УДК 004.934

В.А. Захарьев, А.А. Петровский, Б.М. Лобанов

СИСТЕМА СИНТЕЗА РЕЧИ ПО ТЕКСТУ С ВОЗМОЖНОСТЬЮ НАСТРОЙКИ НА ГОЛОС ЦЕЛЕВОГО ДИКТОРА

Захарьев В.А., Петровский А.А., Лобанов Б.М. **Система синтеза речи по тексту с возможностью настройки на голос целевого диктора.**

Аннотация. В статье представлена схема построения мультиголосового синтезатора речи, основанная на использовании синергетического эффекта от интеграции системы синтеза речи по тексту и конверсии голоса. Такая организация даёт возможность одновременно выполнять действия синтеза и модификации речевого сигнала на основе комплексного подхода, позволяя снизить количество ошибок и артефактов, которые влияют на качество речевого сигнала. Применение данного подхода обеспечивает реализацию функции настройки синтезатора речи на голос целевого диктора без существенных затрат трудоёмкости на обучение речевой базы данных, для добавления новых голосов.

Ключевые слова: синтез речи по тексту, конверсия голоса, функция конверсии голоса

Zahariev V.A., Petrovsky A.A., Lobanov B.M. **Text to speech synthesis system with the target speaker voice customization capability.**

Abstract. A scheme of constructing multi-voice speech synthesizer based on the use of the synergies of integration of the text to speech and voice conversion systems are presented in this article. Such organization of the system allows simultaneous synthesis and modification actions in speech signal, based on an integrated approach to its treatment, significantly reducing the number of errors and artifacts that affect the resulting quality. Applying this approach let to implement a function for multivoice speech synthesizer without significant labor costs for training speech database to add new voices.

Keywords: text to speech synthesis, voice conversion, voice conversion function.

1. Введение. На данном этапе развития систем синтеза речи по тексту (ССРТ) ставится вопрос уже не столько об обеспечении хороших уровней основных показателей систем этого класса, например, разборчивости синтезируемой речи, сколько о более сложных характеристиках, таких как, натуральность синтезируемой речи, поддержка множества языков и различных голосов дикторов. Последний аспект — создание мультиголосовых систем синтеза речи по тексту — требует особого подхода и внимания, поскольку в существующих ССРТ перенастройка системы на нового диктора требует больших материальных и временных затрат от разработчиков системы [1].

В данной статье предлагается рассмотреть возможность решения задачи построения мультиголосового синтезатора речи с использованием технологии конверсии голоса, на основе системы синтеза речи по тексту, разработанной авторами.

2. Архитектура мультиголосовой ССРТ. Система конверсии голоса (КГ) базируется на технологии обработки речевого сигнала,

позволяющей реализовать процесс трансформации параметров голоса, характеризующих речь исходного диктора (ИД), в параметры целевого диктора (ЦД). Объектами конверсии голоса, как технологии обработки сигналов, являются стабильные во времени свойства говорящего, проявляющиеся в речевом сигнале через изменение его акустических параметров [2–4].

Вполне очевидной является идея применения данной технологии в синтезаторах речи по тексту для решения задачи добавления функций мультиголосового синтеза. В простейшем случае системы СРТ и КГ являются полностью независимыми: выходной сигнал, поступающий от ССРТ используется в качестве входного сигнала для системы КГ [5,6]. К положительным сторонам данного подхода можно отнести устойчивость и универсальность данной архитектуры, в её рамках легко может быть заменена любая составляющая, без нарушения общей работоспособности всей системы. Например, при смене модели синтеза или конверсии на более совершенную.

Однако, реализация данного подход может привести к существенной потере качества т.к. изменение просодических характеристик речи (частоты основного тона и длительности звуков) осуществляется дважды: первый раз просодическим процессором синтезатора речи по тексту и второй раз системой конверсии голоса.

Второй подход [7], развиваемый авторами в данной работе, подразумевает тесное объединение двух типов систем, путём включения элементов системы конверсии голоса в состав синтезатора речи по тексту, и использования ожидаемого синергетического эффекта от такого варианта построения интегрированной архитектуры системы. Реализация данного подхода достигается путём внедрения модуля конверсии голоса в блок акустического процессора и рационального разделения задач конверсии параметров голоса между двумя видами систем с учётом особенностей этапов обработки информации в каждой из них.

Предлагаемая архитектура мультиголосовой ССРТ представлена на рисунке 1. В ней были учтены следующие важные нюансы. Во-первых, аспекты конверсии голоса учитываются при выборе единиц компиляции. Во-вторых, все алгоритмы преобразования и конверсии (спектральные и просодические) выполняются единым блоком, это означает, что характеристики сигнала модифицируются только один раз. В-третьих, конкатенация и реконструкция синтезированного речевого сигнала выполняются после конверсии голоса исходного диктора в голос целевого.

3. Модель представления речевого сигнала. В качестве основного метода для выполнения операций анализа-синтеза и параметризации речевого сигнала, мы предлагаем использовать модель сигнала на базе адаптивной интерполяции взвешенного спектра — Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrum (STRAIGHT), первоначально разработанную и построенную профессором Хайдеки Кавахара для исследования механизмов восприятия речи человека, основанную на её вокодерном представлении. STRAIGHT использует процедуры, которые могут быть сгруппированы в три подсистемы: экстрактор информации о сигнале возбуждения, экстрактор сглаженного частотно-временного представления сигнала, и движок синтеза, состоящий из источника возбуждения и системы фильтров с изменяющимися во времени характеристиками [8]. Результаты работы двух первых блоков вычисления параметров речевого сигнала представлены на рисунке 2.

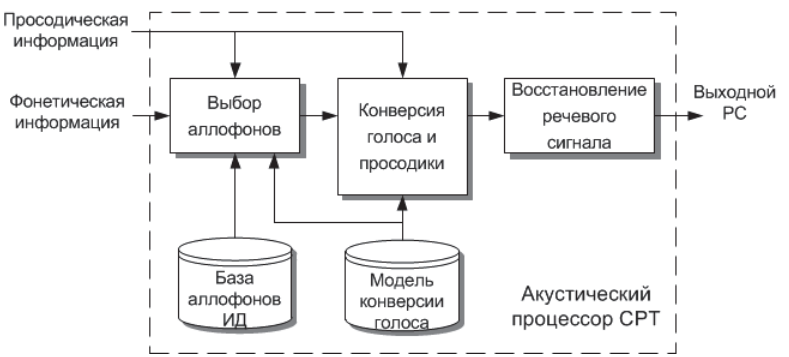
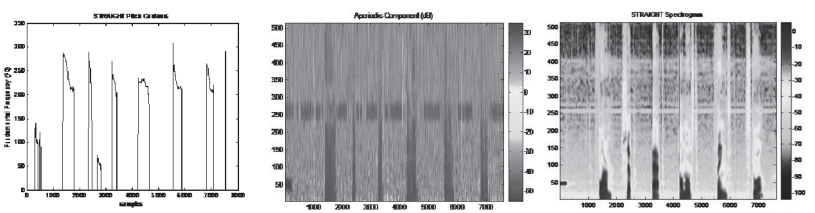


Рис. 1. Архитектура мультиголосового СРТ



а) б) в)

Рис. 2. Параметры сигнала, полученные на основе модели STRAIGHT: а — контур ЧОТ; б — спектрограмма апериодической компоненты; в — спектрограмма периодической компоненты

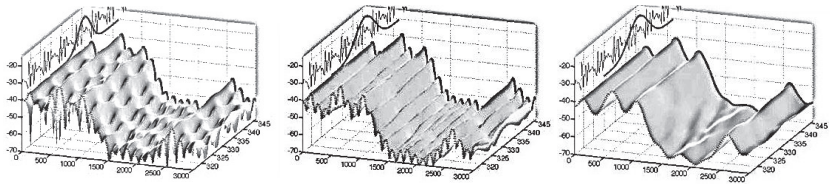
Рассмотрим основные особенности этапов работы подсистем данной модели. Для того чтобы иметь возможность адекватно описывать, трансформировать, и восстанавливать речь, очень важной является возможность точного определения контура частоты основного тона (ЧОТ), который бы не имел никаких следов интерференций, вызванных неправильным размером и формой окна анализа сигнала.

На первом этапе подсистема извлечения сигнала возбуждения определяет частоту основного тона как мгновенную частоту фундаментальной компоненты сигнала. Она определяется как отображение фиксированных точек в область мгновенной частоты из коэффициентов оконного преобразования Фурье.

Процедура, основанная на использовании нормализованной автокорреляционной функции, была интегрирована с предыдущей процедурой поиска мгновенной частоты для уменьшения числа ошибок определения ЧОТ. Представленный способ экстракции контура ЧОТ предполагает, что сигнал имеет близкую к гармонической структуру. Тем не менее, практически всегда существуют отклонения от строгой периодичности, которые вводят в сигнал дополнительные негармонические компоненты. Таким образом, существует возможность найти меру апериодичности, принимая за её величину отношение энергии составляющих на частотах, нарушающих гармонический строй, к полной энергии сигнала. Апериодическая компонента $A(w,t)$ оценивается как остаток между исходным сигналом и его периодической составляющей [9].

Второй этап, экстракция сглаженного частотно-временного представления (спектрограммы) сигнала является “ядром” данной модели. В случае с простыми вокодерными системами периодическое возбуждение фильтров является эффективной стратегией для передачи резонансной информации и увеличения отношения сигнал-шум. Однако, это периодическое возбуждение вносит дополнительные интерференции как во временной так и в частотной областях, как это показано в левой части рисунка 3. Поэтому возникает необходимость в реконструкции гладкой поверхности частотно-временного плана из искаженного представления.

Следующие два шага последовательно выполняются для решения данной проблемы. Первым шагом является сокращение оставшейся временной периодичности из-за фазовой интерференции между соседними гармоническими составляющими.



а) б) в)

Рис. 3. Процесс экстракции сглаженного частотно-временного представления: а — спектр оригинального сигнала; б — спектр сигнала с устранённой темпоральной волатильностью; в — спектр сигнала в модели STRAIGHT

Это достигается за счёт применения, помимо основного окна анализа, дополнительного комплементарного временного окна. Комплементарное окно $w_c(t)$ к окну анализа $w(t)$ определяется согласно выражению:

$$w_c(t) = w(t)\sin\frac{\pi t}{T_0},$$

где T_0 — период основного тона сигнала. Спектр комплементарного сигнала $P_c(w,t)$, вычисленный с помощью комплементарного окна, имеет пики там, где спектр оригинального сигнала $P(w,t)$, вычисленный с помощью обычного окна (например, на рисунке 3а с помощью окна Гаусса), имеет провалы. Далее рассчитывается спектр сигнала с устранённой темпоральной волатильностью $P_R(w,t)$, который определяется путем вычисления взвешенной суммы двух типов спектров $P(w,t)$ и $P_c(w,t)$ с весовым коэффициентом ξ, отыскиваемым в процессе решения задачи оптимизации [10]. Рисунок 3б показывает спектр сигнала $P_R(w,t)$, полученного при оптимальном значении весового коэффициента. Легко видеть, что на данном рисунке отсутствуют провалы в частотно-временном плане, наличие которых очевидно на левом графике, где представлен спектр оригинального сигнала $P(w,t)$.

Вторым шагом является обратная фильтрация на основе сплайнов. Она необходима для того, чтобы удалить оставшиеся паразитные периодичности в частотной области, одновременно с сохранением распределения первоначальных спектральных уровней по частотам

гармоник. STRAIGHT реализует операцию сглаживания с использованием базисной функции на основе Б-сплайна второго порядка. Поскольку спектр сигнала, вычисленный с использованием комплементарного набора окон, не представляет собой линейный спектр, а также дает некоторое размытие энергии сигнала по частоте, используется соответствующее сглаживающее ядро h_Ω для восстановления "смазанных" значений на частотах гармоник. Использование такой весовой функции h_Ω приводит к тому, что мы получаем спектр, менее чувствительный к ошибкам определения ЧОТ, а также с хорошей локализацией энергии по частоте, практически соответствующий линейному спектру. Следующее выражение является расчётным для определения восстановленной спектрограммы $P_{ST}(w,t)$ (спектрограммы STRAIGHT):

$$P_{ST}(w,t) = \left(\int h_\Omega(\lambda / w_0) P_R^\gamma(w-\lambda,t) d\lambda \right)^{1/\gamma},$$

где w_0 — циклическая частота от F_0. Параметр γ определяет нелинейность характеристики и был выбран равным 0,3 на основе экспериментов. Форма весовой функции h_Ω вычисляется путём решения системы линейных уравнений, полученных из $w(t)$, $w_c(t)$, ξ и γ [10]. В крайней правой части рисунка 3в показан спектр сигнала в модели STRAIGHT для гласного звука /а/, произнесённого диктором мужского рода. Необходимо обратить внимание, что в процессе выполнения вышеописанных шагов (при переходе от графика к графику слева направо) постепенно снижается степень проявления интерференции и различных искажений частотно-временного плана, с одновременным сохранением детализации спектральной картины на частотах гармонических составляющих сигнала.

После выполнения всех необходимых преобразований для реконструкции сигнала, аппарату синтеза необходим соответствующий набор параметров (спектр STRAIGHT $P_{ST}(w,t)$, карты апериодичности $A(w,t)$, контура ЧОТ F_0 с отметками о вокализованности $f_0(t)$. Все эти параметры имеют строго определённый физический смысл, связанный с процессом речеобразования, и в тоже время позволяют проводить с ними разнообразные и независимые манипуляции, без опасения внесения различного рода несоответствий между параметрами.

4. Модель конверсии голоса. Для настройки мультиголосового СРТ на целевого диктора необходимо предварительное обучения модели конверсии голоса. Данный этап состоит из ряда шагов:

— Подготовки априорной речевой информации об ИД и ЦД в виде набора фонограмм. Вся необходимая информация об ИД, позволяющая сгенерировать произвольный набор обучающих аудиозаписей фраз, уже имеется в акустической базе аллофонов ССРТ. Для ЦД возможны варианты в зависимости от выбранного пользователем способа обучения системы. В случае текстозависимого способа, обучающая выборка аудиозаписей формируется путём озвучивания фонетически сбалансированного текста, предлагаемого пользователю модулем обучения системы. В случае текстонезависимого способа, пользователь может сам вводить в систему любой текст, параллельно произнося его в микрофон, или загрузить произвольный набор фонограмм и их стенограмм. Например, фрагменты аудиокниги и соответствующие им фрагменты текста. В случае невыполнения условия фонетической сбалансированности, предоставляемого пользователем речевого и текстового материала, система автоматически сформирует пользователю запрос на предоставление дополнительной информации [11,12].

— Проведения анализа и параметризации данного набора фонограмм с помощью модели STRAIGHT и получения последовательностей векторов параметров характеризующих голос ИД и ЦД.

— Определения соответствия между этими последовательностями векторов параметров с помощью метода динамического программирования, при текстозависимом обучении, или аппарата скрытых Марковских моделях, для текстонезависимого способа обучения [13,14]. Объединения данных двух отмасштабированных и выровненных последовательней в общую матрицу, фактически, представляющую собой совместное пространство векторов параметров дикторов.

— Кластеризации получившегося совместного пространства векторов параметров с помощью статистических методов, в частности методов мягкой классификации на основе модели Гауссовых смесей (ГС). Характеристики данных классов или кластеров, выступают в дальнейшем в качестве параметров функции конверсии, которая является ядром модели и выполняет сам процесс трансформации признаков голоса.

Подробно шаги этапа обучения описаны в работах авторов и литературе [15], а далее рассмотрим особенности реализации функции

конверсии как наиболее важного элемента использующегося на этапе выполнения системы.

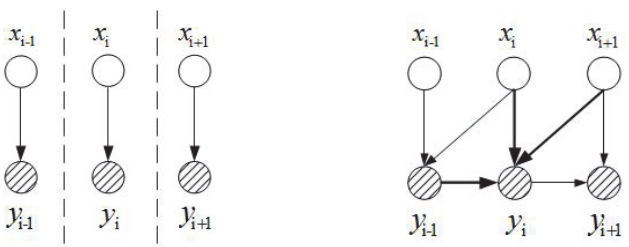
За основу функции конверсии, являющейся ядром системы, было принято выражение на основе линейной регрессионной модели первого порядка, которая использует в качестве своих коэффициентов регрессии параметры ГС, полученные на этапе обучения [16]. Эта функция доказала свою эффективность, особенно по сравнению с подходами, основанными на жесткой кластеризации пространства акустических параметров дикторов [17]. Однако, детальный её анализ выявил ряд недостатков. Сложность выбора порядка модели ГС, с ростом которой увеличивается вычислительная сложность, а также быстро деградирует выходное качество речевого сигнала в результате переобучения и чрезмерного усреднения параметров моделью. Кроме того, статистические отношения рассматриваются исключительно между одной парой векторов параметров ИД и ЦД в каждый i-ый момент времени (рис. 4а). В данной работе предлагается расширенный вариант функции конверсии, основанный на модели ГС и свойствах эргодичности сигнала. Это позволяет учитывать не только пространственные, но и временные корреляционные связи между соседними векторами параметров ИД и ЦД, тем самым придавая процедуре конверсии свойства Марковского процесса (рис 4б):

$$y_i = \sum_{q=1}^{Q} p_q(x_i, y_{i-1}, x_{i+1})[\mu_q + \Phi_q \bar{x}_i^q + \Psi_q \bar{y}_{i-1}^q + \Omega_q \bar{x}_{i+1}^q],$$

где i — порядковый номер фрейма сигнала $i = 1,...,T$, q — индекс смеси в составе модели ГС $q = 1,..., Q$, $x = [x_1, x_2,..., x_T]^T$, $y = [y_1, y_2,..., y_T]^T$ — последовательность векторов параметров ИД и ЦД $x_i, y_i \in \mathrm{R}^{1\times p}$, p — размерность вектора параметров сигнала, $p_q(x_i, y_{i-1}, x_{i+1})$ — апостериорная вероятность того, что входной вектор x_i и вектора y_{i-1} и x_{i+1} принадлежат к q-ой компоненте ГС, $\mu_q = [\mu_1, \mu_2,..., \mu_Q]^T$ — вектор математических ожиданий для каждой компоненты смеси ЦД $\mu_q \in \mathrm{R}^{1\times p}$, $\{\Phi, \Psi, \Omega\}$ — матрицы регрессионных коэффициентов для всех компонент смеси независимой перемен-

ной x_i, а также предикторов x_{i+1} и y_{i-1} где $\Phi_q, \Psi_q, \Omega_q \in \mathrm{R}^{p \times p}$. Тогда задача нахождения неизвестных параметров $\{\mu, \Phi, \Psi, \Omega\}$ формулируется как задача оптимизации, решение которой можно найти по методу наименьших квадратов [18].

Вторым улучшением модели конверсии стало применение в её рамках метода спектрального взвешивания, имеющего глубинную связь с физической природой речевого сигнала. Это позволяет на базе имеющейся информации о входном векторе параметров x, производить конверсию спектральной огибающей с последующим её взвешиванием специальной масштабирующей функцией в виде линейной комбинации Q базисных функций $\{W_q(f)\}$, используя апостериорные вероятности p_q как коэффициенты в данной линейной комбинации.



а) б)

Рис. 4. Отношения между парами векторов в обучающей последовательности: а — независимая модель; б — Марковский процесс

Таким образом, осуществляется обработка параметров сигнала после процедуры конверсии с целью решения проблемы чрезмерного усреднения огибающей в результате конверсии или возможного переобучения системы.

Для преобразования просодических характеристик голоса диктора предлагается применить новый метод, использующий параметрическое представление контура частоты основного тона. Он основан на поиске особых точек контура частоты, в соответствии с методикой предложенной Паттерсоном [19]. Далее, используя кусочно-линейную функцию конверсии ЧОТ специального вида, которая имеет различные свойства в различных областях значений ЧОТ, выполняется схема преобразования просодических единиц [20]. Эта манипуляция обеспечивает лучшее качество конверсии просодических элементов без су-

щественного увеличения вычислительных затрат, по сравнению со стандартной процедурой нормализации статистических оценок ЧОТ в классических алгоритмах конверсии.

Блок реконструкции формы сигнала производит восстановление речевого сигнала из набора параметров на основе модели STRAIGHT. Таким образом, по входному тексту на выходе системы, мы получаем реконструированную фразу озвученную голосом целевого диктора, что позволяет с помощью представленных методов решить поставленную задачу мультиголосового синтеза речи по тексту.

5. Экспериментальная часть. В данном разделе статьи представлена экспериментальная оценка качества работы мультиголосового синтезатора речи по тексту. Основная идея эксперимента заключалась в определении того, насколько речь синтезируемая голосом целевого диктора, с использованием конверсии, соответствует голосу реального целевого диктора, или же в речевом сигнале всё еще присутствуют черты исходного диктора. Такой вывод можно сделать, основываясь на степени близости между речевым сигналом синтезируемым системой мультиголосового синтеза и оригинальных аудиозаписей тестовых фраз для целевых дикторов с одной стороны, и озвученных исходным диктором (т.е. синтезированных системой, с уже имеющимся в ней диктором) с другой. Степень близости, в свою очередь мы предлагаем определять метрикой на базе мел-кепстрального расстояния, как одной из наиболее адекватных объективных оценок определения акустического расстояния между векторами [21]. Более подробно методика эксперимента изложена далее.

Из специального фонетически сбалансированного текста для создания макси-набора аллофонов, диаллофонов аллослогов русской речи [1] сформированы один обучающий (30 предложений), а также пять тестовых (по 4 предложения) наборов текстов различных фраз. На основе данных наборов обучающих и тестовых текстов фраз были записаны их фонограммы для четырёх различных целевых дикторов: двух женщин (ДЖ1, ДЖ2) и двух мужчин (ДМ1, ДМ2). Полученные аудиофайлы были закодированы в формате wav, с частотой дискретизации 16000кГц и разрядностью сетки квантования в 16 бит. Далее, на основании имеющихся фонограмм и стенограмм тренировочной выборки фраз было проведено обучение и настройка системы мультиголосовой ССРТ в текстозависимом режиме, на каждого из четырёх целевых дикторов. Затем, для каждого тестового набора фраз, уже непосредственно в режиме функционирования мультиголосовой системы, был осуществлён их синтез, вначале голосом исходного диктора,

имеющегося в синтезаторе, а затем голосами целевых дикторов с использованием конверсии голоса. Далее, над результирующими аудиоданными, полученными на тестовом наборе фраз, т.е. оригинальными и синтезированными фонограммами для целевых дикторов, а также фонограммами синтезированными голосом исходным диктора системы, выполнялся мел-кепстральный анализ сигнала, с размерностью вектора кепстральных коэффициентов, для одного фрейма, равным двадцати. В параметрическом представлении наборы фраз были объединены в последовательности векторов параметров мел-кепстральных коэффициентов. А затем для каждого диктора по каждому набору фраз рассчитывались две оценки близости, синтезированного системой сигнала, к исходному диктору $\varepsilon_{ИД}$, и вторая — к целевому $\varepsilon_{ЦД}$. Они рассчитывались как средняя квадратичная ошибка между векторами кепстральных коэффициентов для всей последовательности векторов одного набора фраз, согласно выражению [21]:

$$\varepsilon_{ИД} = \frac{1}{N}\sum_{n=1}^{N} \| c_n^{synt} - c_n^{src} \|^2,$$

$$\varepsilon_{ЦД} = \frac{1}{N}\sum_{n=1}^{N} \| c_n^{synt} - c_n^{trg} \|^2,$$

где N — количество векторов кепстральных коэффициентов в одном наборе фонограмм параметризированных аудиозаписей тестовых фраз, c_n^{src} — n-ый вектор кепстральных коэффициентов параметризированной последовательности фонограмм, одного набора, синтезированных голосом исходного диктора (имеющегося в синтезаторе), c_n^{synt} — n-ый вектор кепстральных коэффициентов параметризированной последовательности фонограмм, одного набора, синтезированных голосом текущего диктора (не которого настроен синтезатор в текущий момент), c_n^{trg} — n-ый вектор кепстральных коэффициентов параметризированной последовательности фонограмм, одного набора, озвученных голосом целевого диктора (реального диктора, на которого была настроена система). Результирующие оценки представлены в таблице.

Таблица. Результаты экспериментальной оценки качества работы системы

Диктор	Набор тестовых фраз									
	I		II		III		IV		V	
	$\varepsilon_{ид}$	$\varepsilon_{цд}$	$\varepsilon_{ид}$	$\varepsilon_{цд}$	$\varepsilon_{ид}$	$\varepsilon_{цд}$	$\varepsilon_{ид}$	$\varepsilon_{цд}$	$\varepsilon_{ид}$	$\varepsilon_{цд}$
ДЖ1	0,89	0,62	0,82	0,76	0,90	0,67	0,87	0,63	0,80	0,59
ДЖ2	0,99	0,78	0,96	0,91	0,99	0,78	0,98	0,72	0,91	0,68
ДМ1	0,75	0,56	0,71	0,57	0,78	0,57	0,78	0,52	0,70	0,50
ДМ2	0,73	0,54	0,74	0,63	0,80	0,62	0,73	0,49	0,76	0,44
Среднее:	0,84	0,63	0,81	0,72	0,87	0,66	0,84	0,59	0,79	0,55

Исходя из таблицы, общие среднеквадратические ошибки по всем наборам фраз составляют $\varepsilon_{ИД} = 0{,}84$ для голосов исходного и синтезированного системой диктора, а также $\varepsilon_{ЦД} = 0{,}63$ для синтезированного и целевого дикторов. По результатам экспериментов можно наблюдать следующую картину: речевой сигнал, генерируемый системой мультиголосового синтеза речи, настроенной на целевого диктора, оказывается в параметрическом представлении на тридцать процентов ближе к речевому сигналу реального целевого диктора, чем речь, произведённая этой же системой, голосом имеющегося в ней исходного диктора. Это следствие, в свою очередь, позволяет сделать вывод о том, что речь генерируемая системой мультиголосового синтеза будет, с большей степенью вероятности, воспринята слушателем как произнесённая голосом целевого диктора. Данный результат подтверждает возможность создания системы синтеза речи по тексту с возможностью настройки на голос целевого на базе предложенной архитектуры системы, а также методов и моделей её реализации.

6. Заключение. В работе изложены новые подходы и принципы создания многоголосого синтезатора речи по тексту с использованием технологии конверсии голоса. Представлено обоснование выбора рациональной архитектуры системы на базе интерактивного подхода между двумя типами систем. Предлагаемая схема, на основе синергетического эффекта от интеграции двух типов систем, позволяет в полной мере использовать полезные свойства обоих, и решает проблему создания многоголосого СРТ с улучшенными показателями качества конверсии. Расширенная регрессионная функция конверсии применяется для повышения точности преобразования векторов параметров, а метод спектрального взвешивания — для решения проблемы чрезмерного усреднения спектральной огибающей. Всё это в итоге повышает качественные характеристики результирующего восстановленного сигнала, сохраняя вместе с тем высокие уровни узнаваемости голоса целевого диктора для преобразованной речи.

Литература

1. *Лобанов Б. М., Цирульник Л. И.* Компьютерный синтез и клонирование речи // Минск: Белорусская наука, 2008. 344 с.
2. *Abe M., Nakamura S., Shikano K.* Voice conversion through vector quantization // Proc. of International Conference on Acoustics, Speech and Signal Processing. New York, 1988. pp. 655–658.
3. *Valbret H., Moulines E., Tubach J.P.* Voice transformation using PSOLA technique // Proc. of International Conference on Acoustics, Speech and Signal Processing. 1992. vol. 1. pp. 145–148.
4. *Moulines E., Sagisaka Y.* Voice conversion: State of the art and perspectives // Speech Communication. 1995. pp. 125–224.
5. *Kain A., Macon M. W.* Text-to-speech voice adaptation from sparse training data // Proc. of International Conference on Spoken Language Processing. 1998. pp. 2847 – 2850.
6. *Sundermann D., Hoge H., Bonafonte A.* Text-independent voice conversion based on unit selection // Proc. of International Conference on Acoustics, Speech and Signal Processing. 2006. vol. 1.
7. *Azarov E, Petrovsky A.A, Lobanov B, Tsirulnik L.* Text-to-speech system with acoustic processor based on the instantaneous harmonic analysis // SPECOM. 2009. pp. 414–418.
8. *Kawahara H., Morise M.* Technical foundations of tandem-straight, a speech analysis, modification and synthesis framework // SADHANA. Academy Proceedings in Engineering Sciences, 2011. pp. 713–722.
9. *Kawahara H., Katayose H., Cheveigne A.* Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity // Proc. Eurospeech'99. 1999. pp. 2781–2784.
10. *Kawahara H., Masuda I., Cheveigne A.* Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous–frequency–based F0 extraction // Speech Communication. 2004. pp. 187–207.
11. *Arslan L.* Speaker transformation algorithm using segmental codebooks // Speech Communication. 1999. vol. 28, no. 3. P. 211–226.
12. *Narendranath M*., *Murthy H., Rajendran S., Yegnanarayana N.* Transformation of formants for voice conversion using artificial neural networks // Speech Communication. 1995. vol. 16, no. 2. pp. 207–216.
13. *Рабинер Л. Р.* Скрытые марковские модели и их применение в избранных приложениях при распознавании речи: Обзор // Труды института инженеров по электронике и радиоэлектронике. 1989. Т. 77, № 2. С. 86–120.
14. *Азаров И. С., Петровский А. А.* Система конверсии голоса в реальном масштабе времени с текстонезависимым обучением на основе гибридного параметрического описания речевых сигналов // Цифровая обработка сигналов. 2012. № 2. С. 15–23.
15. Анализаторы речевых и звуковых сигналов: методы, алгоритмы и практика (с MATLAB примерами) / под редакцией А. А. Петровского. // Минск: Бестпринт, 2009. 456 с.
16. *Stylianou, Y., Cappe O., Moulines E.* Statistical methods for voice quality transformation // Proc. of European Conference on Speech Communication and Technology. Madrid, 1995. pp. 447–450.
17. *Павловец А.С., Лившиц М.З. , Личачев Д. С., Петровский А. А.* Конверсия голоса с использованием модели сепарации речевого сигнала на компоненты “гармоники + шум” и переходные фреймы // Речевые технологии. 2008. №4. С. 37–50.
18. *Stylianou Y., Cappe O. , Moulines E.* Continuous probabilistic transform for voice conversion // Proc. of International Conference on Acoustics, Speech and Signal Processing. 1998. pp. 2451–2455.
19. *Patterson D.* A linguistic approach to pitch range modeling // PhD dissertation. Scotland: University of Edinburgh, 2000. 201 p.

20. *Захарьев В.А., Петровский А.А.* Конверсия просодических характеристик диктора на основе методов параметризации контура частоты основного тона. Доклады БГУИР // Минск, 2013. С. 39–46.
21. *Hu Y., Loizou P.C.* Evaluation of objective quality measures for speech enhancement // IEEE Transactions on Audio, Speech & Language Processing. 2008. vol. 16, no. 1. pp. 229–238.

References

1. Lobanov B. M., Cirul'nik L. I. *Komp'juternyj sintez i klonirovanie rechi* [Computer speech synthesis and cloning] Minsk: Belorusskaja nauka, 2008. 344 p. (In Russ.).
2. Abe M., Nakamura S., Shikano K. Voice conversion through vector quantization. Proc. of International Conference on Acoustics, Speech and Signal Processing. New York, 1988. pp. 655–658.
3. Valbret H., Moulines E., Tubach J.P. Voice transformation using PSOLA technique. Proc. of International Conference on Acoustics, Speech and Signal Processing. 1992. vol. 1. pp. 145–148.
4. Moulines E., Sagisaka Y. Voice conversion: State of the art and perspectives. Speech Communication. 1995. pp. 125–224.
5. Kain A., Macon M. W. Text-to-speech voice adaptation from sparse training data. Proc. of International Conference on Spoken Language Processing. 1998. pp. 2847–2850.
6. Sundermann D., Hoge H., Bonafonte A. Text-independent voice conversion based on unit selection. Proc. of International Conference on Acoustics, Speech and Signal Processing. 2006. vol. 1.
7. Azarov E, Petrovsky A.A, Lobanov B, Tsirulnik L. Text-to-speech system with acoustic processor based on the instantaneous harmonic analysis. SPECOM. 2009. pp. 414 – 418.
8. Kawahara H., Morise M. Technical foundations of tandem-straight, a speech analysis, modification and synthesis framework. SADHANA. Academy Proceedings in Engineering Sciences, 2011. pp. 713–722.
9. Kawahara H., Katayose H., Cheveigne A. Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity. Proc. Eurospeech'99. 1999. pp. 2781–2784.
10. Kawahara H., Masuda I., Cheveigne A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous–frequency–based F0 extraction. Speech Communication. 2004. pp. 187–207.
11. Arslan L. Speaker transformation algorithm using segmental codebooks. Speech Communication. 1999. vol. 28, no. 3. pp. 211–226.
12. Narendranath M ., Murthy H., Rajendran S., Yegnanarayana N. Transformation of formants for voice conversion using artificial neural networks. Speech Communication. 1995. vol. 16, no. 2. pp. 207–216.
13. Rabiner L.R. [Hidden Markov Models and their applications in speech recognition]. *Trudy instituta inzhenerov po jelektronike i radiojelektronike – Proceedings of the Institute of Engineers on Electronics*. 1989. vol. 77, no. 2. pp. 86–120. (In Russ.)
14. Azarov I.S., Petrovskij A.A. [Real-time voice conversion system with text independent learning based on the hybrid parametric descriptions of speech signals]. *Cifrovaja obrabotka signalov – Digital Signal Processing*. 2012. no. 2. pp. 15–23. (In Russ.).
15. *Analizatory rechevyh i zvukovyh signalov: metody, algoritmy i praktika (s MATLAB primerami)* [Analyzers of speech and audio signals: methods, algorithms and practice (with MATLAB examples)]. edited by A.A. Petrovskii. Minsk: Bestprint, 2009. 456 p. (In Russ.).
16. Stylianou, Y., Cappe O., Moulines E. Statistical methods for voice quality transformation. Proc. of European Conference on Speech Communication and Technology. Madrid, 1995. pp. 447–450.
17. Pavlovec A.S., Livshic M.Z., Lichachev D. S., [Voice conversion using a model of the speech with signal separation into "harmonic + noise" components and transitional frames]. *Rechevye tehnologii – Speech Technologies*. 2008. no. 4. pp. 37–50.

18. Stylianou Y., Cappe O., Moulines E. Continuous probabilistic transform for voice conversion.Proc. of International Conference on Acoustics, Speech and Signal Processing. 1998. pp. 2451–2455.
19. Patterson D. A linguistic approach to pitch range modeling. PhD dissertation. Scotland: University of Edinburgh, 2000. 201 p.
20. Zahar'ev V.A., Petrovskij A.A. Konversija prosodicheskih harakteristik diktora na osnove metodov parametrizacii kontura chastoty osnovnogo tona [Conversion of prosodic speaker features based on parameterization of the pitch contour]. *Doklady BGUIR – Proceedings – Reports of the BSUIR*. Minsk, 2013. pp. 39–46. (In Russ.).
21. Hu Y., Loizou P.C. Evaluation of objective quality measures for speech enhancement. IEEE Transactions on Audio, Speech & Language Processing. 2008. vol. 16, no. 1. pp. 229–238.

Захарьев Вадим Анатольевич — аспирант кафедры электронных вычислительных средств Белорусского государственного университета информатики и радиоэлектроники (БГУИР). Область научных интересов: цифровая обработка сигналов, методы машинного обучения, конверсия голоса, синтез и распознавание речи. Число научных публикаций — 3. zahariev@bsuir.by; БГУИР, П. Бровки, д. 6, г. Минск, 220013, РБ; р.т. +37(517) 293-8946.

Zahariev Vadim Anatol'evich — Ph.D. student of Computer Engineering Department of the Belarusian State University of Informatics and Radioelectronics (BSUIR). Research interests: digital signal processing, machine learning, voice conversion, speech synthesis and recognition. The number of publications — 3. zahariev@bsuir.by; BSUIR, P. Brovky 6, Minsk, RB; office phone +37(517) 293-8946.

Петровский Александр Александрович — д-р техн. наук., профессор, заведующий кафедрой электронных вычислительных средств Белорусского государственного университета информатики и радиоэлектроники (БГУИР). Область научных интересов: цифровая обработка сигналов, системы кодирования аудио и видео данных, мультипроцессорные системы реального времени для мультимедиа приложений. Число научных публикаций — более 600. palex@bsuir.by; БГУИР, П. Бровки, д. 6, г. Минск, 220013, РБ; р.т. +37(517) 293-8946.

Petrovsky Aleksandr Aleksandrovich — Ph.D., Dr. Sci., professor, head of Computer Engineering Department of the Belarusian State University of Informatics and Radioelectronics (BSUIR). Research interests: digital signal processing, speech and video coding systems, multiprocessor real-time systems for multimedia applications. The number of publications — more than 600. palex@bsuir.by; BSUIR, P. Brovky 6, Minsk, RB; office phone +37(517) 293-8946.

Лобанов Борис Мефодьевич — д-р техн. наук., профессор, главный научный сотрудник лаборатории распознавания и синтеза речи Объединённого института проблем информатики Национальной академии наук Республики Беларусь (ОИПИ НАН РБ). Область научных интересов: методы автоматического анализа, синтеза и распознавания речевых сигналов, человеко-машинные системы речевого общения. Число научных публикаций — 350. lobanov@newman.bas-net.by; ОИПИ НАН РБ, Сурганова 6, г. Минск, 220012, РБ; р.т. +37(517) 284-2773, факс +37(517) 331-8403.

Lobanov Boris Mefod'evich — Ph.D., Dr. Sci., professor, chief researcher of the United Institute of Informatics Problems National Academy of Sciences of Belarus (UIIP NAS RB). Research interests: methods of automatic analysis, synthesis and speech recognition, human-machine speech communication systems. The number of publications — 350. lobanov@newman.bas-net.by; UIIP NAS RB, Surganova 6, Minsk, 220012, RB, office phone . +37(517) 284-2773, fax +37(517) 331-8403.

РЕФЕРАТ

Захарьев В.А., Петровский А.А., Лобанов Б.М. **Система синтеза речи по тексту с возможностью настройки на голос целевого диктора.**

В данной статье предлагается один из подходов построения систем синтеза речи по тексту, обладающих возможностью настройки на голос целевого диктора, и получивших название мультиголосовых, с применением методов и моделей, использующихся в речевых системах другого класса — конверсии голоса. Система конверсии голоса позволяют автоматически выполнять процесс преобразования параметров голоса, характеризующих речь исходного диктора, в параметры целевого. Объектами конверсии голоса, как технологии обработки сигналов, являются стабильные во времени свойства говорящего, проявляющиеся в речевом сигнале через изменение его акустических параметров. Объединение данных двух типов систем позволяет в контексте их взаимодействия эффективно решать задачу быстрой настройки системы синтеза на произвольного диктора, без внесения существенных изменений в существующую систему синтеза, или длительного и трудозатратного процесса подготовки речевой базы для нового диктора.

Представлена архитектура системы на базе интерактивного взаимодействия между двумя типами систем, которая позволяет получить определённый синергетический эффект от их интеграции. Основными особенностями предлагаемого типа архитектуры являются: выполнение операций конкатенации и реконструкции синтезированного речевого сигнала после конверсии голоса исходного диктора в голос целевого, аспекты конверсии голоса учитываются при выборе единиц компиляции, все алгоритмы преобразования и конверсии (спектральные и просодические) выполняются единым блоком, это означает, что характеристики сигнала модифицируются только один раз.

В статье рассмотрены нюансы выбранной модели представления и параметризации речевого сигнала, понимание принципов работы которой являются ключевым момент для успешной реализации этапов анализа, конверсии и восстановления результирующего речевого сигнала со свойствами целевого диктора. В нашем случае за основу была взята модель вокодерного типа на основе взвешенной интерполяции спектра — STRAIGHT. Она позволяет разложить сигнал на три относительно независимые параметризируемые компоненты: контур ЧОТ, а также периодическую и апериодическую составляющую сигнала, и предоставляет большие возможности по гибкому изменению параметров речевого сигнала в широком диапазоне.

Также в статье были представлены расширенная регрессионная функция конверсии, применяемая для повышения точности преобразования векторов параметров, и методика постобработки сигнала на основе спектрального взвешивания — для решения проблемы чрезмерного усреднения спектральной огибающей. Всё это в итоге повышает качественные характеристики результирующего восстановленного сигнала, сохраняя вместе с тем высокие уровни узнаваемости голоса целевого диктора для преобразованной речи.

SUMMARY

Zahariev V.A., Petrovsky A.A., Lobanov B.M. **Text to speech synthesis system with the target speaker voice customization capability.**

An approach of constructing of the text to speech synthesis systems is proposed in this article. Such synthesizers have the abilities to customize the original system voice on the target speaker, using the methods and models of voice conversion technology. Voice conversion system can automatically perform the process of converting voice parameters characterizing the source speaker speech to the target speaker parameters. Voice conversion sites as signal processing technology, works with a stable over time speaker properties, manifested in the speech signal through a changes in its acoustic parameters. Combining these two types of systems we can effectively solve the problem of tuning the synthesizer on the arbitrary speaker without significant changes in the existing synthesis system, or a long and labor-intensive process of preparation of the speech database for a new speaker .

The architecture of the multivoice synthesizer based on the interactivity between the two types of systems, which allows us to get some synergy effect from such integration was presented. The main features of the proposed type of architecture are: implementation of concatenation operations and reconstruction of the synthesized speech signal after the conversion of the original speaker's voice into the voice of the title, aspects of the voice conversion are considered when choosing compilation units, all conversion and transformation algorithms (spectral and prosodic) are performed in a single block is means that the signal characteristic is modified only once.

In the article the nuances of chosen the representation and parameterization model of the speech signal were discussed. Understanding of these principles is a key to successful implementation of analysis, conversion stages and recovery of the resulting speech signal with a voice features of the target speaker. In our case speech model was based on the vocoder-type method known as a speech transformation and representation by adaptive interpolation of weighted spectrum — STRAIGHT. It allows us to decompose signal into three relatively independent parameterized components: pitch contour, maps of periodic and aperiodic component of the signal, and provides a great opportunity for flexible changing of the speech signal characteristics in a wide range.

Also in the article the extended regression conversion function that is used to improve the accuracy of the transformation parameter vectors, and a signal post-processing method based on spectral weighing to solve the problem of excessive averaging spectral envelope, were presented. All of this ultimately increases the quality characteristics of the resulting reconstructed signal, maintaining at the same time high levels of similarity of the target speaker voice for the transformed speech.