Синтез пения для русского языка

Л.И. Цирульник,
кандидат технических наук, доцент

А.С. Ломов,
студент

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 являются языконезависимыми и могут применяться для синтеза пения на других языках.

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], французском, португальском, итальянском языках: программа компании Myriad [5], 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], обладают радом 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. Указанные 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 4 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 – эта 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 актером.

Для того чтобы синтезировать голос с хорошими вокальными данными, нужно выделить отличия профессионального пения от любительского. Наиболее заметное отличие проявляется в более четком выделении первой, второй и третьей форманты у профессиональных певцов. Кроме этого, 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 четкую структуру.

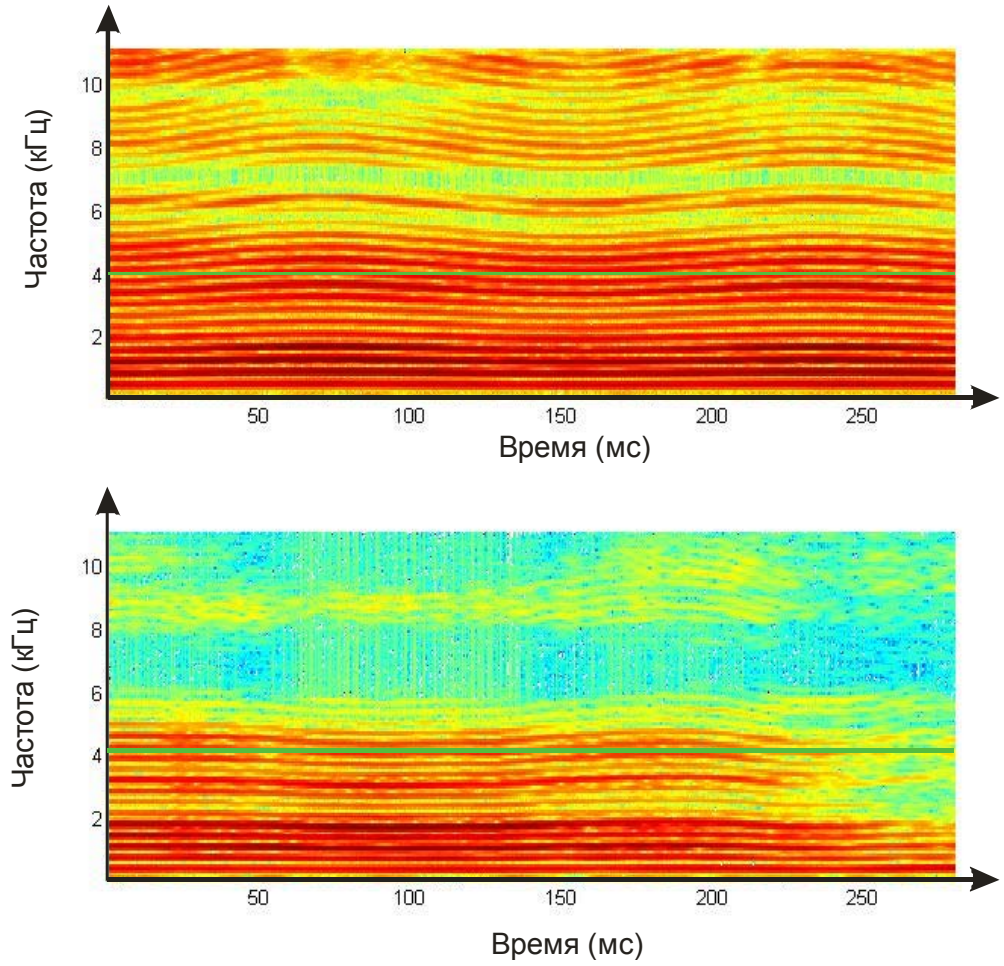

Рисунок 1 – Спектрограммы исполнения одного и того же песенного фрагмента опытным вокалистом (сверху) и певцом без подготовки (снизу)

Для моделирования певческого голоса с большим уровнем естественности звучания следует остановить внимание на приемах, которые используются при пении. Одним из широко распространенных и часто используемых вокальных приемов, как в академической школе, так и при эстрадном исполнении, является 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.

На рисунке 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 приемы 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 объеме 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.

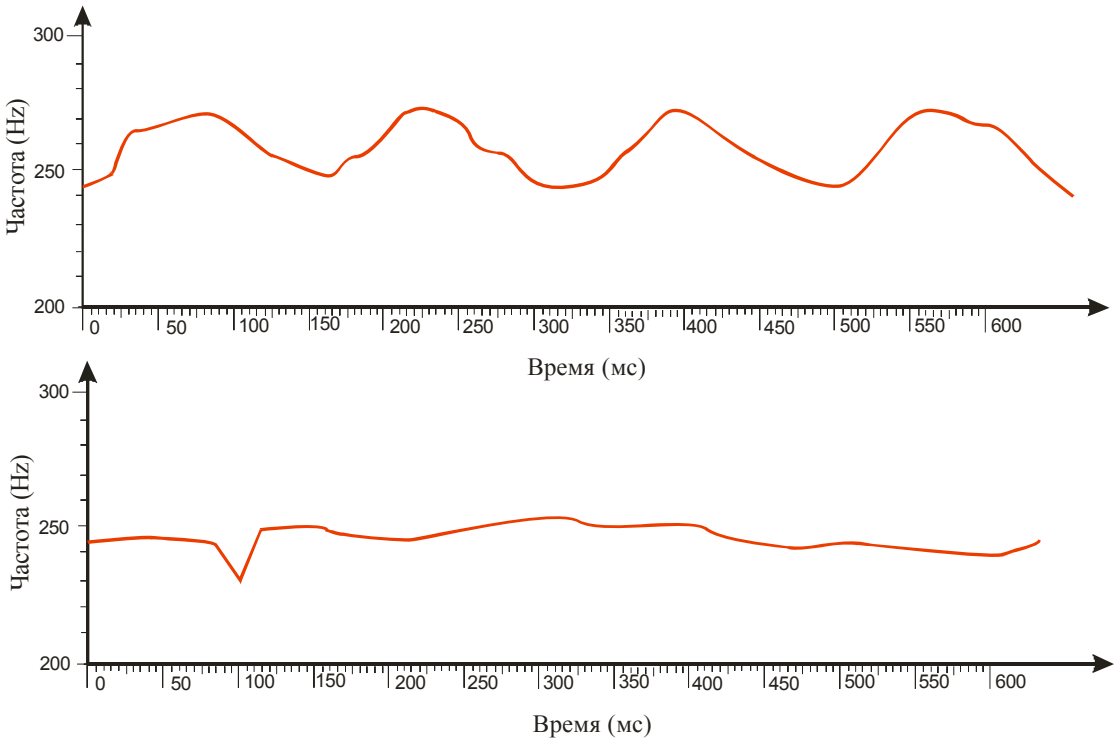

Рисунок 2 – Графики зависимостей ЧОТ от времени при исполнении вибрато опытным вокалистом (сверху) и певцом без подготовки (снизу).

2. Описание системы 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 подается на вход системы (рис. 3), затем из него выделяется музыкальная нотация и текст песни.

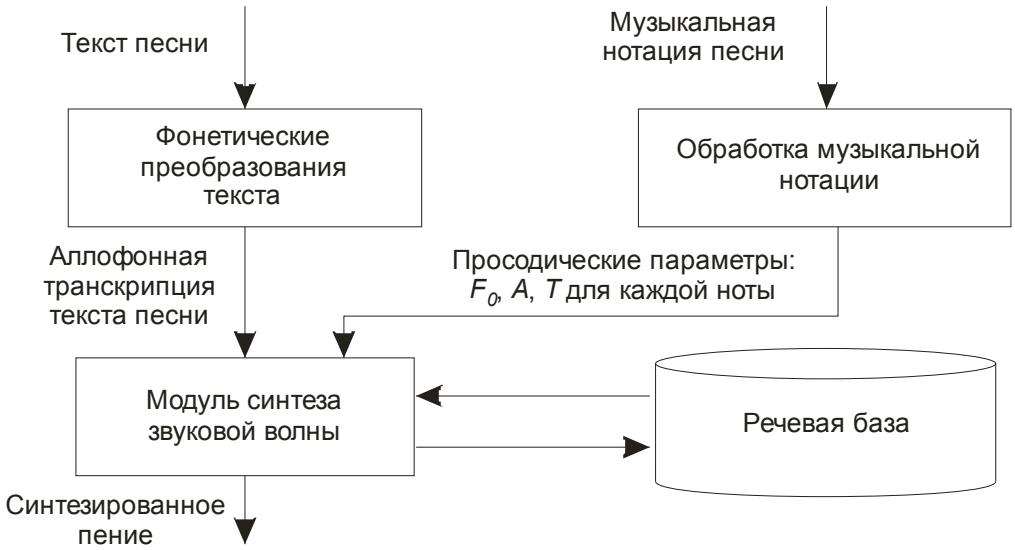

Рисунок 3 – Общая схема работы системы синтеза пения

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.

Результатом обработки текста фонетическим анализатором является 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: F_0, A, T. Существует множество форматов представления музыкальных произведений 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 просто

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:

$$F_0 = f_0 \cdot 2^{n/12} \quad (1)$$

где f_0 – частота исходной ступени,

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

Длительность звучания ноты T вычисляется по формуле

$$T = 4 \cdot r \cdot t_0 \quad (2)$$

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

t_0 – 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 подается текст, разделенный на слоги. Внутри процессора он проходит три этапа: расстановку ударений, преобразование «буква-фонема» и преобразование «фонема-аллофон». Выходными данными является последовательность аллофонов, разделённая на слоги.

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.

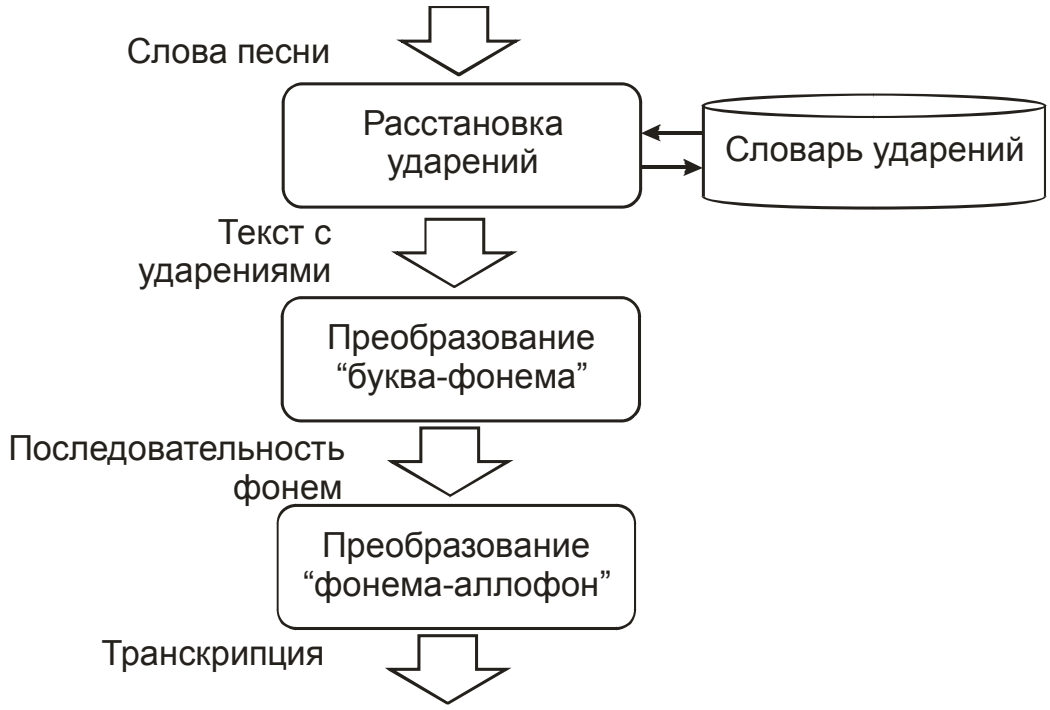

Рисунок 4 – Схема работы фонетического анализатора.

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 осуществления 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.

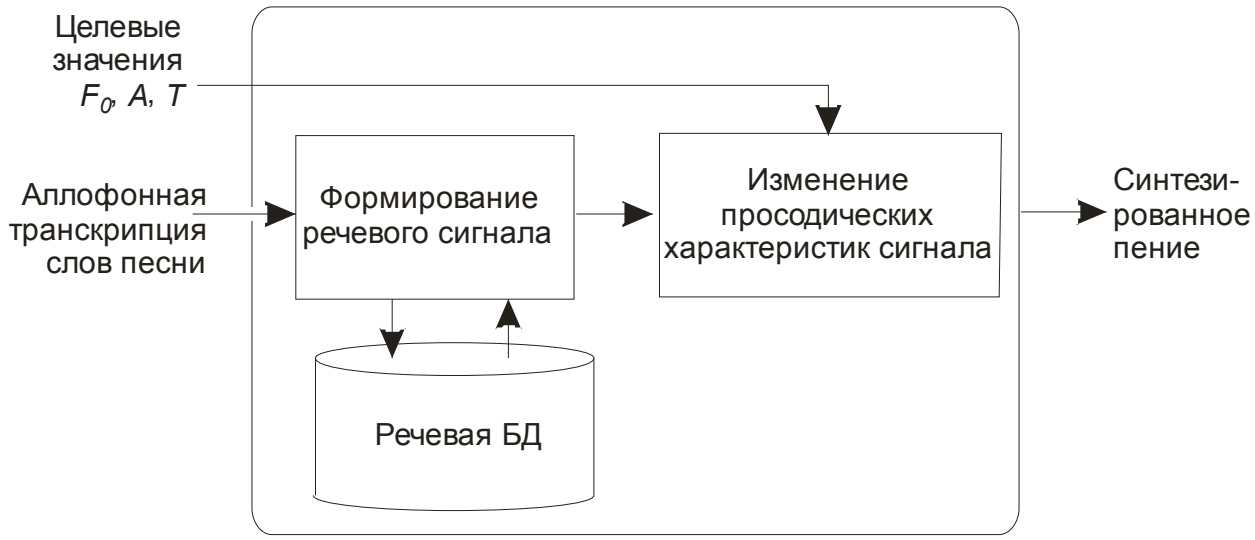

Рисунок 5 – Схематическое представление модуля синтеза сигнала.

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, достоинствами которого являются 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:

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

где $s(n)$ – это отрезок исходного сигнала длительностью в один период основного тона;

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

$k(n)$ – это 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 приведен пример увеличения длительности одного из периодов основного тона фонемы /a/. В этом случае длительность периода увеличивается с 241 до 361 отсчетов.

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

Данный алгоритм дает 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.

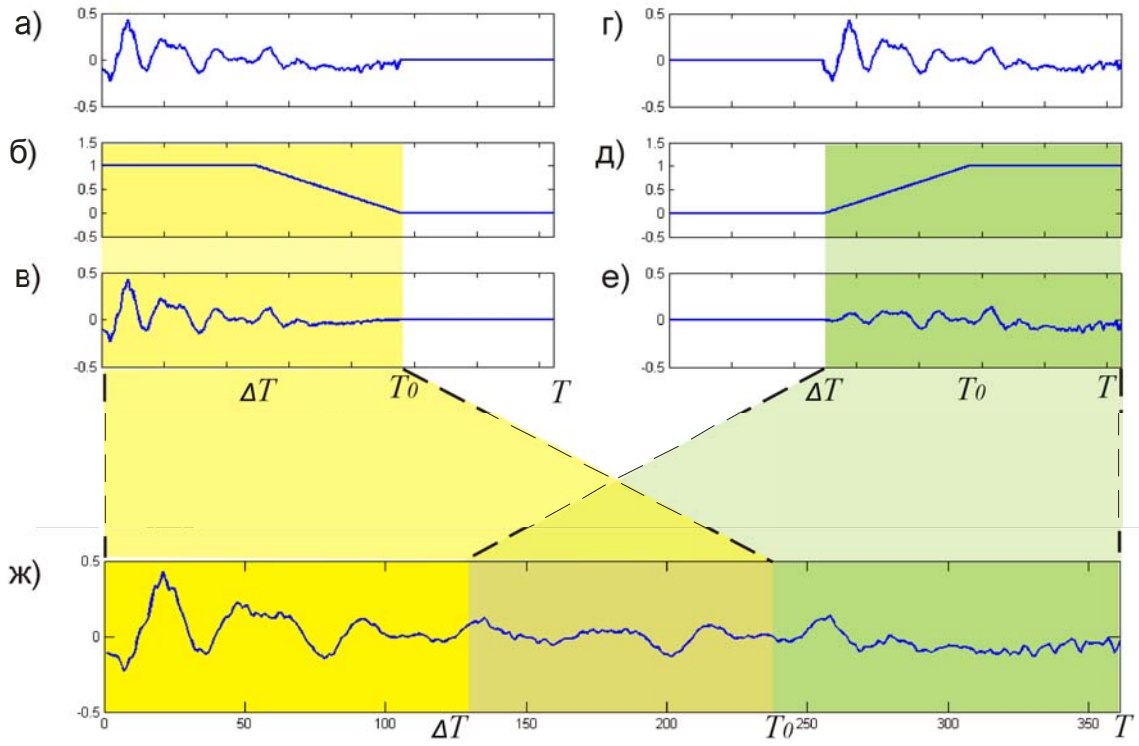

Рисунок 6 – Иллюстрация последовательной обработки исходного сигнала $s(t)$ методом «плавной сшивки» при увеличении длительности периода основного тона: а) исходный сигнал $s(t)$; б) кусочно-линейная функция $k(t)$; в) первое слагаемое результирующего сигнала $s(t)*k(t)$; г) сдвинутый сигдал $s(t+\Delta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

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 приведен пример внешнего вида программы. Информация о словах песни, их транскрипция и осциллограмма синтезированного звука отображается 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.

В системе используется речевая БД мужского голоса, содержащая 3000 речевых отрезков. Среднее значение ЧОТ вокализованных элементов БД – 100 Гц. Таким образом, в соответствие 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.

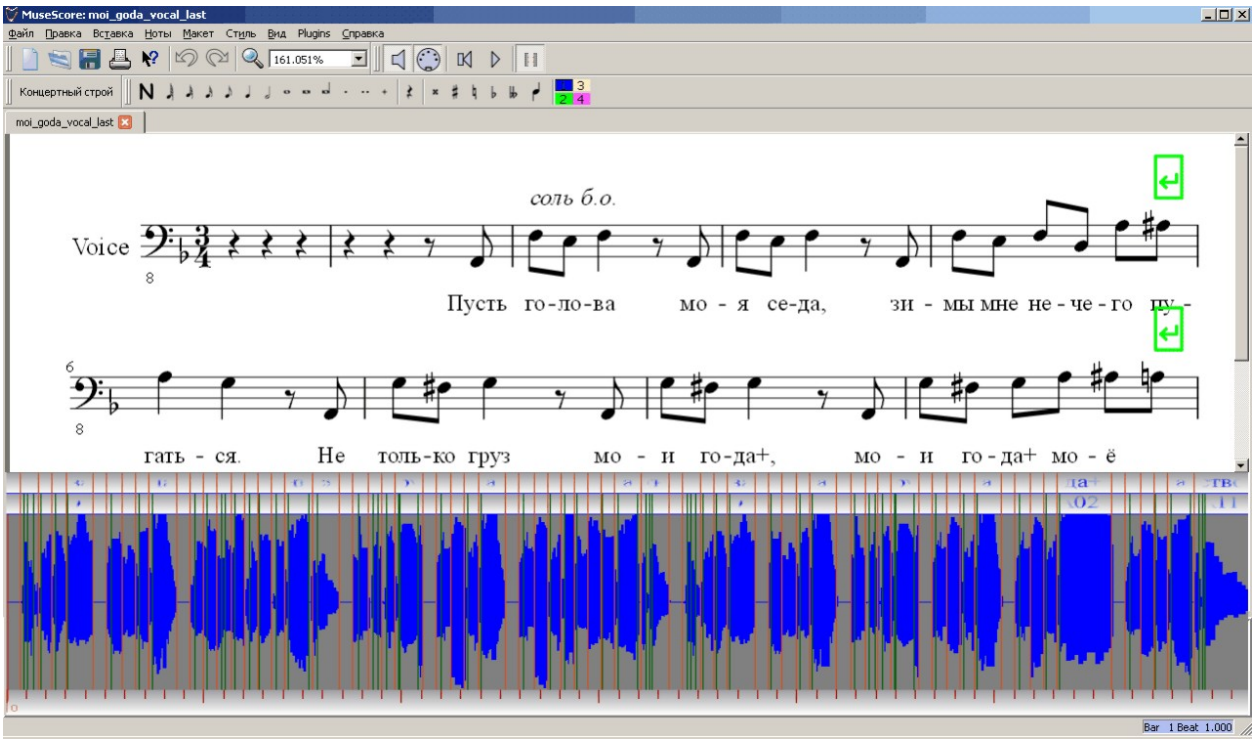

Рисунок 7 – Внешний вид окна программы синтеза пения

Заключение

В данной работе описана система синтеза пения для русского языка, реализованная впервые. Описанные алгоритмы являются языконезависимыми 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.

Список источников:

1. Иванов, А. Искусство пения. / А. П. Иванов – Голос-Пресс, 2006. – 235 стр.
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.kaelabs.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 – 127 p.
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. – 7ое изд. – Москва: «Музыка», 1975. – 228с.
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. – 342 стр.
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.