АЛГОРИТМЫ СИНТЕЗА ПРОСОДИЧЕСКИХ ХАРАКТЕРИСТИК РЕЧИ ПО ТЕКСТУ В СИСТЕМЕ «МУЛЬТИФОН»[1]

ALGORITHMS OF SPEECH PROSODIC CHARACTERISTICS SYNTHESIS IN "MULTIPHONE" TTS SYNTHESIS SYSTEM

Цирульник Л.И.(liliya_tsirulnik@ssrlab.com), Жадинец Д.В.(dmitry@ssrlab.com)

Лобанов Б.М.(lobanov@newman.bas-net.by), Сизонов О.Г. (oleg@ssrlab.com)

Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

В работе представлена ПАЕ-модель синтеза просодических характеристик речи. Описаны принципы создания просодических портретов акцентных единиц синтагмы, приведена структура подсистемы синтеза просодических характеристик речи, показано применение ПАЕ-модели в многоязычной системе синтеза речи по тексту «Мультифон».

Введение

Для определения просодических параметров при синтезе речи по тексту используются разнообразные просодические модели, в частности, автосегментная метрическая модель, которая представляет интонацию как последовательность уровней тона [1], акустико-фонетическая суперпозиционная модель, интерпретирующая интонацию как последовательность событий с перекрывающимися областями [2], модель IPO, передающая интонацию как последовательность дискретных событий [3], и модель Tilt, которая использует непрерывную параметризацию контуров частоты основного тона F_0 [4].

Большинство моделей интонации, упомянутых выше, были разработаны и применены для английского, французского, немецкого, голландского и некоторых других языков. Существует только небольшое число примеров разработки и использования этих моделей для русского языка. Эффективное перцептивное описание интонации русского языка в соответствии с моделью IPO было разработано С. Оде [5]. Суперпозиционная модель была использована для русскоязычного синтеза речи компанией Bell Labs - Lucent Technologies [6].

Основной принцип синтеза просодических характеристик, используемый нами, основан на модели, которая ближе всего к Tilt-модели, но отличается от неё методом представления просодических характеристик синтагмы. Синтагма в нашей модели представляется последовательностью просодических портретов акцентных единиц (ПАЕ). Данная модель была предложена более десятка лет назад [7] и с тех пор успешно использовалась в нескольких системах синтеза русской речи по тексту. В модели аккумулированы как собственные данные о просодике русской речи, так и полученные рядом известных исследователей русской интонации: Е.А. Брызгуновой, С.В. Кодзасовым, О.Ф. Кривновой, Н.Д. Светозаровой и некоторыми другими авторами. Описанное в данной работе исследование выполнено в контексте создания единой системы синтеза речи по тексту для славянских языков [8], а также в рамках продолжающихся исследований по компьютерному клонированию персональных характеристик речи диктора, которое включает просодические исследования.

В работе описываются принципы создания просодических ПАЕ синтагм, приведена структура подсистемы синтеза просодических характеристик речи, показано применение ПАЕ-модели в многоязычной многоголосой системе синтеза речи по тексту «Мультифон».

1. Основные принципы синтеза просодических характеристик речи

В соответствии с ПАЕ-моделью [8], минимальной просодическим компонентом, из которого составляется интонация синтагмы, является Акцентная Единица (АЕ). АЕ может состоять из одного или более фонетических слов, но должна иметь в своём составе только один полноударный слог. Каждая АЕ, в свою очередь, состоит из ядра (полноударная гласная фонема), предъядра (все фонемы, предшествующие полноударной гласной) и заядра (все фонемы за полноударной гласной). Основное предположение ПАЕ-модели состоит в том, что для определенного типа интонации синтагмы топологические свойства просодических параметров каждой АЕ не изменяются (или изменяются незначительно) с изменениями её фонетического содержания и количества слогов в предъядре и заядре АЕ. Этот факт иллюстрируется рис. 1, где показаны контуры F_0 для однословных вопросительных синтагм с различным положением словесного ударения, т.е. с различным количеством слогов в предъядре и заядре АЕ.

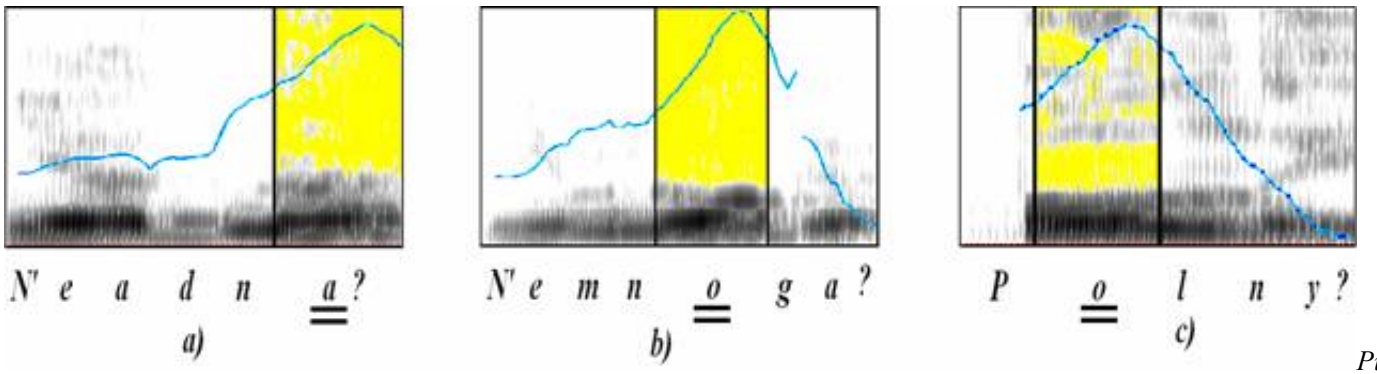

Рис. 1. Контуры F_0 для однословных вопросительных синтагм: a) "Не одна?", b) "Не много?, c) "Полный?" (ударные гласные подчеркнуты двойной чертой)

АЕ может состоять также и из более чем одного фонетического слова, но при условии, что она содержит лишь один полноударный слог. Это иллюстрируется на рис 2, где представлены контуры F_0 для трёхсловных вопросительных синтагм с тремя различными положениями полноударного слова в синтагме.

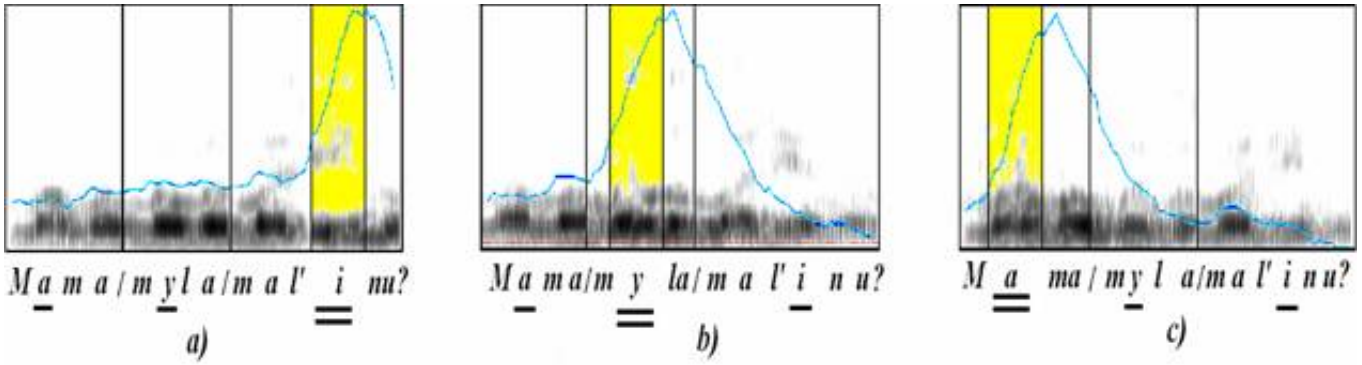

Рис. 2. Контуры F_0 вопросительной синтагмы "Мама мыла малину?" с полноударными словами a) "малину", b) "мыла", c) "мама" (полноударные гласные подчеркнуты двойной чертой, частично ударные - одинарной)

Как видно из рис. 2, каждая из этих синтагм состоит только из одной АЕ, а поведение контура F_0 подобно поведению на ядре, предъядре и заядре однословной синтагмы, показанной на рис. 1.

Все упомянутое выше дает нам серьезные основания к тому, чтобы представить ПАЕ контура F_0 в нормированном пространстве «частота-время» с равной относительной длительностью трёх частей АЕ - ядра, предъядра и заядра. На рис. 3a показан контур F_0 ПАЕ однословной синтагмы, а на рис. 3b – трёхсловной синтагмы, полученные из рис.1, 2. Как видно из рис. 3a и 3b, различие в их ПАЕ малосущественно.

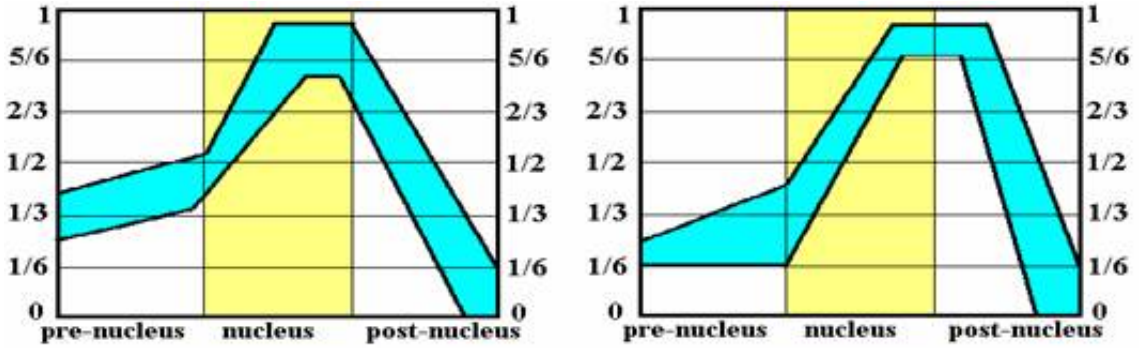

Рис. 3. ПАЕ вопросительного типа интонации для a) однословных синтагм и b) трёхсловных синтагм

Как показал опыт, отмеченные выше закономерности создания F_0-ПАЕ для вопросительного типа интонации справедливы также для других интонационных типов: завершённости, незавершённости, вводности и др. Подобное заключение может быть также сделано относительно возможности создания ПАЕ для динамических (A_0-ПАЕ) и ритмических (T_0-ПАЕ) характеристик просодики.

Рассмотренные примеры касались только одноакцентных (содержащих одну АЕ) синтагм. Однако синтагма может включать также 2 и более АЕ. Основные принципы создания интонационных ПАЕ для синтагмы, состоящей из 3-х АЕ, проиллюстрированы на рис. 4 на примере синтагмы: *"которые могут быть представлены*", произнесённой с интонацией незавершённости.

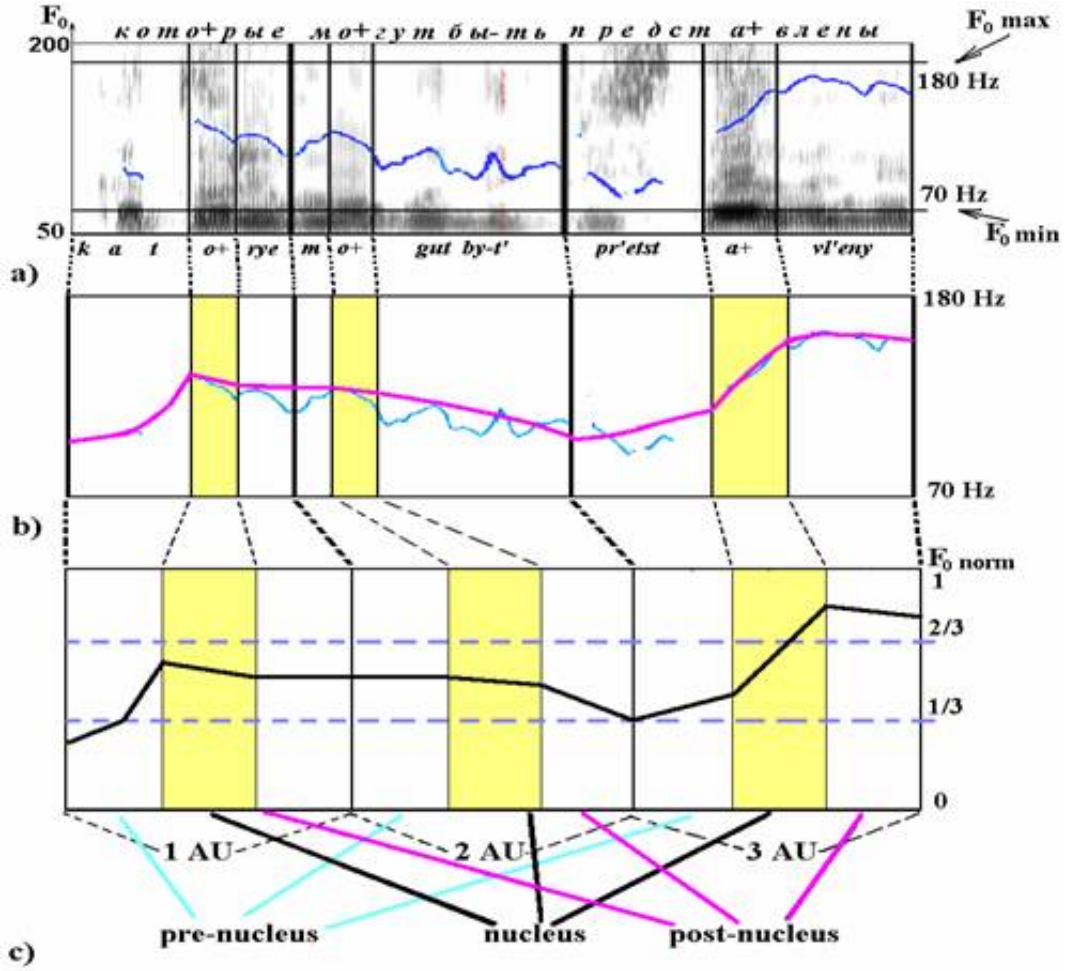

Рис. 4. Этапы создания ПАЕ: a) вычисление значений F_0, b) интерполяция F_0, c) нормализация контура F_0

На первом этапе создания ПАЕ вычисляются значения F_0 для каждого вокализованного речевого сегмента (рис. 4 a). Затем отмечаются границы каждой АЕ, а также области предъядра, ядра, и заядра для каждой АЕ. Значения F_0 интерполируются на невокализованные участки (рис. 4 b). На конечном этапе осуществляется нормализация контура по частоте и длительности (рис. 4 c).

Для нормализации по частоте определяются минимальное ($F_{0\ min}$) и максимальное ($F_{0\ max}$) значения F_0 на всей анализируемой фонограмме. Как правило, значение $F_{0\ max}$ расположено на ядре АЕ восклицательной синтагмы, в то время как $F_{0\ min}$ связано с ядром АЕ синтагмы, находящейся в конце абзаца. Для нормализации F_0 используется следующая формула:

$$F_{0norm} = \frac{F_0 - F_{0min}}{F_{0max} - F_{0min}}$$

Для данного диктора $F_{0\ min}$ была равна 70 Гц, $F_{0\ max}$ – 180 Гц, что отражено на рис. 4 a. Нормализация длительностей элементов АЕ осуществляется путём приведения к стандартной длине предъядерных, ядерных и заядерных участков (см. рис. 4 c).

Таким образом, мы получаем ряд нормализованных ПАЕ для синтагм различных интонационных типов. Эти нормализованные последовательности ПАЕ используются затем системой синтеза речи по тексту независимо от фонетического содержания конкретных АЕ.

2. Общая структура подсистемы синтеза просодических характеристик речевого сигнала

Структура подсистемы синтеза просодических характеристик в синтезаторе речи «Мультифон» показана на рис. 5. Орфографический текст для синтеза речи до подачи на вход подсистемы синтеза просодических характеристик подвергается процедуре нормализации, в процессе которой осуществляется преобразование к текстовому виду следующих символов:

- многоразрядных целых и дробных чисел;
- римских цифр;
- времени и даты;

- телефонных номеров;
- математических знаков;
- аббревиатур;
- сокращений;
- интернет адресов;
- букв латинского алфавита.

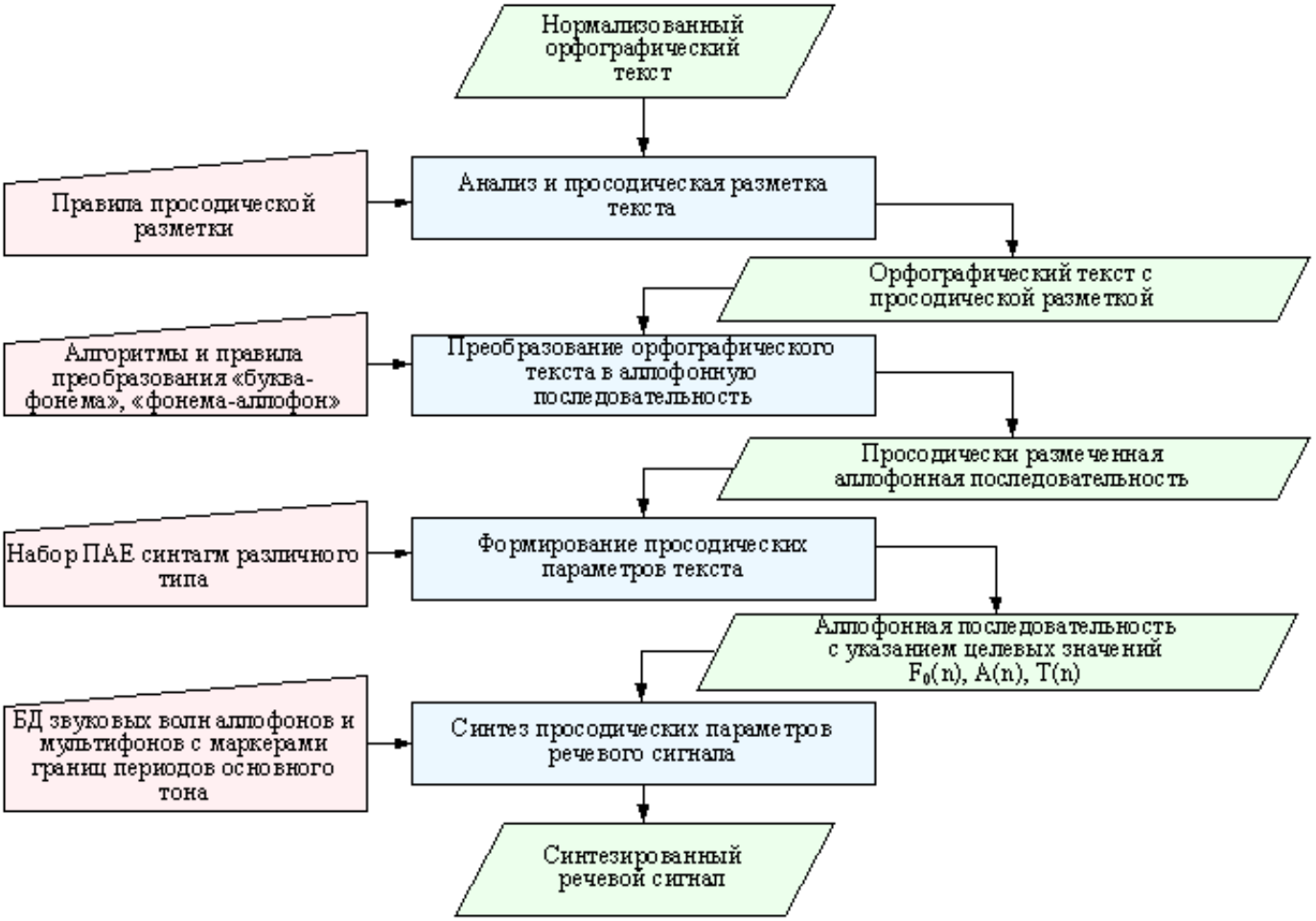

Рис. 5. Общая структура подсистемы синтеза просодических характеристик речевого сигнала

Первый из блоков подсистемы синтеза просодических характеристик речевого сигнала, используя языко-зависимые ресурсы и правила, осуществляет анализ и просодическую разметку нормализованного орфографического текста. Подробная структура блока анализа и разметки текста представлена на рис. 6.

Второй блок – блок преобразования орфографического текста в аллофонную последовательность – использует языко-зависимые алгоритмы и правила преобразования «буква-фонема», «фонема-аллофон», подробно описанные в работе [9].

Действия, осуществляемые в двух последующих блоках – блоках формирования просодических параметров текста и синтеза просодических параметров речевого сигнала – являются фактически обратными действиям, показанным на рис. 4, т.к. здесь происходит “натягивание” нормированных ПАЕ на полученные текстовые АЕ в соответствии с интонационным типом синтагм.

Ресурсом блока формирования просодических параметров текста является полный набор ПАЕ синтагм всех используемых интонационных типов и подтипов. В этом блоке, в соответствии с указанным интонационным типом синтагмы и положением АЕ в синтагме для каждого *n*-го аллофона устанавливается от 2-х до 8-ти значений частоты основного тона – $F_0(n)$ – и по одному значению длительности – $T(n)$ и *амплитуды* – $A(n)$. Структура блока формирования просодических параметров текста представлена на рис. 7.

Полученный аллофонный текст с указанием значений $F_0(n)$, $T(n)$, $A(n)$ поступает на вход блока синтеза просодических характеристик речевого сигнала. В соответствие с поступившим аллофонным текстом из БД звуковых волн элементов компиляции выбирается необходимая последовательность мультифонов, размеченных предварительно на периоды основного тона (питчи). Синтез просодически модифицированного речевого сигнала осуществляется путём изменения акустических характеристик звуковых волн мультифонов в соответствие с целевыми значениями $F_0(n)$, $T(n)$, $A(n)$. Структура блока синтеза просодических характеристик речевого сигнала представлена на рис. 8.

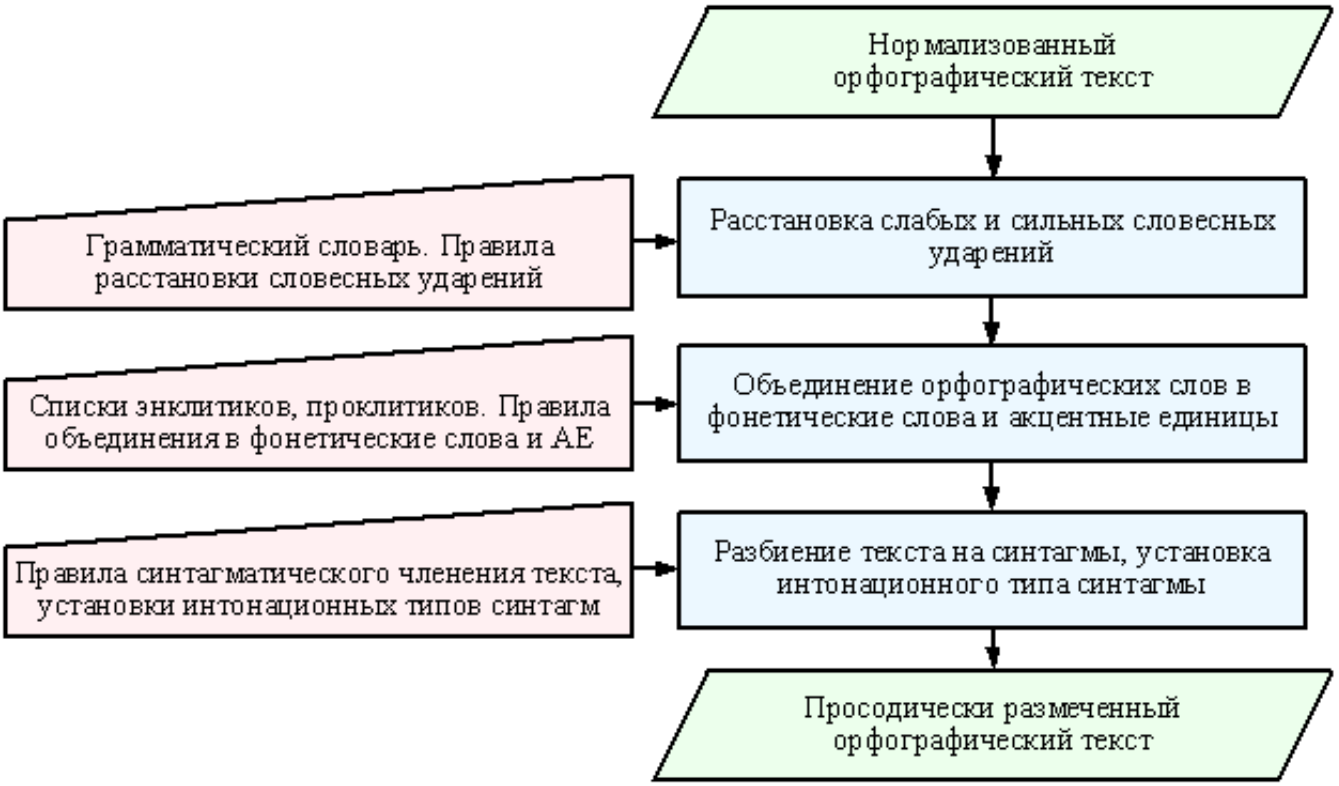

Рис. 6. Структура блока анализа и просодической разметки текста

Анализ и просодическая разметка текста (рис.6) происходит в несколько этапов. На первом этапе осуществляется расстановка сильных и слабых словесных ударений, для чего используется грамматический словарь словоформ, содержащий пометы позиции ударения каждой словоформы, а также правила расстановки ударений, которые учитывают, в частности, принадлежность слова к знаменательным или служебным частям речи, его положение в предложении и ближайшее окружение. На следующем этапе – этапе объединения орфографических слов в фонетические слова и АЕ – используются списки энклитиков и проклитиков, а также правила объединения в фонетические слова и АЕ, которые также учитывают принадлежность «смежных» слов к определённым частям речи. На этапе разбиения текста на синтагмы и установки интонационного типа синтагм – завершающем этапе анализа и просодической разметки – используются правила синтагматического членения текста, согласно которым количество АЕ в синтагме не может превышать четырёх. Правила синтагматического членения и определения интонационных типов используют явные маркеры границ синтагм в тексте: знаки препинания, а также неявные, в частности, сочинительные и подчинительные союзы. Выходным данным блока анализа и просодической разметки является текст с пометами позиций ударения, границ фонетических слов, АЕ, синтагм и указанием интонационных типов каждой синтагмы.

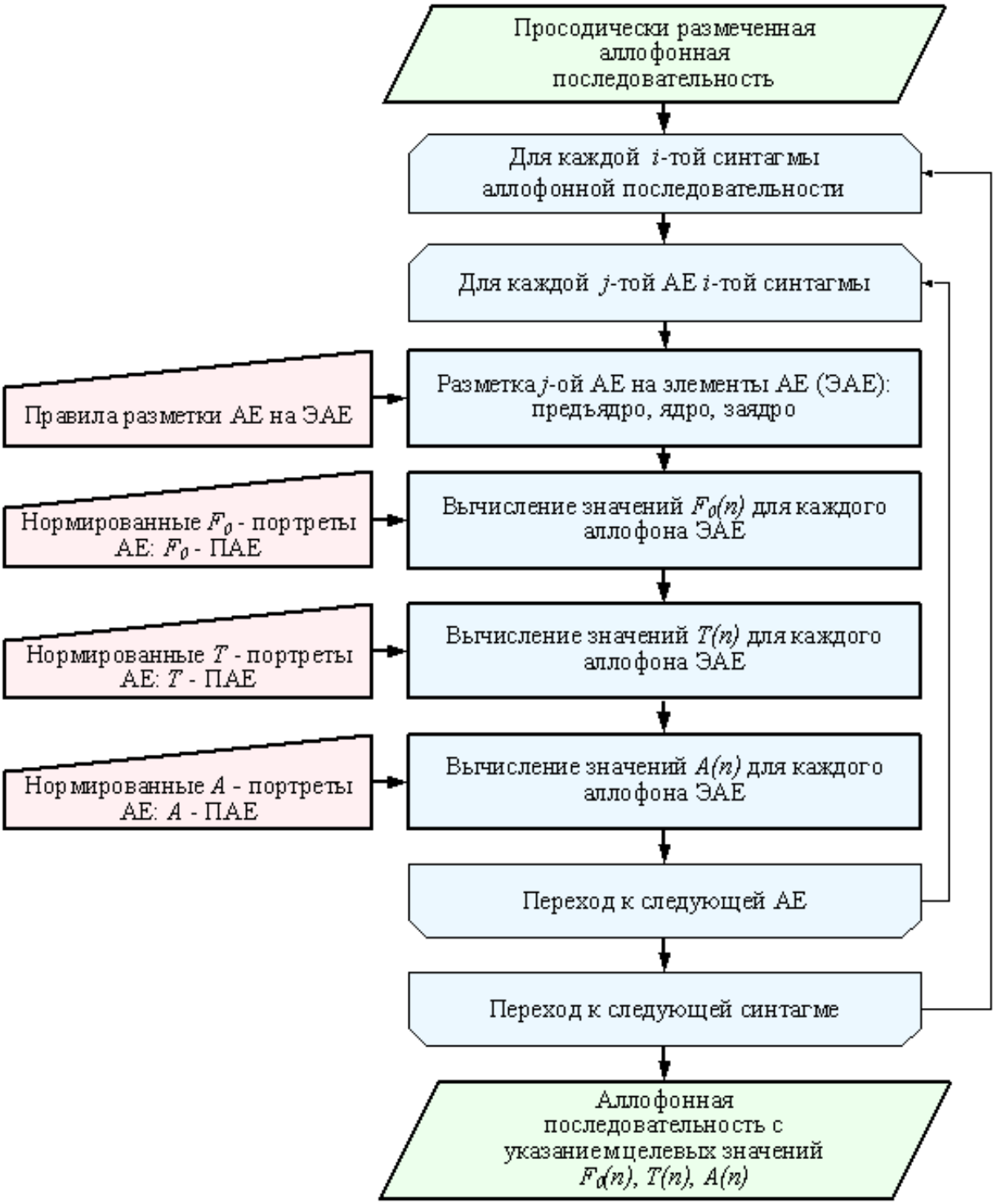

Рис. 7. Структура блока формирования просодических параметров

Формирование просодических параметров текста (рис. 7) осуществляется последовательно для каждой синтагмы. На первом этапе осуществляется разметка каждой АЕ синтагмы на ЭАЕ: предъядро, ядро, заядро. Ядром синтагмы, согласно используемым правилам, является полноударный гласный; все аллофоны, предшествующие полноударному гласному, являются предъядерным участком, все следующие за полноударным гласным аллофоны – заядерным участком. Затем с использование нормированных портретов F_0 *-ПАЕ*, *A - ПАЕ*, *T* - ПАЕ для синтагмы соответствующего интонационного типа осуществляется вычисление значений $F_0(n)$, $A(n)$, $T(n)$ каждого n-го аллофона элементов предъядра, ядра, заядра j-той АЕ.

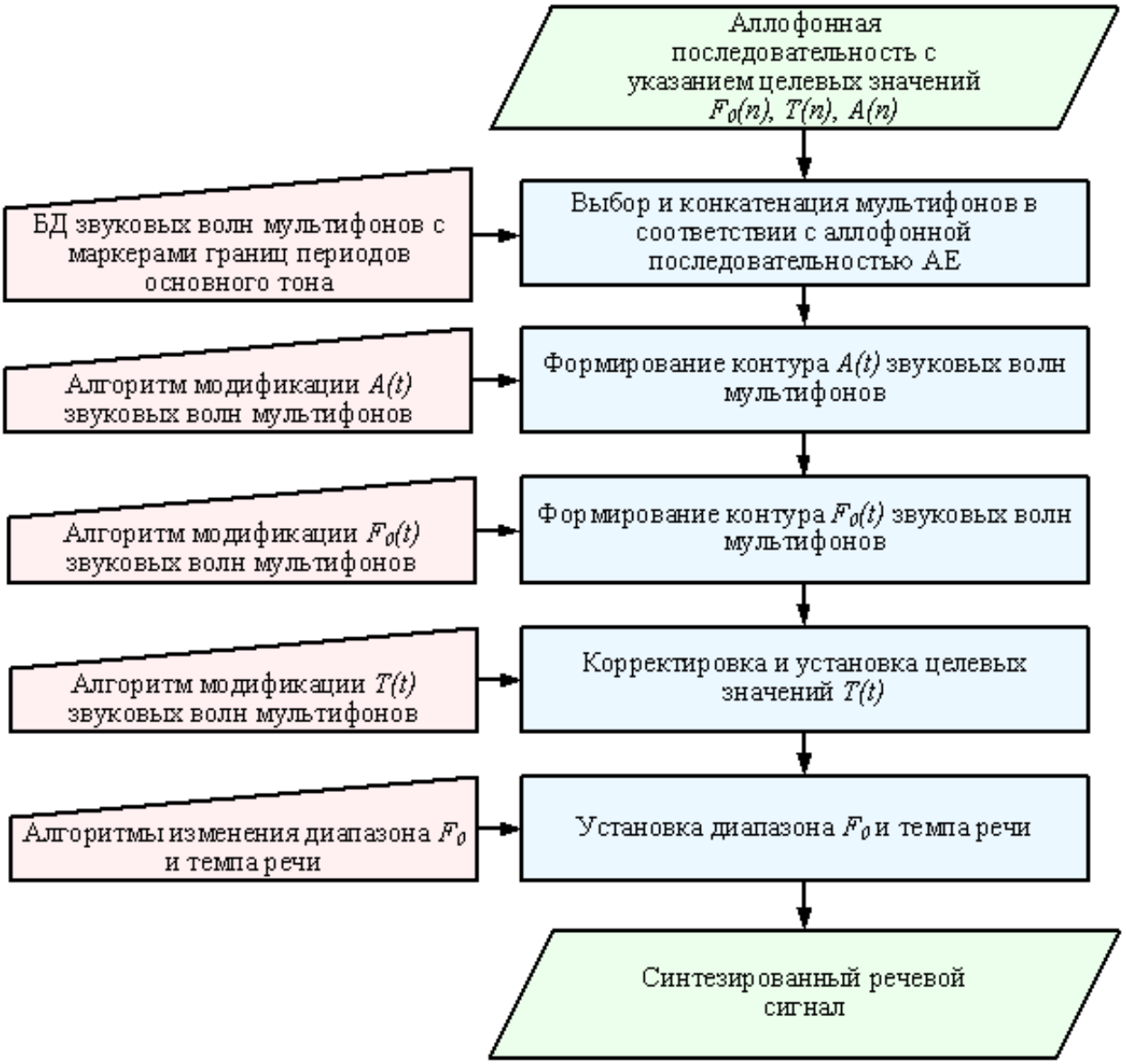

Рис. 8. Структура блока синтеза просодических характеристик речевого сигнала

Процедура синтеза речевого сигнала с заданными просодическими характеристиками $A(t)$, $F_0(t)$, $T(t)$ представлена на рис. 8. В соответствие с поступающим аллофонным текстом из БД звуковых волн выбирается требуемая последовательность мультифонов и осуществляется их конкатенация. Далее на основе заданных на основе ПАЕ целевых значений $A(n)$, $F_0(n)$, $T(n)$ формируются контуры $A(t)$, $F_0(t)$ и устанавливаются длительности $T(t)$ звуковых волн аллофонов предъядра, ядра и заядра. При формировании контуров $A(t)$ учитывается фактор просодической изменчивости силы звука, представляющего каждый конкретный аллофон. Для формирования мелодического контура $F_0(t)$ используется SL-алгоритм [10], который позволяет осуществлять "щадящую" модификацию ЧОТ путём "плавной сшивки" ("Soft Lacing") соседних периодов естественного сигнала на интервалах открытой голосовой щели, сохраняя речевой сигнал неизменённым на остальных участках. Установка значений $T(t)$ осуществляется в соответствии с заданными целевыми значениями длительности звуковых элементов АЕ и корректируется затем с учетом её качественного и количественного состава. В блоке предусмотрены регулировки диапазона изменения $F_0(t)$ и темпа речи. Регулировка диапазона изменения $F_0(t)$ осуществляется путём перерасчёта нормированных значений в соответствие с формулой:

$$F_{0РС}(t) = F_{0НАК}(t) * (F_{0_{max}} - F_{0_{min}}) + F_{0_{min}},$$

где $F_{0\ min}$, $F_{0\ max}$ задают требуемый диапазон изменений $F_0(t)$. Регулировка темпа речи осуществляется путём корректировки длительности звуковых элементов АЕ и межсинтагменных пауз с учётом коэффициента "податливости" каждого конкретного звука темповым изменениям.

3. Реализация просодического блока в системе синтеза речи по тексту

Реализация подсистемы синтеза просодических характеристик осуществляется с помощью просодического

блока, интерфейс которого показан на рис. 9. Настройки блока включают выбор одного из имеющихся наборов ПАЕ – просодических стилей (“Prosody style”), указание диапазона изменения частоты основного тона (“Frequency range”), опции использования интонационных, ритмических и динамических контуров ПАЕ при синтезе просодических параметров речевого сигнала (“Use frequency”, “Use rhythm”, “Use energy”), а также указание значений темпа речи (“Tempo”) и уровня громкости (“Volume”).

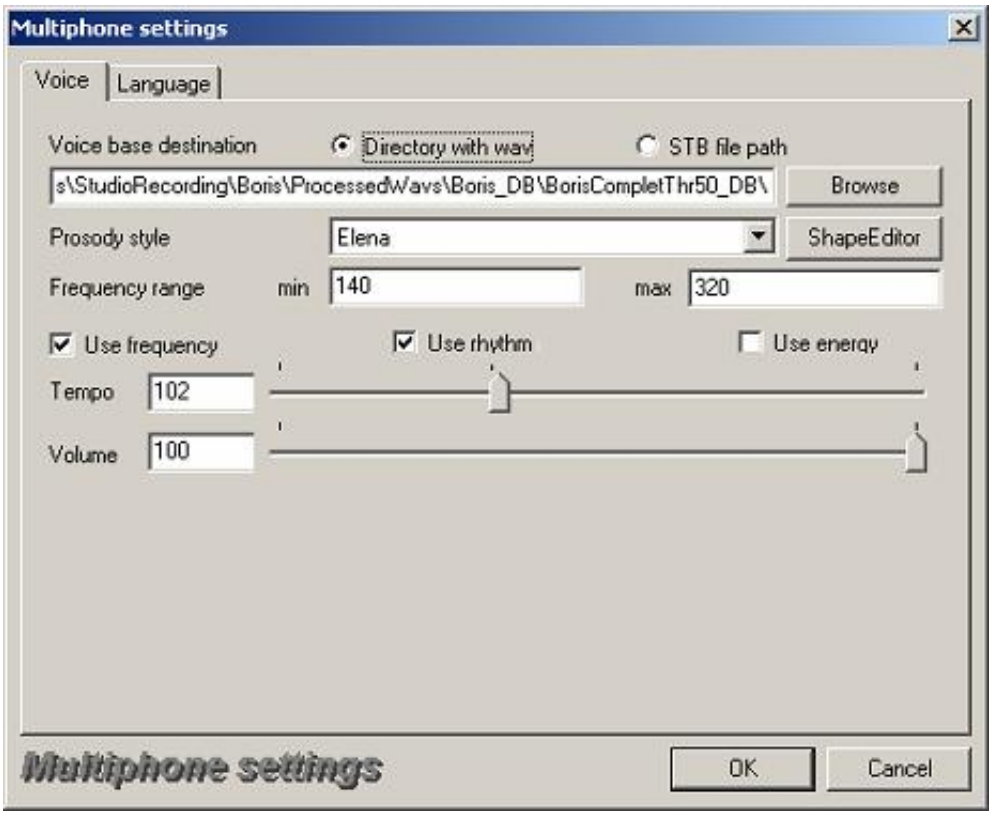

Рис. 9. Интерфейс просодического блока в системе синтеза речи по тексту

Используемые просодические стили отражают языкозависимые и дикторозависимые особенности интонации. Каждый стиль содержит наборы ПАЕ следующих интонационных типов: завершённость, незавершённость, вопрос, восклицание. Кроме того, каждый интонационный тип имеет несколько подтипов. В настоящее время используется 14 вариантов интонации завершённости, 15 вариантов незавершённости, 7 вариантов вопроса и 4 варианта восклицания. Каждый из интонационных подтипов, в свою очередь, содержит нормированные контуры F_0, A, T акцентных единиц, составляющих синтагму. Количество акцентных единиц в синтагме, согласно ограничениям разработанного просодического блока, не может быть более четырёх.

В системе реализована возможность добавления и редактирования контуров ПАЕ. Интерфейс блока редактирования ПАЕ (рис. 10) позволяет просматривать и изменять нормированные контуры F_0, A, T, выполняя тем самым эффективное регулирование просодических портретов.

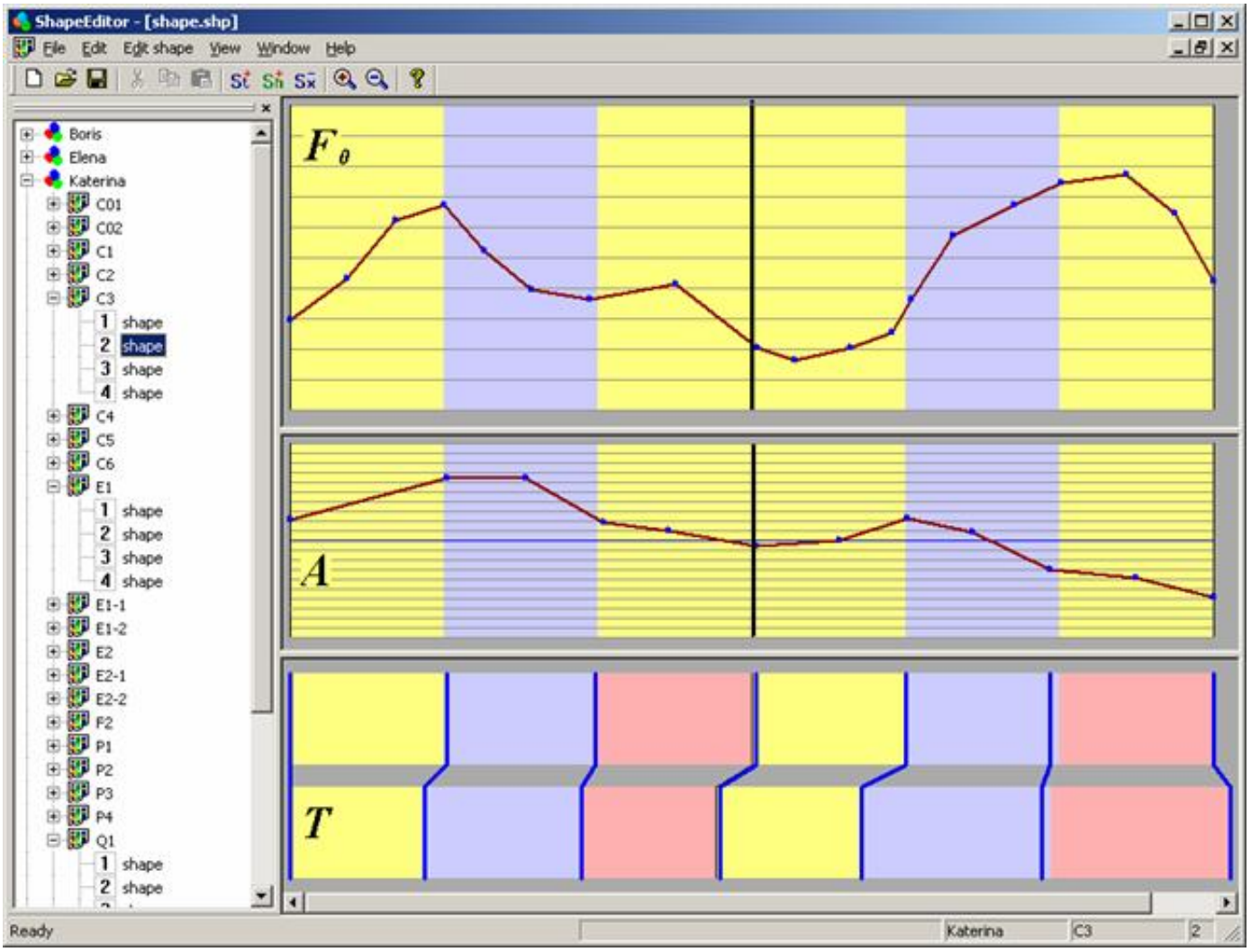

Рис.10. Интерфейс блока редактирования ПАЕ в системе синтеза речи по тексту (пример контуров F_0, A, T для двухакцентной синтагмы незавершённого типа)

Заключение

Просодическая модель ПАЕ, описанная в работе, эффективно используется в многоязычной системе синтеза речи по тексту «Мультифон». В настоящее время в рамках ПАЕ-модели созданы 4 различных просодических стиля, отражающих языкозависимые и дикторозависимые характеристики речи. Образцы синтезированной речи будут продемонстрированы участникам конференции во время доклада.

Список литературы

1. Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J., Hirschberg J. TOBI: a standard for labelling English prosody // International Conference on Spoken Language Processing: proc. of the ICSLP'92. Alberta: 1992. V. 2. P. 867-870.
2. Fujisaki H. Prosody, Models, and Spontaneous Speech // Computing Prosody, ed. Campbell N., Norio H. New York: Springer, 1996. Ch.2. P. 27–42.
3. de Pijper J. Modelling British English Intonation. Foris: Dordrecht., 1983.
4. Taylor P. Analysis and synthesis of intonation using the Tilt model // J. Acoust. Soc. of America: 2000. V. 107. Is.3. P.1697-1714.
5. Ode C. Russian intonation: a perceptual description. Amsterdam: Rodopi, 1990.
6. Pavlova E., Pavlov Y., Sproat R., Shih Ch., van Santen J. Bell laboratories Russian text-to-speech system // 5-th European Conference on Speech Communication and Technology: proc. of Eurospeech'97. Rhodes:1997. P. 2451-2454.
7. Lobanov B. The Phonemophon Text-to-Speech System // 11-th International Congress of Phonetic Sciences: proc. of the ICPhS'87. Tallin: 1987. P. 61-64.
8. Lobanov B., Tsirulnik L., Zhadinets D., Karnevskaya E. Language- and speaker specific implementation of intonation contours in multilingual TTS synthesis // Speech Prosody: proceedings of the 3-rd International conference. Dresden: 2006. V. 2. P. 553-556.
9. Лобанов Б.М., Пьорковска Б., Рафалко Я., Цирульник Л.И., Шпилевский Э. Фонетико-акустическая база данных для многоязычного синтеза речи по тексту на славянских языках // "Компьютерная лингвистика и интеллектуальные

технологии”: труды междунар. конф. Диалог’2006. М.: 2006. – С. 357–363.
10. Lobanov B., Tsirulnik L. Phonetic-Acoustical Problems of Personal Voice Cloning by TTS // Speech and Computer: proc. of the IX International Conference SPECOM’2004. S.-Petersburg: Anatolya, 2004. – P. 17 – 21.

---

[1] *Работа выполнена при поддержке европейского фонда INTAS в рамках проекта «Разработка многоголосовой и многоязыковой системы синтеза и распознавания речи (языки: белорусский, польский, русский)» в соответствии с грантом INTAS № 04-77-7404*