Система подготовки нового голоса для системы синтеза «VitalVoice»

A voice building system for the hybrid VitalVoice russian TTS system

Продан А. И. (prodan@speechpro.com),
Таланов А. О. (andre@speechpro.com),
Чистиков П. Г. (chistikov@speechpro.com)

ООО «Центр речевых технологий», (Санкт-Петербург, Россия)

Рассматривается технология создания нового голоса заданного диктора для работы в системе синтеза VitalVoice. Описана система автоматизированной подготовки голоса, выбор текстового материала, особенности процесса записи речи, создание базы Unit Selection, настройка параметров подбора элементов.

1. Введение

Система автоматизированной подготовки голоса используется при подготовке речевых данных для синтезатора речи по тексту VitalVoice ООО «Центр речевых технологий» (ЦРТ) [4, 11, 12].

Существуют различные подходы к организации автоматического синтеза речи по тексту. К основным можно отнести синтез по правилам (формантный синтез), артикуляторный синтез, компилятивный синтез, синтез на основании статистических моделей (HMM-синтез). Синтез методом Unit Selection [2, 3] — один из видов компилятивного синтеза. Его отличительной особенностью является то, что синтезированная речь составляется не из базы специально записанных аллофонов, дифонов или других элементов, каждый из которых представлен единственным вариантом, а из элементов, взятых из естественных предложений, и для каждого элемента производится выбор наиболее подходящего кандидата из множества вариантов. Данная технология позволяет достичь очень высокой естественности синтезированной речи. В рамках работы по созданию новой системы синтеза русской речи, осуществляемой ЦРТ, создан синтезатор на основе использования технологии Unit Selection, совмещенной с аллофонным синтезом.

Характерной особенностью синтеза методом Unit Selection является его критическая зависимость от состава и полноты речевого корпуса. Качественный синтез определённым голосом возможен только на основе полного, сбалансированного и корректно размеченного речевого корпуса. В ЦРТ для разметки речевой базы для синтеза Unit Selection была разработана специальная многоуровневая система [13].

Задача добавления нового голоса, безусловно, является очень актуальной для любой системы синтеза речи. В особенности это актуально для синтеза методом Unit Selection, поскольку для этого метода это крайне ресурсоемкая задача. Именно поэтому она является предметом разных направлений исследований. С одной стороны, она интересна с точки зрения изучения тех или иных характеристик голоса и речи определённого диктора, которые влияют на качество синтезированной речи, с другой — с точки зрения задачи максимального приближения синтезированной речи по своим характеристикам к оригинальной речи самого диктора, то есть имитации. Отдельной интересной подзадачей является создание голоса по уже имеющемуся речевому материалу без участия диктора в записи (например, имитация голоса известных актёров).

Поэтому очень важно сделать процесс подготовки голоса по возможности максимально быстрым и удобным. Причём инструмент, помогающий в подготовке звуковой базы и настройке голоса, должен подходить как для специалиста (то есть подразумевает наличие ручных настроек всех процессов, их корректировки и т. п.), так и для человека, имеющего только самые общие представления о фонетике и синтезе речи (или получившего их из прикладываемой к программе документации и пошаговой инструкции), — то есть свести ручную корректировку и настройку к минимуму, в идеале к «нажатию одной кнопки».

Система подготовки нового голоса (СПГ) предназначена для автоматизации работы по созданию голоса для системы синтеза VitalVoice. В результате работы СПГ формируется установочный файл голоса, который работает с программой синтеза русской речи VitalVoice.

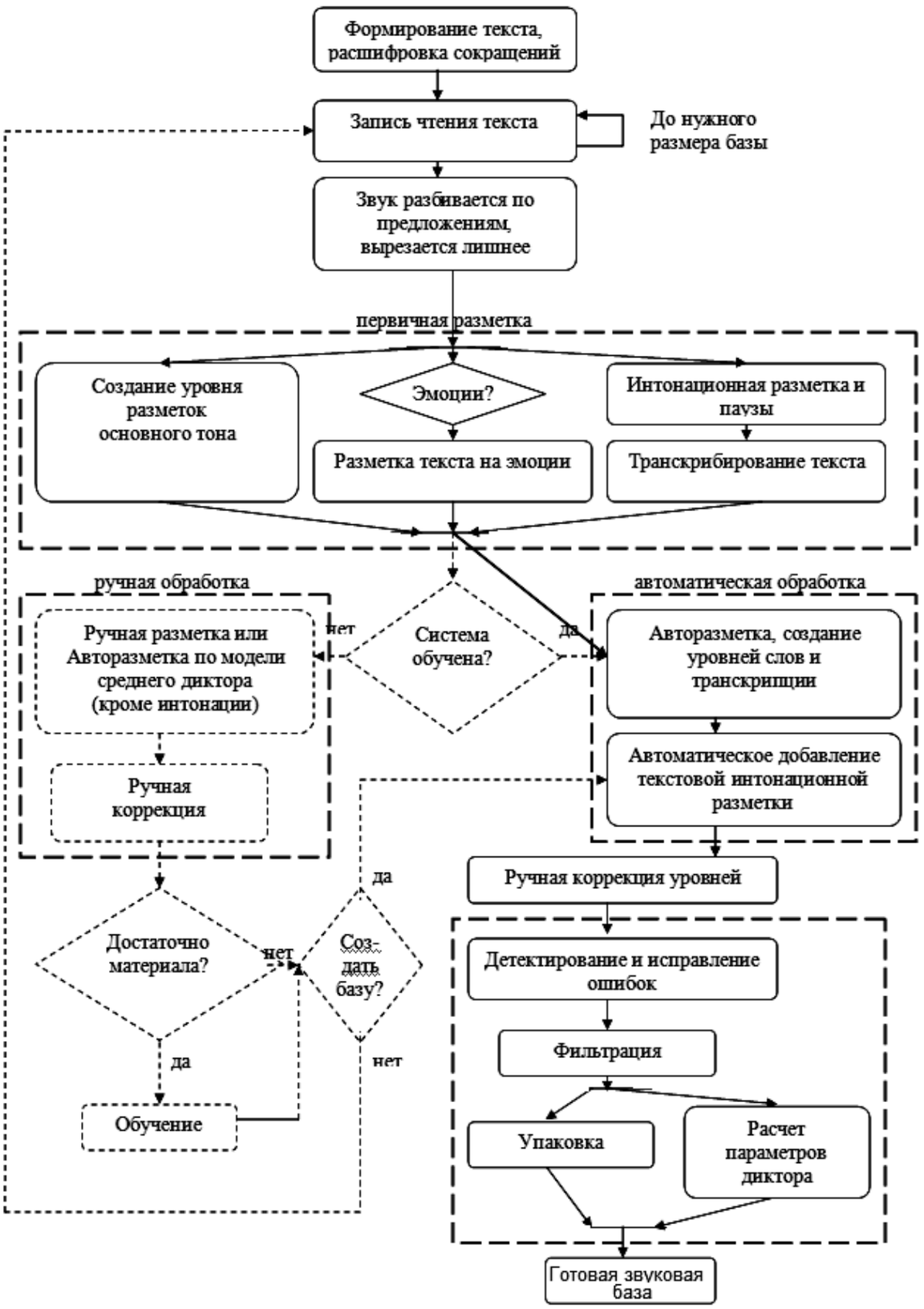

Рис. 1. Алгоритм создания звуковой базы

Система состоит из следующих частей:

- сама система подготовки голоса;
- звуковой редактор;
- транскриптор;
- программа автоматической разметки;
- программа автоматизированной проверки разметки базы;
- упаковка базы, создание установочного файла;
- расчёт и настройка параметров диктора;
- пошаговая инструкция по созданию и настройке голоса.

2. Основные шаги СПГ

Процесс создания звуковой базы представляет собой последовательность взаимосвязанных и при необходимости повторяющихся действий, объединённых в блоки, которые в разной степени поддаются автоматизации. На рис. 1 изображён алгоритм создания звуковой базы Unit Selection для синтеза Vital Voice.

Автоматизированное создание голоса происходит посредством выполнения последовательности шагов. На каждом шаге есть возможность перейти к следующему с использованием параметров по умолчанию (если не возникло ошибок, в противном случае — исправить только критические или все). Для исполнения каждого шага программа предложит совершить определённые действия. Стандартное окно программы изображено на рис. 2.

На каждом шаге предлагается его краткое описание и ссылка на подробную инструкцию. Рабочий экран содержит элементы управления, необходимые для выполнения действий данного этапа. При открытии каждого этапа кнопка «Дальше» остается недоступной до нажатия кнопки «Принять» и успешного прохождения всех проверок.

Флаг «автозапуск процессов» служит для установки автоматического запуска процессов обработки, предусмотренных на этапе, запуска проверки полученных результатов и перехода к следующему этапу в случае успеха.

2.1. Запись звуковых файлов

На этом этапе записывается чтение диктором выбранного текста. Если звуковые файлы уже записаны или необходимо создать голос по уже имеющимся записям, указывается текст, который был предложен диктору (если он есть, иначе текстовую расшифровку можно сделать на следующем этапе), и соответствующие ему звуковые файлы.

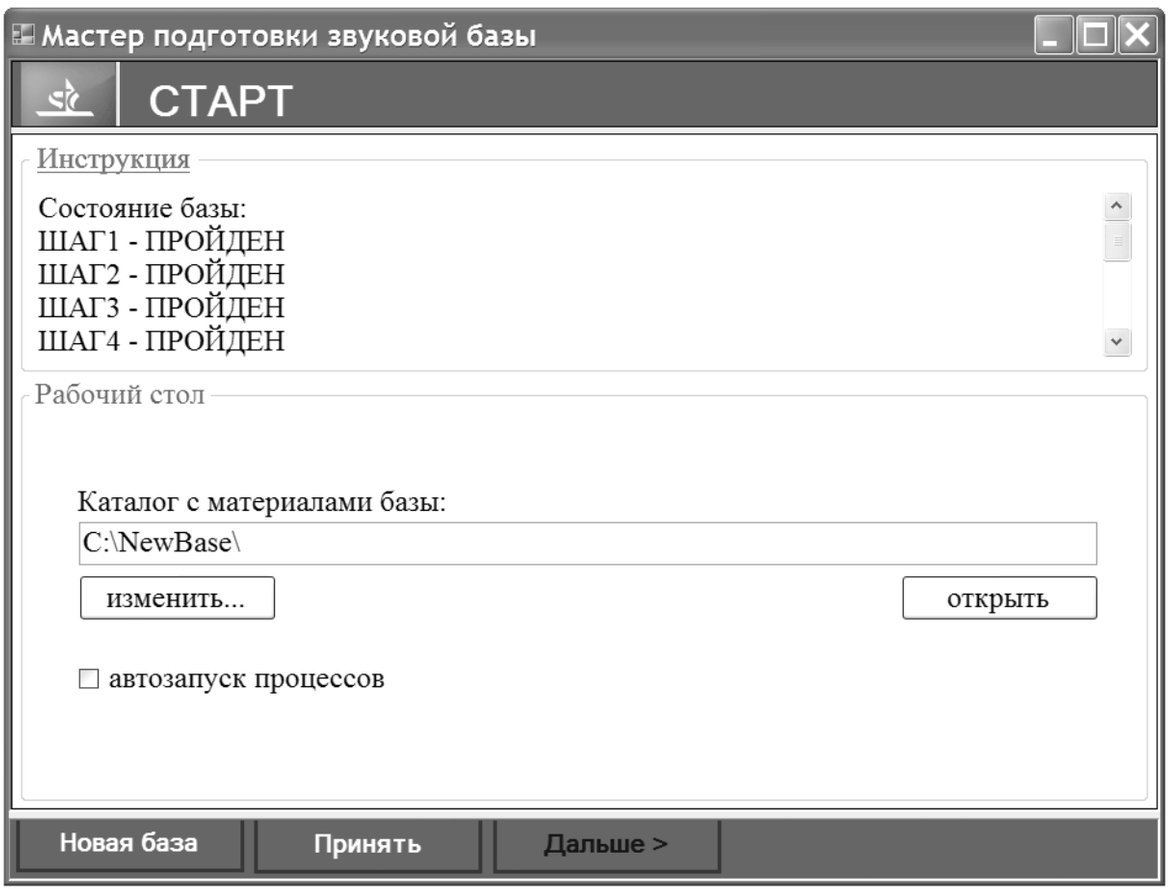

Рис. 2. Основная форма приложения, шаг «Старт»

2.1.1. Требования к тексту

Для создания качественного синтезированного голоса методом Unit Selection важно, чтобы на минимальном объёме звучащей речи было представлено максимальное количество элементов, по которым ведётся поиск, в системе синтеза VitalVoice это аллофоны в разных контекстах (т. е. трифоны) [1, 5]. Если минимальной единицей является не аллофон, а дифон, часть аллофона и т. д., важно, чтобы база была представительной с этой точки зрения. То же требование касается максимально возможной представительности интонационных конструкций. В «Центре речевых технологий» были разработаны специальные инструменты, позволяющие получить статистику наличия трифонов в тексте, а также подобрать фразы, содержащие аллофоны в редком контексте. Если в базе есть интонационная разметка, то по ней также можно получить статистику по использованию типов ИК [9, 10] и их параметрам. Кроме аллофонной и интонационной вариативности в текст также полезно добавить частотные слова, числа, фразы и, особенно если синтез будет использоваться для какой-то определённой задачи, названия и фразы, характерные для этой сферы (например, общение с системой голосового самообслуживания).

2.1.2. Условия записи

Запись диктора желательно производить в студии с хорошей шумоизоляцией — уровень сигнал/шум должен быть не меньше 30–40 дБ. Рекомендуется использовать конденсаторный микрофон. Запись наших баз производилась с частотой дискретизации 22 050 Гц и разрядностью 24 бита. При записи необходимо придерживаться уровня громкости −12 ... −10 дБ.

Чтение диктором текста

Текст лучше читать в нейтральной, сдержанной манере. Реализации различных интонационных конструкций не должны отличаться сильным разнообразием. Манера речи должна быть спокойной, доброжелательной. Темп и ритм чтения ровные. Между предложениями обязательны паузы. Лучше делать их больше естественных, чтобы можно было разбить запись по предложениям автоматически без ручной корректировки, настроив нужную длину паузы для разбиения. Внутри предложений частые паузы также излишни.

Желательно следить за чёткостью артикуляции всех звуков, но не делать её нарочитой и неестественной. Как правило, типичные для диктора темп и тембр устанавливаются после нескольких первых минут записи. При начале каждой новой сессии, особенно если прошлая запись происходила в другой день, надо сверять тембр и темп с предыдущими записями, полезно также дать послушать их диктору. В работе обязательны небольшие перерывы, а если диктор устал, запинается или у него изменился тембр голоса — лучше продолжить запись в другой день. Ни в коем случае нельзя записывать диктора с простудой или охрипшим голосом. При записи желательно постоянно работать с диктором, т. е. непрерывно следить за темпом и тембром и поправлять его.

При выборе готовых записей для создания голоса следует отдать предпочтение чтению перед спонтанной речью (интервью и т. п.), а также обратить внимание на зашумлённость записи и реверберацию. Те участки, где на фоне речи целевого диктора присутствует речь других людей, полностью удаляются.

2.2. Разбивка звуковых файлов по предложениям

На этом шаге звуковые файлы, содержащие озвученный текст, разбиваются на части, содержащие по одному предложению или синтагме, в соответствии с текстом (или просто по паузам). Некачественный материал (зашумлённые участки, оговорки, речь других людей и повторы, если они по каким-либо причинам не нужны) удаляется. Удаляются длинные паузы в начале итоговых звуковых файлов. Также на этом этапе удобно делать текстовую расшифровку записи, если диктор не читал заранее подготовленный текст.

Разбить исходный звуковой файл можно автоматически. Если разбивка прошла неудачно (слишком короткие/длинные куски), изменяются в соответствующую сторону параметры паузы, максимальная амплитуда и минимальная длина.

Необходимо синхронизировать предложения в тексте и пронумерованные звуковые файлы. Лишние файлы удаляются. Слишком короткие можно объединить.

Для удобства все предложения в тексте нумеруются. При переходе на следующий шаг автоматически проверяется наличие соответствующих номерам предложений звуковых файлов и наоборот, а также, исходя из среднего темпа чтения, при сильных отклонениях выдаются предупреждения о слишком коротких или длинных звуковых файлах, если текст намного короче или длиннее звука соответственно. На этом подготовительном этапе участие оператора для проверки и синхронизации звуковых файлов и текста необходимо, так как в ином случае будут неизбежны ошибки в разметке и составлении базы.

2.3. Разметка на периоды основного тона

На этом шаге для каждого файла, отобранного на предыдущем этапе, делается разметка на периоды основного тона. Это можно сделать с помощью программы WaveAssistant вручную или автоматически в режиме пакетной обработки.

Для более высокого качества синтеза и автоматической разметки на аллофоны желательно после автоматической разметки на периоды ОТ производить ручную правку тона для каждого файла.

2.4. Транскрибирование текста

Для каждого предложения автоматически создаётся файл с транскрипцией и текстовой расшифровкой в специальном формате для программы автоматической разметки. Правила транскрипции и правила лингвистической обработки большей частью задаются во внешних текстовых файлах, при необходимости (например, чтобы учесть какие-либо индивидуальные особенности диктора) в них можно оперативно внести изменения. Изначально вся обработка до стадии получения транскрипции совпадает с обработкой текста программой синтеза речи [8, 14], но вынесена в отдельное приложение, в свою очередь встроенное в систему подготовки нового голоса.

2.5. Авторазметка на аллофоны

Разметка на аллофоны — основной шаг при подготовке нового голоса.

Разметка может быть произведена как вручную, так и при помощи программы автоматической сегментации, разработанной в «Центре речевых технологий» на основе компонентов системы автоматического распознавания речи. При помощи неё создаются уровни идеальной, реальной транскрипции и уровень слов.

На уровне идеальной транскрипции расставляются метки аллофонов, полученные на шаге транскрибирования текста. На уровне реальной транскрипции выполняется пофонемное распознавание, при окончательном выборе аллофона учитывается его соответствие аллофону на уровне идеальной транскрипции. Возможные варианты аллофонов идеальной транскрипции или их пропуск задаются в подгружаемой из внешнего файла таблице соответствий.

Более точной является разметка, основанная на акустических моделях, построенных на основе ручной сегментации речи заданного диктора на аллофоны, но для этого требуется размеченная база голоса объёмом не менее полутора часов. Если такого объема ручной сегментации нет, то программа автоматической разметки будет использовать акустические модели, обученные по любым другим речевым базам (совпадающим по техническим характеристикам сигнала). При этом в процессе обработки других баз будут использованы только дикторы с похожими голосами. Тот же принцип программа авторазметки использует в случае нехватки звукового материала при построении акустических моделей редких трифонов.

2.6. Коррекция ошибок разметки

На этом этапе производится проверка полученной на прошлом шаге автоматической разметки. Для синтеза речи заданным голосом с высоким качеством звучания желательно выполнять проверку разметки каждого файла как вручную, так и автоматически. В систему автоматической проверки разметки звукового корпуса [13] были введены некоторые новые проверки, относящиеся в основном к анализу автоматической разметки: проверка аллофонов, выбивающихся по длительности, более жёсткая проверка соответствий реальной и идеальной транскрипции и т. д. Кроме этого, ряд ошибок, препятствующих сборке базы голоса, выделен в отдельный блок, причём переход на следующий шаг осуществляется только после их исправления. Остальные блоки ошибок отсортированы по степени важности, и решение об их исправлении выносится в зависимости от наличия времени и ресурсов.

2.7. Фильтрация звука

На этом шаге каждый звуковой файл базы фильтруется. Это делается для того, чтобы по возможности максимально избежать при стыковке эффекта реверберации, то есть в основном следов предшествующих гласных на глухих согласных и звонких смычных. Глухие согласные фильтруются ФВЧ с полосой задерживания от 1500 Гц, звонкие после гласных с большой амплитудой — ФНЧ с полосой задерживания от 450 Гц.

2.8. Упаковка базы, расчёт параметров диктора и создание установочного файла голоса

Все необходимые для упаковки файлы к этому этапу уже готовы. Упаковка базы производится автоматически. Во время сборки подсчитываются различные характеристики (средняя длительность, амплитуда аллофонов, средний основной тон диктора и т. д.), которые потом можно сразу использовать как параметры подбора аллофонов для диктора. Если была произведена интонационная разметка, можно также заменить стандартные настройки типичными для диктора.

Настройка параметров подбора аллофонов

Параметры подбора элементов [6, 7] для лучшего качества синтеза зависят как от характеристик речи диктора и точности разметки базы, так и от размера базы. По умолчанию выбираются параметры для средней базы диктора заданного пола.

Далее приводятся общие рекомендации по параметрам подбора элементов. Чем меньше размер

базы, тем больший вес следует поставить на соединение по тону и спектру, но меньше вес на неразрывность последовательности аллофонов. В базе небольшого объёма больше вероятность того, что аллофон в целевом контексте найден не будет, поэтому нужна большая свобода выбора неточного контекста. Но, с другой стороны, если разметка неточна, то есть границы аллофонов сильно смещены, этот вес снижать не рекомендуется. В целом, если более приоритетной является естественность голоса, следует увеличивать стоимость связи, а если важнее точная передача интонации и максимальная разборчивость в ущерб естественности — стоимость замены.

3. Выводы и перспективы

На данный момент в системе синтеза VitalVoice десять основных голосов (четыре мужских и шесть женских), созданных на базах разного размера — от полутора до восьми часов звучащей речи. Опыт создания голосов подтверждает, что сбалансированной базы размером в полтора-два часа достаточно для довольно качественного синтеза, но для максимальной естественности желательно увеличить объём до 6–7 часов в зависимости от характеристик диктора.

Автоматизированная система создания голоса на текущий момент была опробована на небольших объёмах речевого материала (менее получаса), но уже позволила получить практически важные результаты: с минимальной ручной корректировкой разметки достигнута почти полная разборчивость речи и практически стопроцентная узнаваемость исходного диктора.

Развитие системы предусматривает как развитие каждого отдельного компонента, так и улучшение их взаимодействия и интеграции в системе подготовки голоса. В будущем планируется уделить наибольшее внимание автоматической разметке, дальнейшей автоматизации настройки параметров подбора элементов в зависимости от характеристик речи диктора и размеров базы, а также удобству использования самой программы.

Литература

1. *Black A. W.* Perfect Synthesis for all of the people all of the time // Keynote, IEEE TTS Workshop Santa Monica, CA, 2002.
2. *Black A. W., Hunt A. J.* Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database // In Proceedings of ICASSP 96. Atlanta, Georgia, 1996. Vol. 1, pp. 373–376.
3. *Clark R. A. G., Richmond K., King S.* Multisyn: Open-domain unit selection for the Festival speech synthesis system // Speech Communication, 2007. Vol. 49, issue 4. P. 317–330.
4. *Oparin I., Talanov A.* Outline of a New Hybrid Russian TTS System // Proc. of the 12th International conference on Speech and Computer, SPECOM 2007, Moscow, Russia, 2007. P. 603–608.
5. *Tatham M, Morton K.* Developments in Speech Synthesis // John Wiley & Sons Ltd, 2005.
6. *Vepa J., King S.* Subjective evaluation of join cost functions used in unit selection speech synthesis // In Proceedings of the International Conference on Speech and Language Processing 2004. Jeju, Korea, 2004. P. 1181–1184.
7. *Vepa J.* Join Cost for Unit Selection Speech Synthesis // University of Edinburgh, 2004.
8. *Аничкин И. М., Чистиков П. Г.* Формализация правил автоматического снятия омонимии в системе синтеза речи по тексту // Материалы XXXVIII Международной филологической конференции.
9. *Брызгунова Е. А.* Интонация // Русская грамматика. М.: 1980.
10. *Вольская Н. Б., Скрелин П. А.* Моделирование интонации для синтеза речи по тексту // Уфа: 1998.
11. *Киселёв В. В., Чижденко В. А., Таланов А. О., Опарин И. В.* Архитектура системы синтеза русской речи по тексту нового поколения // Материалы XXXVII Международной филологической конференции.
12. *Корольков Е. А., Главатских И. А., Таланов А. О., Киселев В. В., Опарин И. В.* Синтез естественной русской речи при помощи метода Unit Selection // Материалы XXXVII Международной филологической конференции.
13. *Продан А. И., Корольков Е. А., Опарин И. В., Таланов А. О.* Особенности использования многоуровневой разметки звукового корпуса Unit Selection в системе гибридного синтеза «Живой голос» // Материалы международной конференции Диалог 2009.
14. *Хомицевич О. Г., Рыбин С. В., Таланов А. О., Опарин И. В.* Автоматическое определение места ударения в незнакомых словах в системе синтеза речи // Материалы XXXVII Международной филологической конференции.