входной орфографический текст в следующей последовательности: очистка текста, преобразование символов (аббревиатур, сокращений, чисел и др.), расстановка словесных ударений и грамматических признаков словоформ. Преобразованный текст поступает на входы просодического, а затем фонетического процессора. В результате работы просодического процессора, текст делится на синтагмы, акцентные единицы (АЕ), который далее размечается на элементы акцентных единиц (ЭАЕ): интонационное предъядро, ядро и заядро.

И наконец, последняя функция просодического процессора — установка в соответствии с БД просодических «портретов» синтагм значений амплитуды (А), длительности фонем (Т) и частоты основного тона (F0) для каждого ЭАЕ. Задача фонетического процессора заключается в преобразовании орфографического текста в

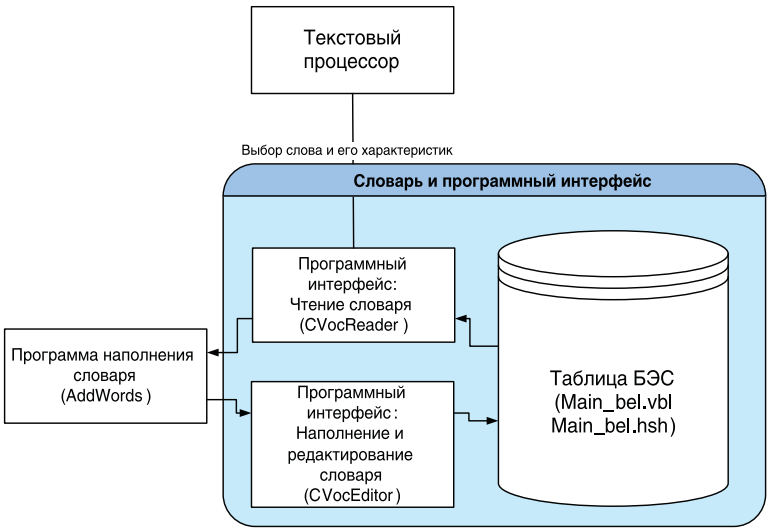


Рис. 2. Структурная схема белорусского электронного словаря

фонемный, а также в генерации позиционных и комбинаторных аллофонов. Акустический процессор на основе информации о том, какие аллофоны необходимо синтезировать, а также какие просодические характеристики должны быть приписаны каждому аллофону, генерирует речевой сигнал путём компиляции отрезков естественных звуковых волн соответствующих аллофонов и мультифонов (диаллофоны и аллослоги).

Текстовый, просодический, фонетический и акустический процессоры синтезатора речи в своей основе языконезависимые, а языковая специфика (в нашем случае специфика белорусского языка) задаётся соответствующим набором баз данных и знаний (правил). На *рис. 1* языконезависимые блоки представлены прямоугольниками в левой части, а языкозависимые — в правой части. Три из четырёх блоков синтезатора (см. *рис. 1*) — просодический, фонетический и акустический процессоры — в наименьшей степени требуют коррекции языковых баз данных и знаний при переходе от синтеза русской речи к белорусской. Коренные изменения связаны, в основном, с созданием электронного белорусского словаря.

Таблица 2

Сравнительные характеристики словарей белорусского и русского языков

Часть речи	*Количество словоформ (белорусский)*	*Количество словоформ (русский)*
Существительное	591 036	465 230
Глагол	624 501	989 064
Прилагательное	874 997	626 842
Наречие	6 541	1 353
Числительное	571	410
Предлог	121	87
Междометие	96	180
Частица	58	86
Союз	46	68
Всего	2 097 967	2 083 320

3. Белорусский электронный словарь

Полная словарная база содержит 2 097 967 словоформ белорусской речи. В *табл. 2* приведены сравнительные характеристики состава используемых словарей для белорусского и русского языков. Словарь белорусского языка построен на базе словаря М.В. Бирылы [3], русского языка — А.А. Зализняка [4].

Архитектура белорусского электронного словаря (БЭС) представляет собой базу данных с конкретными функциями для её чтения и наполнения (*рис. 2*).

Таблица базы данных белорусского электронного словаря наполняется записями с тремя полями: орфографическое слово S1, позиции ударений в слове S2, теги слова S3 (*табл. 3*). На диске компьютера таблица БЭС сохраняется в файлах main_bel.vbl (данные записей) и main_bel.hsh (данные проиндексированных записей).

Для чтения БЭС используется программный интерфейс класса CVocReader. Его архитектура приводиться в *табл. 4*.

Таблица 3

Архитектура базы данных БЭС

Название поля	*Тип*
Орфографическое слово	Строка (<=255 символов)
Позиция ударения	Вектор целых чисел
Тег	Строка (<=255 символов)

Таблица 4

Архитектура программного класса CVocReader для чтения словаря

Функция класса CVocReader	*Комментарий*
Открытие/Закрытие	Словарь переводится в режим чтения/общего доступа
Получить количество слов в словаре	Реальное количество слов
Поиск слова	Выбирается запись (3 поля) из таблицы БД
Поиск следующего слова	Выбираются остальные записи. (В словаре могут быть слова омографы, поэтому эта функция покажет все хранимые варианты слов)
Получить последнюю ошибку	Во время чтения словаря эта функция постоянно должна проверяться для следующей корректной работы всего процесса синтезирования речи

Для редактирования БЭС используется программный интерфейс класса CVocEditor. Его архитектура приводится в таблице 5.

Таблица 5

Архитектура программного класса CVocEditor для редактирования словаря

Функция класса CVocEditor	*Комментарий*
Открытие/Закрытие	Словарь переводится в режим редактирования/общего доступа
Добавить слово	Добавляется запись с ЛГК (3 поля) в таблицу БД

Удалить слово	Удаляется запись с ЛГК (3 поля) из таблицы БД
Поиск слова	По орфографической записи слова ищется соответствующая запись с ЛГК в БЭС
Поиск слова по маске	По условной записи (через регулярные записи) слова ищется соответствующая запись с ЛГК в БЭС
Получить последнюю ошибку	Во время чтения словаря эта функция постоянно должна проверяться для следующей корректной работы всего процесса синтезирования речи
Обновить/Перестроить индекс БЭС	Синхронизация индексов для поиска слов в БЭС. (Должна исполняться после любых редактирований таблицы)

В *табл. 6* представлен фрагмент используемой словарной базы.

Таблица 6

Список словарной базы с указанием ударения (`) и граматических категорий (Тэг). (фрагмент, всего — 1 260 794)

Слова_Тэг (существительные)	*Слова_Тэг (глаголы)*
…	…
зака`зчык_ NNAMO	падвыша`ць_ VIC
зака`зчыка_ NNAMG	падвыша`ю_ VIIR1
зака`зчыку_ NNAMD	падвыша`еш_ VIIR2
зака`зчыка_ NNAMA	падвыша`е_ VIIR3
зака`зчыкам_ NNAMI	падвыша`ем_ VIIR1P
зака`зчыку_ NNAMR	падвыша`еце_ VIIR2P
зака`зчыкі_ NNAMPO	падвыша`юць_ VIIR3P
зака`зчыкаў_ NNAMPG	падвыша`й_ VIM2
зака`зчыкам_ NNAMPD	падвыша`йце_ VIM2P
зака`зчыкаў_ NNAMPA	падвыша`ў_ VIIPM
зака`зчыкамі_ NNAMPI	падвыша`ла_ VIIPF
зака`зчыках_ NNAMPR	падвыша`ла_ VIIPN

Отдельная группа алгоритмов разработана для преобразования чисел в порядковые и количественные числительные.

В т*абл. 7–9* приведены правила преобразования чисел в числительные с соответствующими базами данных, а в таблице 10 — соответствующие текстовые примеры.

Таблица 7

Список соответствий — Tls «окончание =>склонение порядкового числительного» (фрагмент, всего — 480).

Число	*Окончание*	*Окончание*
…	…	…
12	-га	дванаццатага
12	-е	дванаццатае
12	-ы	дванаццаты
12	-м	дванаццатым
12	-мі	дванаццатымі
12	-му	дванаццатаму
12	-ае	дванаццатае
12	-ай	дванаццатай
12	-аю	дванаццатаю
12	-ую	дванаццатую
12	-х	дванаццатых
12	-ыя	дванаццатыя
12	-ым	дванаццатым
…	…	…

Таблица 8

Список Tls — «число, склонение =>количественное числительное» (фрагмент, всего 388).

Лік	*Склонение*	*Числительное*
…	…	…
3	Т	Трыма
30	Т	Трыццаць
300	Т	Трыстамі
4	Т	Чатырма
40	Т	Сарака
400	Т	Чатырстамі
5	Т	Пяццю
50	Т	Пяццюдзесяццю
500	Т	Пяццюстамі
6	Т	Шасцю
60	Т	Шасцюдзесяццю
600	Т	Шасцюстамі
8	Т	Васьмю
80	Т	Васьмюдзесяццю
800	Т	Васьмюстамі
9	Т	Дзевяццю
90	Т	Дзевяноста
…	…	…

Таблица 9

Список Ttl: «число=>название триады»

Число	*Название триады*
1	Тысяча
2…4	Тысячы
5..20, 30, 40, … 90	Тысяч
100, 200, … 900	Тысяч

4. Программная реализация синтезатора белорусской речи

Модульная структура программной реализации синтезатора белорусской речи представлена на рис. 3, где модули, управляющие последовательностью действий других модулей,

называются контроллерами, в то время как модули, реализующие алгоритмы обработки текста или речевого сигнала, называются процессорами.

Главный контроллер системы управляет последовательностью преобразований входных данных, получая на каждом этапе промежуточный результат и передавая его на обработку на следующий этап. Контроллер нормализации текста удаляет из текста символы, не нужные для синтеза речи, убирает случайное дублирование знаков препинания, заменяет похожие символы на один из них, очищает входной текст от недопустимых символов, используя для этого списки букв языка, знаков препинания, а также правила замены символов.

Таблица 10

Примеры преобразования чисел в числительные

Входной текст	*Выходной текст*
Количественные числительные	
Настаўнік атрымаў заробак 784921 рубель.	Настаўнік атрымаў заробак семсот восемдзесят чатыры тысячы дзевяцьсот дваццаць адзін рубель.
Кіраўніцтва зацвердзіла аб'ёмы асноўных відаў прадукцыі для 5 галінаў гаспадаркі.	Кіраўніцтва зацвердзіла аб'ёмы асноўных відаў прадукцыі для пяці галінаў гаспадаркі.
Порядковые числительные	
Завод выпусціў 234000-ы аўтамабіль.	Завод выпусціў двухсоттрыццацічатырохтысячны аўтамабіль.
У нас няма 123-га байца.	У нас няма сто дваццаць трэцяга байца.
У 2010-ым годзе адбудзецца алімпіяда па…	У дзве тысячы дзесятым годзе адбудзецца алімпіяда па …
198000-ая скрынка з цукеркамі выйшла з вытворчага цэха.	Стодзевяноставасьмітысячная скрынка з цукеркамі выйшла з вытворчага цэха.

«Очищенный» таким образом текст подаётся на вход лингвистического контроллера, выходом которого является просодически размеченный текст. Преобразованный текст поступает на вход фонетического процессора, который осуществляет преобразования «буква-фонема», «фонема — аллофон», затем — на вход просодического процессора, который устанавливает текущие значения амплитуды, частоты основного тона и длительности каждого аллофона. Просодически размеченный аллофонный текст поступает затем на вход акустического процессора, который генерирует речевой сигнал с использованием БД звуковых волн аллофонов и мультифонов. Результат преобразований поступает в контроллер формата выходных данных, который осуществляет преобразование в нужный формат: wav или mp3.

На *рис. 4* представлен программный интерфейс отладочной версии синтезатора белорусской речи.

В нижней части представлен орфографический текст, подаваемый на вход синтезатора. В верхней части первый столбец отображает порядковый номер просодических синтагм этого текста. Во втором столбце указан интонационный тип каждой синтагмы (Q — разновидности вопросительной интонации, P — повествовательной). В третьем столбце указано количество акцентных единиц в каждой из синтагм. В четвёртом столбце приведён фонемный текст каждой синтагмы, в которой знаком «ъ» отражен факт объединения служебных

Рис. 3. Программная реализация синтезатора белорусской речи

и знаменательных слов в одно фонетическое слово, знаки «+» и «=»указывают, соответственно, положение в словах полного и частичного ударения, а знак «/» — положение границ акцентных единиц синтагмы. В последнем, пятом, столбце приведён аллофонный текст синтагм.

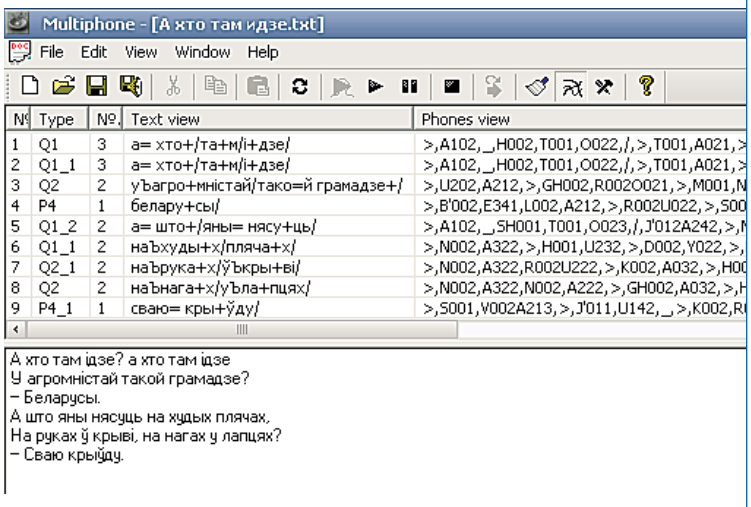


Рис. 4. Программный интерфейс синтезатора белорусской речи

Заключение

Специальное подразделение ЮНЕСКО, занимающееся сохранением языков мира как живым представителем культурного наследия планеты, отнесло белорусский язык к вымирающим — «потенциально угрожаемым» языкам [6]. Если современная тенденция кардинально не изменится, в ближайшие десятилетия мы можем потерять еще один живой язык Земли — славянский белорусский язык. В свете этого факта значимость создания белорусскоязычного синтезатора речи трудно переоценить. В нём программным образом аккумулированы и сохранены для будущих поколений лексика белорусского языка — арфография, акцентуация, словоизменение —

фонетика и правила чтения, интонация и звуки современной белорусской речи. В то же время современным пользователям компьютеров и Интернета этот программный продукт сможет помочь в овладении и совершенствовании знания и навыков белорусской речи, а также использовать его в различных прикладных системах.

В заключение авторы выражают глубокую благодарность заведующему лабораторией интеллектуальных систем БГУ профессору И.В. Совпелю и его сотрудникам за передачу базовой версии белорусского электронного словаря и помощь в его адаптации для решения задач синтеза белорусской речи.

Работа выполнена при поддержке гранта № Ф10Р-006 БРФФИ.

Литература

1. http://www.speech.cs.cmu.edu/comp.speech/
2. Лобанов Б.М. Компьютерный синтез и клонирование речи // Б.М. Лобанов, Л.И. Цирульник / Минск: Белорусская Наука, 2008.
3. Зализняк А.А. Грамматический словарь русского языка: Словоизменение. Ок 100 000 слов. 2-е изд., стереотип. М.: Рус. Яз., 1980.
4. Слоўнік беларускай мовы: Арфаграфія. Арфаэпія. Акцэнтуацыя. Словазмяненне / Ін-т мовазнаўства імя Я. Коласа АН БССР; Пад рэд. М.В. Бірылы. Мн.: БелСЭ, 1987..
5. http://ru.wikipedia.org/wiki — Белорусский язык.

Гецевич Ю.С. —

окончил факультет прикладной математики и информатики Белорусского государственного университета, факультет математики и информатики университета в Манхейме (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.
E-mail: mix1122@gmail.com

Лобанов Б.М. —

доктор технических наук. 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.

Синтез пения для русского языка

Л.И. Цирульник,
кандидат технических наук
А.С. Ломов,
магистрат

Работа посвящена описанию реализации синтеза пения для русского языка. Приводится информация о певческом голосе, показаны основные тембральные и просодические особенности певческих голосов. Приведена общая структурная схема системы синтеза пения, описаны алгоритмы и принципы работы блоков системы, а именно, блока обработки музыкальной нотации, блока фонетических преобразований, блока синтеза речевой волны. Описанные алгоритмы языконезависимы и могут применяться для синтеза пения на других языках.

Abstract

The paper describes the implementation of the singing synthesis software system for Russian language. The information about singing voice is outlined, and the general timbral and prosodic characteristics of singing voices are shown. The paper presents the achitecture of the singing synthesis system and descibes the algorithms and principles of operation of the system components such as the music notation processing, speech phonemic processing, and speech wave synthesis units. The given algorithms are language independent and could be applied for creating singing synthesis systems for other languages.

Введение

Система синтеза пения может использоваться при обучении вокалу, для демонстрации правильного исполнения песни или развития музыкального слуха. Такой компьютерный инструментарий будет полезен композиторам и продюсерам для создания демонстрационных версий песен, добавления в уже имеющиеся записи бэк-вокала и получения других эффектов. Эта система может найти широкое применение в качестве средства для генерации заставок на радио, интерактивной рекламы, звуковых дорожек к различным видеоматериалам.

Идея цифрового синтеза качественного певческого голоса начала привлекать внимание исследователей с 50-х годов прошлого века. Первый синтезированный певческий голос — синтез песни «Daisy Bell» — был создан американским учёным Максом Мэтьюзом [2], который разработал технологию синтеза вокала на основе вокодера. Первой полностью автоматической компьютерной системой, осуществляющей синтез пения, стала программа VocalWriter от компании KAE Labs [3], выпущенная в 1998 году для операционной системы MacOS. К настоящему моменту существуют компьютерные системы, осуществляющие синтез пения на японском языке: программа Vocaloid компании Yamaha [4], на французском, португальском, итальянском языках: программа компании Myriad [5], на немецком языке: программа Virsyn Cantor [6] и вышеназванная программа компании Myriad. Кроме того, все перечисленные программы осуществляют синтез пения на английском языке.

Для русского языка до сих пор не существует профессиональных программных продуктов, осуществляющих синтез пения. Созданные к настоящему моменту системы, одна из которых описана в работе [7], имеют ряд недостатков, в частности, они не реализуют особые правила преобразования «буква-фонема» на стыках слов, не используют при синтезе речевые отрезки длительностью более одного аллофона, а также не работают с наиболее распространёнными форматами записи музыкальной нотации. Эти недостатки влекут сильное снижение качества синтезированного певческого голоса и требуют предварительных преобразований существующих музыкальных нотаций в формат текста и MIDI-файла. В данной работе описана система синтеза пения для русского языка, лишённая указанных недостатков и позволяющая синтезировать высококачественный певческий голос.

1. Общая информация о певческом голосе

Существует множество систем классификации певческих голосов. Одни учитывают силу голоса, другие — насколько подвижен, виртуозен, отчётлив голос певца. Чаще всего используется классификация, учитывающая диапазон голоса певца [1]. Под вокальным диапазоном обычно понимают набор музыкально полезных звуков, которые доступны певцу. «Полезными» называют те звуки, которым певец может придать необходимую длительность, силу и окраску. Как показано в таблице 1, частотный диапазон певческого голоса составляет 80–1050 Гц [1], что в интервальном исчислении составляет четыре октавы. Каждый певческий голос занимает две и более октавы, в то время как диапазон изменения частоты основного тона (ЧОТ) при устной речи, как правило, не превышает одной октавы.

Таблица 1

Классификация певческих голосов по диапазону

Название группы голосов	*Частотный диапазон, Гц*
Бас	80–330
Баритон	110–440
Тенор	130–520
Контральто	165–700
Меццо-сопрано	220–880
Сопрано	260–1050

Другая характеристика голоса — тембр. Подвижный тип резонаторов голосового тракта обеспечивает возможность изменения тембра в процессе пения или речи и, наряду с изменением высоты и силы голоса, используется для выражения эмоций певцом, лектором, драматическим актёром.

Для того чтобы синтезировать голос с хорошими вокальными данными, нужно выделить отличия профессионального пения от любительского. Наиболее заметное отличие проявляется в более чётком выделении первой, второй и третьей форманты у профессиональных певцов. Кроме того, обученные певцы создают резонанс после 3000 Гц [8]. Эти явления продемонстрированы на *рис. 1* на примере партии голоса эстрадной песни «Красная смородина» двух учениц музыкальной студии, одна из которых имеет хорошие вокальные данные и пятилетний опыт музыкальных занятий, другая только начала обучение вокалу. Рисунок демонстрирует, насколько профессиональное пение обогащено дополнительными обертонами выше границы в 4 кГц, в какой степени форманты имеют более чёткую структуру.

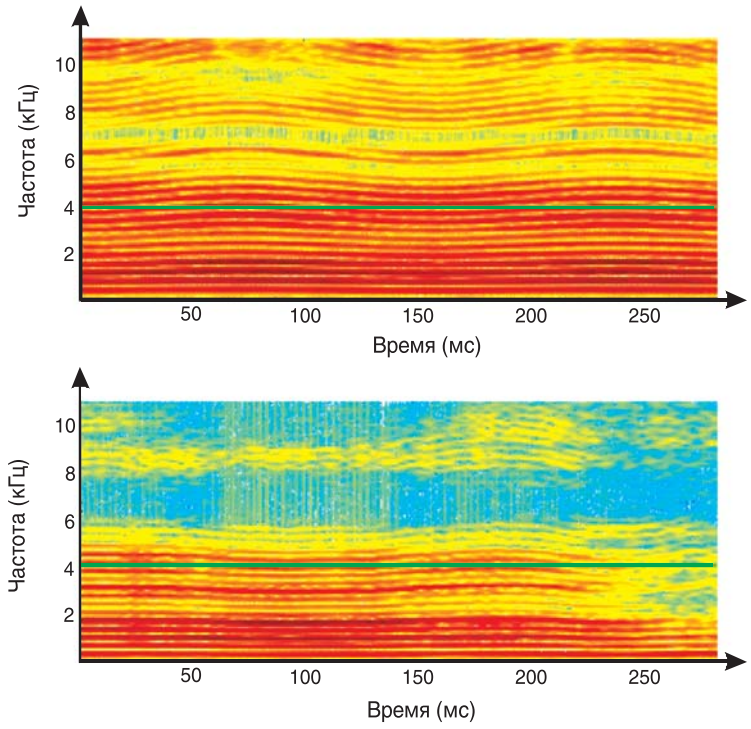


Рис. 1. Спектрограммы исполнения одного и того же песенного фрагмента опытным вокалистом (сверху) и певцом без подготовки (снизу)

Для моделирования певческого голоса с большим уровнем естественности звучания следует остановить внимание на приёмах, которые используются при пении. Один из широко распространённых и часто используемых вокальных приёмов как в академической школе, так и при эстрадном исполнении — вибрато. Вибрато — это периодическое изменение ЧОТ в течение фрагмента речи. Частота изменения ЧОТ обычно 5–8 Гц, а глубина модуляции изменяется в пределах 50–150 центов (под центом в музыке понимается логарифмическая единица измерения относительного изменения частоты, при этом в одной октаве содержится 1200 центов. Две частоты f_1 и f_2 отличаются на 1 цент, если их отношение f_1/f_2 равно $2^{1/1200}$).

Опытные певцы исполняют вибрато с большей частотой и глубиной [8]. Известно, что исполнители баритоном с наиболее приятными голосами поддерживали вибрато в течение 80% времени пения.

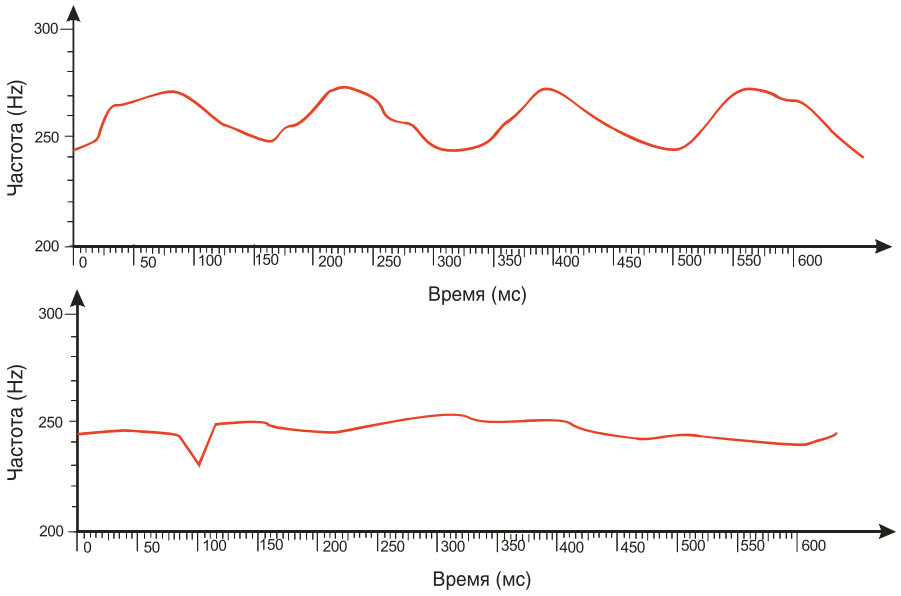


Рис. 2. Графики зависимостей ЧОТ от времени при исполнении вибрато опытным вокалистом (сверху) и певцом без подготовки (снизу)

На *рис. 2* показаны графики изменения ЧОТ при исполнении с помощью вибрато последнего гласного /о/ в слове «домой» певицы с достаточно хорошо поставленным голосом (сверху) и начинающей певицы (снизу). На верхнем графике ЧОТ имеет более выраженные периодические изменения, с большей амплитудой и частотой.

Кроме вибрато, во время пения используются такие приёмы извлечения звука, как пение в грудном регистре и фальцетом. Как известно, в образовании звука главную роль играют поперечные колебания голосовых складок. Именно они в полном объёме имеют место при грудном регистре. Фальцет — это способ формирования высоких звуков, превышающих по частоте естественный грудной регистр [1]. При фальцетном регистре голосовые складки расслабляются, колеблются лишь их края; голосовая щель закрыта не полностью, имеет эллипсоидную форму.

2. Система синтеза пения

Одно из главных отличий пения от устной речи заключается в форме его представления. Музыкальная нотация явно определяет просодические характеристики звуков, в отличие от синтеза речи по тексту, при котором интонацию высказывания нужно определить, для чего используются различные модели и алгоритмы.

Музыкальная нотация имеет множество представлений — от обычно используемых нотных и табулатурных записей до таких необычных нотаций, как невмы [9] и «abc» [10]. Однако общее правило — каждому слогу или звуку сопоставляется последовательность записей, которые определяют высоту тона, длительность и другие параметры звука [11]. Такое представление подаётся на вход системы (*рис. 3*), затем из него выделяется музыкальная нотация и текст песни.

Далее текст поступает на вход фонетического преобразователя, а нотное представление песни переводится в набор целевых (требуемых при синтезе) просодических параметров: частоты основного тона (F_0), амплитуды (A), длительности (T) для каждой ноты в музыкальной нотации.

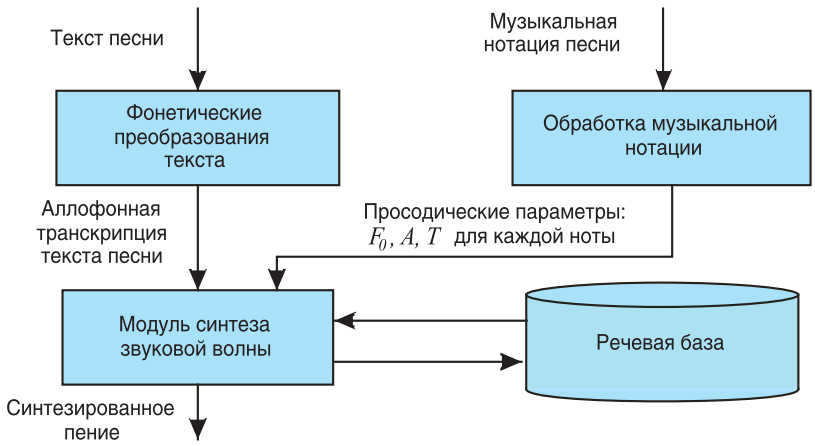


Рис. 3. Общая схема работы системы синтеза пения

Результатом обработки текста фонетическим анализатором становится аллофонная транскрипция слов песни. В модуле синтеза сигнала на основе полученной транскрипции и целевых просодических параметров генерируется звуковой сигнал. При этом модуль использует речевую базу данных (БД), содержание которой определяется методом синтеза речи.

2.1. Обработчик музыкальной нотации

Задача обработки музыкального представления песни заключается в переводе из формата представления музыкальной нотации в целевые значения просодических параметров речи: F0, A, T. Существует множество форматов представления музыкальных произ-ведений в электронном виде, например, такие как gtp [12], MIDI и kar [13], NIFF и SMDL [14]. Однако каждый из них разрабатывался для определённых узких целей, кроме того, большинство из них — коммерческие закрытые форматы. Поэтому в качестве внутреннего формата был выбран MusicXML [14], который является открытым. Этот формат понятен человеку, знакомому с теорией музыки, и редактируется вручную. Формат MusicXML быстро развивается и поддерживается большинством коммерческих и открытых нотных редакторов.

При вычислении целевых просодических параметров на основе нотации в формате MusicXML частота основного тона вычисляется в зависимости от ступени ноты по формуле:

$$F_0 = f_0 \cdot 2^{n\backslash 12}, \qquad (1)$$

где f_0 — частота исходной ступени,

n — количество ступеней от ноты до исходной ступени [11].

Длительность звучания ноты T вычисляется по формуле

$$T = 4 \cdot r \cdot t_0, \qquad (2)$$

где r — относительная длительность текущей ноты (половинная, четвертная, восьмая и т.п.);

t — длительность четвертной ноты в миллисекундах, определяемая темпом произведения [14].

Коэффициент интенсивности вычисляется на основе знаков динамики, присутствующих в нотной записи, например, таких как крещендо, диминуэндо, сфорцандо, меццо-форте, пианиссимо и др. [11].

2.2. Фонетический преобразователь

На вход фонетического обработчика подаётся текст, разделённый на слоги. Внутри процессора он проходит три этапа: расстановку ударений, преобразование «буква-фонема» и преобразование «фонема-аллофон». Выходные данные — последовательность аллофонов, разделённая на слоги.

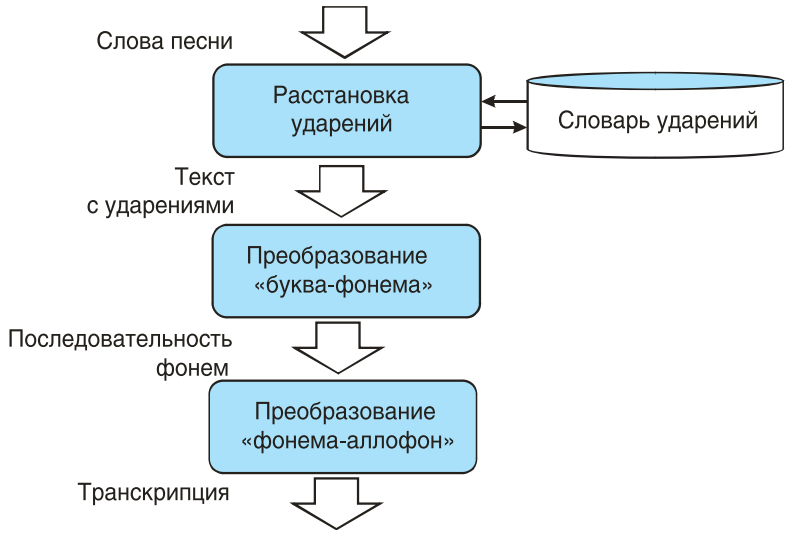


Рис. 4. Схема работы фонетического анализатора

На первом этапе в поступившем на вход тексте расставляются ударения, для чего используется словарь ударений. Затем размеченный текст преобразуется в последовательность фонем с использованием стандартных правил преобразования «буква-фонема» [15]. При преобразовании «фонема-аллофон» генерируются, в отличие от соответствующего преобразования в системе синтеза речи по тексту, аллофоны только полноударных и частично ударных гласных.

2.3. Модуль синтеза речевого сигнала

Несмотря на то, что пение отличается от устной речи, синтез пения имеет много общего с синтезом речи по тексту. Для синтеза речи по тексту используются такие подходы, как артикуляторный, формантный, компиляционный (конкатенативный) и корпусный синтез [16]. В качестве модели для синтеза певческого голоса был выбран компиляционный метод из-за простоты реализации и достаточно хорошего конечного качества.

На вход модуля синтеза речевого сигнала (рис. 5) поступает аллофонная транскрипция текста и набор целевых просодических характеристик: F_0, A, T для каждого аллофона. На первом этапе обработки происходит выбор из речевой БД требуемых речевых сегментов и их конкатенация. При компиляционном синтезе речи БД может содержать не только аллофонные, но и диаллофонные (состоящие из последовательности двух аллофонов) и аллослоговые сегменты, причём использование более длинных сегментов улучшает качество синтезированной речи. В работе [16] показано, что для

достижения наиболее высокого качества синтезированной речи необходимо осуществлять поиск и извлечение из БД диаллофонов в соответствии со следующим приоритетом: ГГ, СГ, СС, ГС (где Г обозначает гласный, С — согласный).

При синтезе пения, однако, поиск и извлечение диаллофонов происходят по другим правилам. Не осуществляется поиск в БД диаллофонов типа ГГ и диаллофонов типа СГ в случае, если согласный — сонорный. Связано это с тем, что в обоих вариантах сложно определить точную границу между двумя звуками. Точное определение границы, однако, очень важно и в первом, и во втором случаях. В первом случае это значимо потому, что две гласные принадлежат к разным слогам и имеют в большинстве случаев разные целевые значения F_0. Во втором случае определение точной границы необходимо потому, что длительность гласных в процессе просодической модификации меняется, в то время как длительность сонорных согласных остаётся неизменной. Как показал опыт разработки системы синтеза пения, искажения, возникающие из-за неточного определения границ двух звуков, заметно ухудшают качество синтезированного пения. Таким образом, из речевой БД осуществляется выбор только следующих типов сегментов: СГ (где С не является сонорным), СС и ГС.

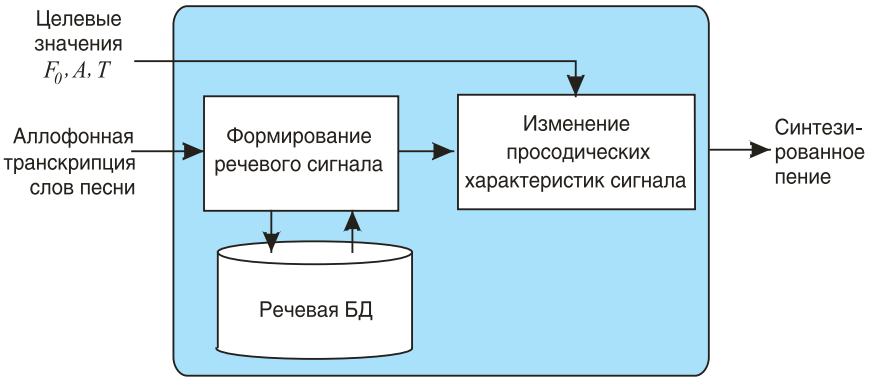


Рис. 5. Схематическое представление модуля синтеза сигнала

Сформированный сигнал подаётся в блок акустической обработки, выполняющий модификацию значений F_0, A, T речевой волны в соответствии с входными значениями просодических параметров. При этом могут использоваться различные алгоритмы модификации сигнала: TD-PSOLA [17], алгоритм плавной сшивки [16], модель «гармоники плюс шум» [18]. В описываемой системе используется алгоритм плавной сшивки, достоинства которого — достаточно хорошее качество модифицированного сигнала, а также линейная вычислительная сложность алгоритма.

Модификация речевой волны при увеличении периода основного тона осуществляется по периодам. Результирующий сигнал одного периода основного тона $\tilde{s}(n)$ вычисляется в соответствии с формулой:

$$\tilde{s}(n) = k(n)s(n) + (1 - k(n))s(n + \Delta T), \quad n = \overline{(1,T)}, \qquad (3)$$

где $\tilde{s}(n)$ — отрезок исходного сигнала длительностью в один период основного тона;

— ΔT — разность между требуемой длительностью периода основного тона T и исходной длительностью периода T_0: $\Delta T = T - T_0$;

$k(n)$ — кусочно-линейная функция, которую можно выразить формулой:

$$k(n) = \begin{cases} 1, & n \le \Delta T; \\ 1 - \dfrac{n - \Delta T}{T_0 - \Delta T}, & \Delta T < n < T_0; \\ 0, & n > T_0; \end{cases} \qquad (4)$$

Ниже приведён пример увеличения длительности одного из периодов основного тона фонемы /a/. В этом случае длительность периода увеличивается с 241 до 361 отсчётов.

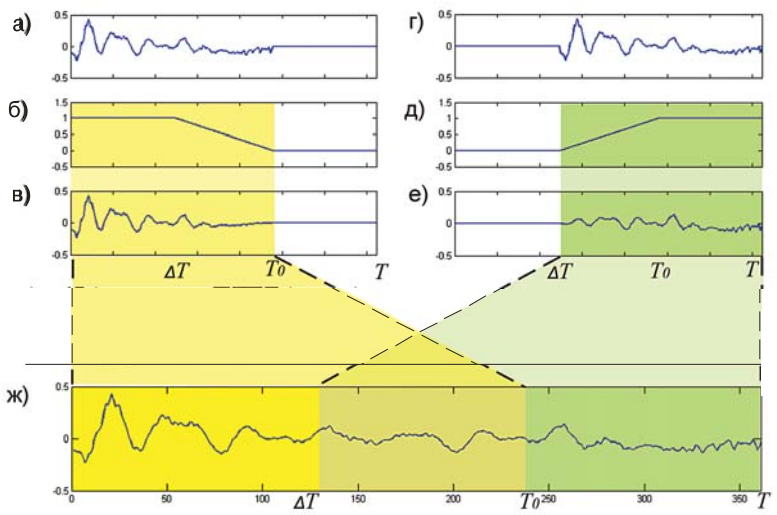


*Рис. 6. Иллюстрация последовательной обработки исходного сигнала s(t) методом «плавной сшивки» при увеличении длительности периода основного тона: а) исходный сигнал s(t); б) кусочно-линейная функция k(t); в) первое слагаемое результирующего сигнала s(t)*k(t); г) сдвинутый сигдал s(t+ΔT); д) кусочно-линейная функция k'(t); е) второе слагаемое результирующего сигнала k'(t)*s(t+ΔT); ж) результирующий сигнал*

При уменьшении длительности периода основного тона лишний участок удаляется и «накладывается» на предшествующий участок по тому же принципу, что и при увеличении длительности.

Алгоритм даёт возможность с хорошим качеством изменять длительность периода основного тона на 50% от длины исходного периода. Изменение ЧОТ при этом находится в интервале от 70% до 200% от исходной частоты.

Изменение длительности в соответствии с целевым значением T происходит только на гласных фонемах. При этом в гласном дублируется или удаляется целое число периодов основного тона. Изменение аллофона начинается с его середины, чтобы сохранить переходные участки между звуками как можно более неизменными.

3. Особенности программной реализации системы

Описанная выше система реализована на языке программирования C++ с использованием инструментария Qt для создания интерфейса. Для работы со звуком выбрана библиотека DirectSound. Программа работает в операционной среде Windows. В качестве входных данных программа использует файлы MusicXML. Результат можно сохранить в файл с расширением wav.

На *рис. 7* приведён пример внешнего вида программы. Информация о словах песни, их транскрипция и осциллограмма синтезированного звука отображаются

друг под другом на разных линейках.

Программа может устанавливаться в виде расширения для редактора музыкальных нотаций MuseScore [19]. В этом случае синтезатор озвучивает составленную в редакторе нотную запись.

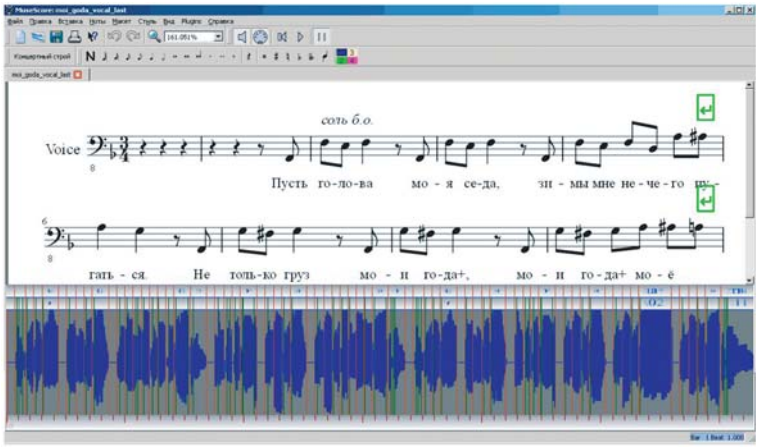


Рис. 7. Внешний вид окна программы синтеза пения

В системе используется речевая БД мужского голоса, содержащая 3000 речевых отрезков. Среднее значение ЧОТ вокализованных элементов БД — 100 Гц. Таким образом, в соответствии с используемым алгоритмом изменения ЧОТ — алгоритмом плавной сшивки — высокое качество синтезированного пения может быть получено в пределах диапазона изменения ЧОТ от 70 до 200 Гц, что полностью соответствует большой октаве. Это значит, что диапазон качественного синтеза системы меньше, чем диапазон любого певческого голоса, но достаточен для исполнения народных, детских и некоторых эстрадных песен.

Заключение

В работе описана система синтеза пения для русского языка, реализованная впервые. Описанные алгоритмы языконезависимые и могут применяться для синтеза пения на других языках при добавлении в систему речевой БД соответствующего языка, правил преобразования «буква-фонема» и «фонема-аллофон», а также словаря ударений.

Использование компиляционного метода синтеза и алгоритма «плавной сшивки» для модификации ЧОТ накладывает ограничения на частотный диапазон синтезируемой песни. Эти ограничения могут быть расширены путём пополнения речевой базы несколькими экземплярами вокализованных аллофонов с различными значениями ЧОТ либо же использованием корпусного метода синтеза речи.

Литература

1. *Иванов А.* Искусство пения. / А. П. Иванов. Голос-Пресс, 2006.
2. Max Mathews [Электронный ресурс]. Электронные данные. Режим доступа: http://en.wikipedia.org/wiki/Max_Mathews. Дата доступа: 01.06.2010.
3. KAE Labs Site [Электронный ресурс]. Электронные данные. Режим доступа: http://www.kalabs.com/index.html. Дата доступа: 01.06.2010.
4. Vocaloid official web site [Электронный ресурс]. Электронные данные. Режим доступа: http://www.vocaloid.com/en/index.html. Дата доступа: 01.06.2010.
5. Myriad: Music Notation Software [Электронный ресурс]. Электронные данные. Режим доступа: http://www.myriad-online.com/en/index.htm. Дата доступа: 01.06.2010.

6. E_CANTOR Site [Электронный ресурс]. Электронные данные. Режим доступа: http://www.virsyn.de/en/E_Products/E_CANTOR/ e_cantor.html. Дата доступа: 01.06.2010.
7. *Жадинец Д.В.* Система пения на основе синтеза речи / *Д.В. Жадинец, В.В. Киселёв* // Известия Белорусской инженерной академии. 2004. № 1. Т. 3. С. 81–84.
8. *Matthew L.* Acoustic Models for the Analysis and Synthesis of the Singing Voice. / Georgia Institute of Technology, 2005.
9. Wikipedia, the free encyclopedia [Электронный ресурс]. Электронные данные. Режим доступа: http://en.wikipedia.org/wiki/Neume. Дата доступа: 01.06.2010.
10. The ABC Music project [Электронный ресурс]. Электронные данные. Режим доступа: http://abcnotation.com/. Дата доступа: 01.06.2010.
11. *Вахромеев В.* Элементарная теория музыки. / В.А. Вахромеев. М.: Музыка, 1975.
12. Guitar Pro File Format (.gtp,.gp3,.gp4) [Электронный ресурс]. Электронные данные. Режим доступа: http://www.music-notation.info/en/formats/GuitarProFormat.html. Дата доступа: 01.06.2010.
13. MIDI Manufacturers Association [Электронный ресурс]. Электронные данные. Режим доступа: http://www.midi.org/. Дата доступа: 01.06.2010.
14. MusicXML 2.0 Tutorial [Электронный ресурс]. Электронные данные. Режим доступа: http://www.recordare.com/xml/tutorial.html. Дата доступа: 01.06.2010.
15. *Цирульник Л.И.* Алгоритм генерации фонемной последовательности по орфографическому тексту в системе синтеза речи / Л.И. Цирульник // Информатика. 2006. № 4. С. 61–70.
16. *Лобанов Б.М.* Компьютерный синтез и клонирование речи. / Лобанов Б.М., Цирульник Л.И. Минск, Белорусская наука, 2008.
17. *Moulines E., Charpentier F.* Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones // Speech Communication. 1990. Vol. 9. P. 453–467.
18. *Laroche J., Stylianou Y., Moulines E.* HNS: Speech modification based on a harmonic + noise model // Acoustics, Speech, and Signal Processing: proceedings of IEEE International conference ICASSP-93, Minneapolis, USA, 27–30 April 1993. Minneapolis, 1993. P. 550–553.
19. MuseScore Project Official Website [Электронный ресурс]. Электронные данные. Режим доступа: http://musescore.org/. Дата доступа: 01.06.2010.

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 50 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. E-mail: liliya.tsirulnik@gmail.com

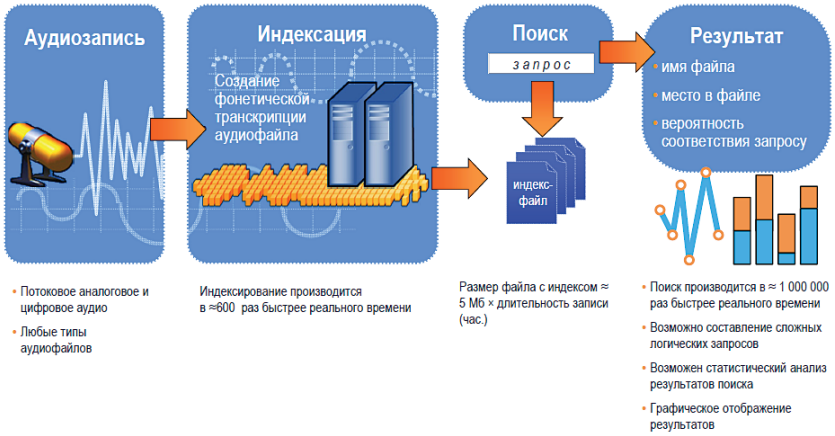
Ломов А.С. —

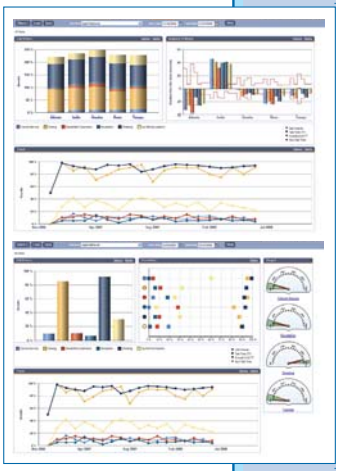
окончил факультет информационных технологий и управления Белорусского государственного университета информатики и радиоэлектроники. Магистрант Института подготовки научных кадров Национальной академии наук Беларуси по специальности прикладная математика. Область научных интересов — теория цифровой обработки сигналов, методы синтеза речи по тексту, речевые компьютерные технологии. E-mail: lomov.as@gmail.com

ЗАО НТЦ «Поиск-ИТ» представляет систему анализа речи на основе технологии фонемного поиска

Информационное пространство, в котором функционируют современные организации, существенно изменилось за последние годы. Крайне высокими темпами растут объёмы создаваемой и хранимой информации, так в 2006 году объём хранимой в мире в цифровом виде информации составлял 161 экзабайт, в то время как в 2010 году, по экспертным оценкам, эта цифра составила уже 988 экзабайт, причём большую часть накопленной в мире информации — около 85% — составляют неструктурированные данные, среди которых наиболее быстрыми темпами растут объёмы аудиовизуальных данных.

В связи с этим за последнее время бурное развитие, как в фундаментальном, так и в прикладном плане получают технологии, позволяющие производить поиск и анализ неструктурированных, и в первую очередь, аудиовизуальных данных.





Научно-технический центр «ПОИСК-ИТ» находится на переднем крае таких исследований и представляет мощное, легко масштабируемое решение для анализа речи, позволяющее за короткое время извлекать и анализировать значимую информацию непосредственно из массивов аудио данных. В данном решении реализована непрерывно совершенствуемая технология фонетического поиска, позволяющая:

- производить поиск любых слов и выражений непосредственно в речевой составляющей файла;
- создавать логически сложные поисковые запросы, позволяющие максимально точно находить нужные сведения;
- проводить статистический анализ результатов поиска, извлекая таким образом новые знания из массивов аудио-данных.

Принципиальной особенностью системы является её способность производить высококачественный поиск по материалам низкого качества (зашумленным записям, записям речи по телефону и пр.).

Немаловажной характеристикой системы является её быстродействие. Так, например, аудиоархив общей продолжительностью звучания 600 часов может быть доступен для поиска менее чем за один час. Непосредственно поиск выполняется со скоростью в 1 000 000 раз превышающей реальное время звучания.

Применение фонемного поиска позволяет значительно повысить эффективность поиска по аудио-данным и открывает пользователям богатейший источник информации.

Предлагаемая технология находит своё применение в различных областях человеческой деятельности, начиная от служб безопасности объектов разного уровня и заканчивая контакт-центрами, обладающими аудио-архивами наибольших объёмов. Применение предлагаемого решения, реализующего технологию фонетического поиска, в контакт-центрах позволяет значительно повысить их эффективность и расширить услуги, предлагаемые заказчикам и абонентам, за счёт возможности содержательного анализа записанных телефонных разговоров, а комбинация такого анализа с традиционными статистическими методами и встроенные аналитические отчёты позволяют поднять на новый уровень качество управления контакт-центром.

ПРЕДСТАВЛЯЕМ КНИГУ

Теория нейронных сетей, развитие которой в значительной степени определяет уровень решения сложных научно-технических задач, связанных с развитием высоких технологий в самых различных отраслях промышленности, народном хозяйстве и военной технике, является важным разделом современных научных исследований.

В августе 2007 г. в издательстве «Springer» вышла из печати на английском языке монография доктора технических наук, профессора А.И. Галушкина «Neural networks Theory», которая содержит результаты многолетних исследований автора в области теории нейронных сетей — логической основы построения принципиально новых, по сравнению с классическими, вычислительных систем — нейрокомпьютеров.

В 2010 г. в издательстве «Горячая Линия — Телеком» опубликована монография А.И. Галушкина «Нейронные сети: основы теории», которая является переводом монографии, выпущенной издательством «Springer».

Монография является одной из немногих, если не единственной монографией российского ученого, в которой представлены предисловия трёх известных учёных с мировым именем:

- Роберта Хехт-Нильсена — ведущего разработчика нейрокомпьютеров в США;
- Лотфи Заде — автора концепции размытой логики;
- Шун-иши Амари — директора института Исследований мозга в Японии.

В частности, Роберт Хехт-Нильсен отмечает следующее: «Эта книга представляет собой долгожданный панорамный обзор советской и российской нейросетевой традиции. Книга является кладезем важных идей и значительных результатов, которые не доступны более нигде в английском варианте. Автор, доктор А.И. Галушкин, является ведущим российским экспертом в области нейронных сетей и был ведущим советским и российским разработчиком нейронных сетей с 1970 г. В этот период доктор А.И. Галушкин имел доступ ко всем важным западным публикациям по нейросетевой тематике. Поэтому ценность этой книги удваивается, так как она написана не просто экспертом, но и человеком, который знает и ссылается на интеллектуальные достижения западной школы. Монография «Нейронные сети: основы теории» является наиболее значимым вкладом в литературу по нейросетевой тематике. Этот найденный клад должен быть использован тысячами исследователей и практиков по всему миру, у которых до сих пор не было возможности воспользоваться плодами советских и российских исследований в области нейронных сетей. Доктора Галушкина следует поздравить и поблагодарить за написание этой монументальной работы — книги, которую мог написать только он. Это по-настоящему дар всему миру».

Лотфи Заде пишет: «Монография профессора А.И. Галушкина имеет множество уникальных свойств, которые в общей сложности делают его работу важным вкладом в литературу по теории нейронных сетей. Он и его издатель заслуживают щедрых благодарностей и поздравлений от всех, кто всерьёз имеет интерес к созданию, развитию и текущему положению дел теории нейронных сетей».

А профессор Амари отмечает: «Профессор А.И. Галушкин, ведущий специалист по теории нейронных сетей в России, использует математические методы в комбинации с теорией сложности, нелинейной динамикой и оптимизацией, а также другими концепциями, крепко укоренившимися в российской научной школе. Его теория очень обширна: она охватывает не только традиционные аспекты, такие как архитектура сети, но также рассматривает континуальные нейронные сети в пространствах функций. Я с большим удовольствием воспринял выход книги, в которой эта теория описана во всей своей полноте. Огромная ценность самой теории и используемого автором метода описания такого сложного явления, как нейросетевая система, не может вызывать никаких сомнений».

От редакции: необходимо отметить, что изданная «Springer» монография является естественным продолжением большого числа работ, опубликованных автором за сорок лет работы в этой области. Результаты работ автора в области теории нейронных сетей стали основой для развития нейроматематики — нового раздела вычислительной математики, связанного с решением сложных математических задач в нейросетевом логическом базисе, а также основой для развития нейроуправления — нового раздела теории управления, ориентированного на сложные нелинейные, многомерные объекты управления с переменными параметрами и структурой.