УДК 621.391.8

АРХИТЕКТУРА МУЛЬТИГОЛОСОВОГО СИНТЕЗАТОРА РЕЧИ ПО ТЕКСТУ

В.А. ЗАХАРЬЕВ, А.А. ПЕТРОВСКИЙ

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Поступила в редакцию 7 июня 2013

Предлагается схема построения мультиголосового синтезатора речи на базе использования синергетического эффекта от интеграции системы синтеза речи по тексту с системой конверсии голоса. Главной отличительной особенностью данного решения является возможность использования лингвистической, фонетической и просодической информации, имеющейся в синтезаторе речи, на этапе обучения системы конверсии голоса. Это позволяет эффективно применить текстонезависимый подход к обучению, улучшив степень качества конверсии голоса. Его использование позволяет добавить функции мультимодальности для синтезатора речи без значительных трудозатрат на подготовку речевых баз для добавления новых дикторов.

Ключевые слова: конверсия голоса, мультиголосовой синтезатор речи по тексту, текстонезависимое обучение, скрытая марковская модель, параметрическая модель представления сигнала.

Введение

На текущем этапе развития систем синтеза речи по тексту (СРТ) ставится вопрос 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 (мультиголосовые системы синтеза речи по тексту (МГСРТ)). Последний аспект требует особого подхода и внимания, поскольку перенастройка системы на нового диктора требует больших материальных и временных затрат от разработчиков системы. В данной статье предлагается рассмотреть возможности решения задачи построения мультиголосового синтезатора речи с использованием технологии конверсии голоса.

Системы синтеза речи и конверсии голоса

Система синтеза речи – это техническая или программная система, которая позволяет на основе входного орфографического текста синтезировать речевой сигнал определенным голосом, как правило, одного, заранее заданного диктора. В процессе развития сменилось три поколения систем синтеза речи по тексту, в основу которых были положены три различных подхода к синтезу фонетических характеристик речи: фонемно-артикуляторно-формантный, фонемно-формантный и фонемно-микроволновой [1]. Большинство из современных систем построены на основе последнего подхода, обобщенная структурная схема такой системы представлена на рис. 1. Как видно из данной структуры, синтезатор речи состоит из ряда процессоров, основная задача которых заключается в поэтапной обработке входного орфографического текста. Рассмотрим кратко представленные на рис. 1 компоненты синтезатора речи.

Первый компонент называется лингвистическим процессором. Он предназначен для преобразования входного орфографического текста в размеченный фонемный текст. Под

Библиотека БГУИР

разметкой понимается разбиение текста на отдельные элементы в следующей иерархии: фонетический период, фразы, синтагмы. Кроме того, лингвистический процессор осуществляет: расстановку ударений, интонационную маркировку. Далее размеченный фонемный текст поступает на вход двух следующих процессоров: просодического и фонетического [1].

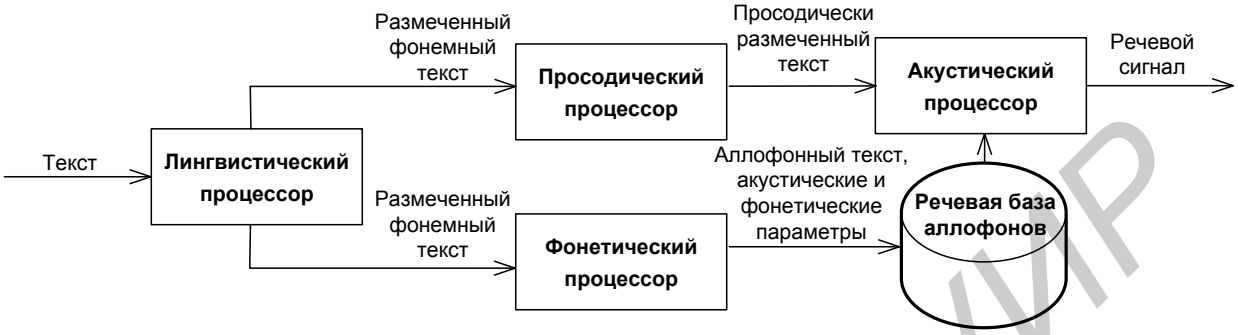

Рис. 1. Обобщенная структурная схема синтезатора речи по тексту

Задача фонетического процессора заключается в том, чтобы на основе специальной базы фонетических правил и алгоритма преобразования фонема-аллофон, выполнить подстановку позиционных и комбинаторных аллофонов в фонемный текст, сформировав тем самым аллофонный текст, который является набором входных команд для речевой базы данных (БД) синтезатора.

В результате работы просодического процессора фонемный текст делится на акцентные группы. Далее осуществляется разметка акцентной группы на элементы акцентных групп: интонационные предъядро, ядро и заядро. Затем производится расстановка значений интенсивности или амплитуды, длительности фонем и частоты основного тона или мелодики для каждого из элементов акцентных групп. Просодический процессор также работает со специальной просодической базой данных и правил.

Акустический процессор на основании информации от соответствующих процессоров о том, какие аллофоны требуется синтезировать, а также какие просодические характеристики должны быть приписаны каждому аллофону, синтезирует речевой сигнал. Акустический процессор использует соответствующую БД, в которой хранятся акустические эталоны аллофонов, правила модификации аллофонов и правила модификации синтезируемого голоса для конкретного диктора. Поэтому для добавления нового диктора необходимо создание новой акустической БД, что несет, в свою очередь, существенные материальные и временные издержки.

Система конверсии голоса – это система, реализующая процесс конверсии параметров голоса, характеризующих исходного диктора (ИД), в параметры целевого диктора (ЦД), без изменения лингвистической составляющей самого сообщения. Типовая обобщенная структурная схема системы конверсии голоса представлена на рис. 2 [2]. Процесс работы системы осуществляется в два этапа: обучения и конверсии.

Рис. 2. Структурная схема системы конверсии голоса

Этап обучения. На вход системы поступают речевые сигналы исходного и целевого дикторов. В соответствующих блоках производится их анализ согласно некоторой модели представления сигнала (авторегрессионная, синусоидальная, гибридная и т.д.), и отыскивается параметрическое описание для каждого фрейма. Далее для двух последовательностей векторов

Библиотека БГУИР

параметров производится операция масштабирования для устранения временного несоответствия между ними. Затем осуществляется разделение пространства параметров сигнала на акустические классы для каждого диктора с использованием выбранной модели кластеризации. Параметры данной модели содержат в себе дикторозависимую информацию о тембральных (огибающая спектра и др.) и интонационных (частота основного тона – ЧОТ и др.) признаках голоса говорящего и в последующем используются в качестве коэффициентов функции конверсии признаков. Этап обучения считается завершенным. В дополнение необходимо отметить, что если исходные фразы совпадают по лингвистическому и фонетическому составу, такое обучение носит название текстозависимого, в противном случае оно называется текстонезависимым [2]. Второй вариант сложнее в реализации, однако он является более перспективным с точки зрения интеграции с СРТ, поскольку позволяет гибко осуществлять настройку системы на целевого диктора, без использования в процессе обучения специально подготовленных параллельных обучающих выборок фраз [3].

Этап конверсии. На вход системы подается речевой сигнал только исходного диктора. Для каждого фрейма производится анализ речевого сигнала в соответствии с тем же методом, что и на этапе обучения. Далее осуществляется непосредственно процесс конверсии – отображение при помощи функции конверсии, характеристического вектора исходного диктора в пространство акустических признаков целевого диктора таким образом, чтобы максимально приблизиться к соответствующему характеристическому вектору целевого диктора. Полученная последовательность модифицированных векторов используется для синтеза речевого сигнала исходного диктора с характерными чертами целевого диктора.

Обоснование выбора архитектуры системы

К рассмотрению предлагаются два варианта архитектуры, представленные на рис. 3 и условно обозначенные как вариант архитектуры на базе суперпозиции систем и вариант на базе их синергии. Первый подход подразумевает существование двух абсолютно независимых систем: синтезатора речи, на вход которого подается предназначенный к озвучиванию текст, а на выходе синтезируется речь диктора, выступающего в качестве исходного, которая затем системой конверсии преобразовывается в речь, произносимую голосом целевого диктора. Таким образом, системы никак не связаны друг с другом, в том смысле, что в процессе работы синтезатор речи передает системе конверсии лишь конечный продукт, который затем обрабатывается системой конверсии без учета принципов обработки его в синтезаторе.

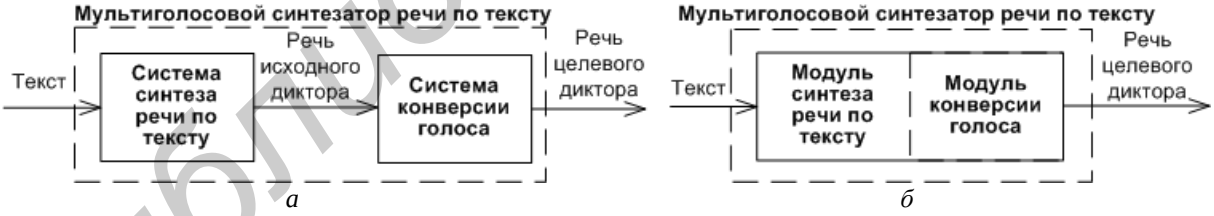

Рис 3. Архитектура мультиголосового синтезатора речи на базе суперпозиции систем (*а*) и на базе синергии систем (*б*)

Положительный момент данного подхода заключается в устойчивости и универсальности данной архитектуры, в рамках которой легко может быть заменена любая составляющая без нарушения общей работоспособности всей системы. Например, при смене модели представления сигнала или конверсии на более совершенную, доработанную модель.

Второй подход, на базе синергии, рассматривает объединение двух типов систем не как их простую сумму, а как интеграцию в рамках единой системы, включающей необходимые модули, взятые из каждой. Данный подход позволяет максимально учесть особенности решаемой задачи и эффективно использовать внутреннюю информацию каждой из систем, сделав ее разделимой между частями. Например, информацию о важных параметрах речи и передаваемого сообщения из синтезатора на всех уровнях (лингвистическом, фонетическом, акустическом) можно передать в систему конверсии и за счет этого улучшить ее качественные характеристики. Однако подход на базе синергии систем требует детальной переработки и

Библиотека БГУИР

изменения принципов построения каждой из них, вычленения необходимых модулей и создания новой структуры, сформированной под решение задачи многоголосого синтеза.

Анализ вышеперечисленных типов архитектур показал, что для построения МГСРТ наиболее подходящим является вариант на базе синергии систем. Поскольку именно данный тип архитектуры за счет использования разделяемой информации из модуля синтеза позволит в максимально сжатые сроки осуществить перенастройку на нового диктора, и выполнять синтез речевого сигнала только один раз, сразу с характеристиками голоса целевого диктора, на выходе модуля конверсии. Исходя из этого, рассмотрение и дальнейшие исследования велись в контексте второй концепции архитектуры системы МГСРТ.

Структура мультиголосовой системы синтеза речи по тексту

На основании концепции синергетического подхода при выборе архитектуры, рассмотренной в предыдущем разделе, была предложена следующая структура системы МГСРТ, представленная на рис. 4. Система включает в себя необходимую информацию и набор модулей из состава синтезатора речи (элементы затушеваны на схеме серым цветом) и системы конверсии голоса. На схеме (рис. 4) одновременно представлены два варианта структуры системы для режимов обучения и, непосредственно, работы, поскольку в зависимости текущего режима набор элементов и используемых данных будет несколько изменяться.

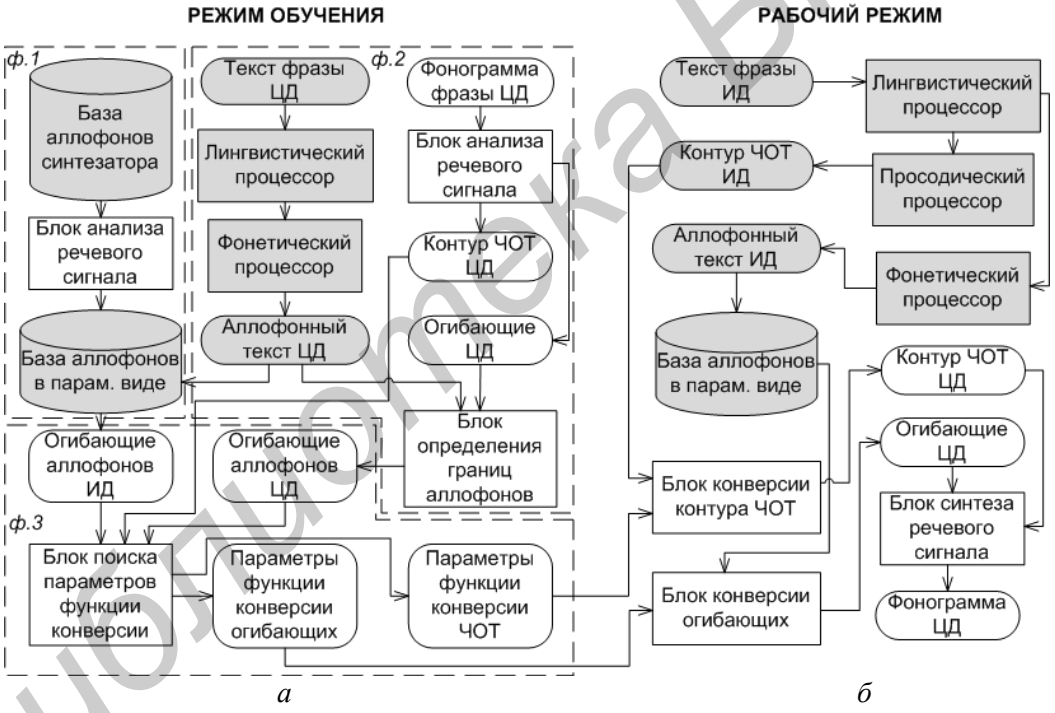

Рис 4. Структурная схема мультиголосового синтезатора речи в режиме обучения (*а*) и в рабочем режиме (*б*)

Характерной особенностью предлагаемой схемы является наличие специальной подготовительной фазы на этапе обучения (рис. 4, *а*, ф. 1), в ходе которой осуществляется анализ записей аллофонов, хранящихся в базе синтезатора, с целью их представления в параметрическом виде. Таким образом, результирующая база аллофонов в параметрическом виде является исчерпывающим хранилищем информации обо всех фонетических и акустических характеристиках голоса ИД, сведения из которого используются в ходе всех последующих этапов. Данное действие выполняется единожды для каждого исходного диктора, и не повторяется при перенастройке системы на нового целевого диктора, что также является достоинством предлагаемой схемы. Анализ аллофонов производится на основании методов и моделей, подробно описанных далее. В ходе второй фазы (рис. 4, *а*, ф. 2),

Библиотека БГУИР

параллельно выполняются два процесса: подготовка и анализ речевой и текстовой информации о фонетических и акустических особенностях целевого диктора. Априорной информацией для этих процессов являются последовательность фонограмм обучающих фраз $Wav=(w_1, w_2,..., w_n)$ и соответствующая ей последовательность орфографической записи этих фраз $Torpho=(t_1, t_2,..., t_n)$, где n – количество фраз обучающей выборки. Далее над текстовой информацией лингвистическим процессором осуществляется преобразования орфографического текста в фонемный вид L: $Torpho \rightarrow Tphono \in D^{s\times n}$, где s – количество фонетически различимых единиц в одной фразе, а затем фонетический процессор выполняет преобразование фонемного текста, F: $Tphono \rightarrow Tallo \in D^{a\times n}$, где a – количество аллофонов (комбинаций фонем) в одной фразе, в последовательность индексов аллофонов. Необходимо отметить, что алфавит данных индексов совпадает для всех дикторов, поскольку их состав строго определен и неизменен для представителей одной языковой группы. Алгоритмы, реализующие данные преобразования, подробно изложены в литературе [1].

Блок анализа речевого сигнала выполняет над последовательностью фонограмм W преобразование A, основанное на дискретном преобразовании Фурье согласованном с изменением ЧОТ [4], общий вид которого можно записать:

$$A \Leftrightarrow X(k) = \sum_{i=0}^{I-1} x(i) e^{-j\varphi(i,k)},\ k = 1...K, \qquad (1)$$

$$\varphi(i,k) = \frac{2\pi i k}{F_s}\left(F_0 + \frac{\Delta F_0 i}{2I}\right), \qquad (2)$$

где $X(k)$ – k-ый коэффициент Фурье, $x(i)$ – i-ый отсчет входного сигнала, I – длина фрейма анализа в отсчетах, K – количество гармоник сигнала, F_0 – частота основного тона, ΔF_0 – изменение ЧОТ, F_s – частота дискретизации. Поиск частоты основного тона F_0 производится на основе метода поиска максимума нормализованной автокорреляционной функции. Модифицированное ядро преобразования (2) позволяет учесть линейное изменение ЧОТ в пределах фрейма анализа. Использование выражения (1) для анализа сигнала позволяет получить более четкую локализацию энергии в спектре сигнала. Для уменьшения вычислительной сложности последующих этапов и результирующей модели конверсии Фурье-спектр заменяется своей огибающей в параметрическом виде с использованием линейных спектральных частот (ЛСЧ). Таким образом, преобразование, выполняемое блоком анализа речевого сигнала, формально можно определить как

$$A: Waw \rightarrow Prm \in D^{p\times m\times n} \mid Prm(m,n) = \{F_0, \Delta F_0, a_1, a_2, ..., a_p\},$$

где p – количество параметров ЛСЧ, m – номер фрейма сигнала, n – номер фонограммы в выборке.

Далее последовательности индексов аллофонов $Tallo$ и векторов параметров сигнала Prm одновременно поступают на входы блока определения границ аллофонов, в котором производится установление оптимального соответствия между ними. Это необходимо для сопоставления индекса каждого аллофона набору векторов параметров сигнала, что в дальнейшем позволяет сформировать совместную последовательность обучения для блока поиска параметров функции конверсии, на основе равенства алфавитов индексов аллофонов для исходного и целевого дикторов. Данная задача эффективно может быть решена с использованием аппарата скрытых марковских моделей (СММ), в рамках которого, последовательность $Tallo$ будет являться последовательностью состояний, а Prm – последовательностью наблюдений. С точки зрения СММ данный процесс заключается в том, чтобы связать оптимальную последовательность состояний с текущей последовательностью наблюдений для модели (3). Решение данной задачи, формализованной в виде выражения (4), возможно с помощью использования итерационного алгоритма Витерби [5].

$$H: (Tallo, Prm) \rightarrow (Tallo^{opt}, Prm), \qquad (3)$$

Библиотека БГУИР

$$\arg\max P(Tallo, Prm \mid \lambda),\ \lambda \in (A, B, \pi), \quad (4)$$

$$Ballo^{trg} = \{\forall t \mid Tallo^{opt}(t)\} \Leftrightarrow Prm^{trg} = \{\forall p \mid Prm(p)\}, \quad (5)$$

где λ – скрытая марковская модель, A – матрица состояний СММ, B – матрица наблюдений СММ, π – матрица начального распределения состояний СММ. Далее, путем объединения статистики всех наблюдений по каждому из состояний, по всем фразам обучающей выборки (2.3), возможно найти соответствие векторов параметров, относящихся к определенному аллофону, благодаря равенству аллофонных алфавитов с точки зрения его состава $Ballo^{src}(i) = Ballo^{trg}(i) = Ballo(i) \Rightarrow Prm^{src}(i) \Leftrightarrow Prm^{trg}(i)$. Таким образом, все вышеперечисленные действия в результате выполнения двух этапов, позволяют сформировать совместную последовательность векторов параметров сигнала по фонетическому принципу $Z = (\{Prm^{src}(i), Prm^{trg}(i)\}_i),\ \forall i \in N,\ i = \overline{1, I}$ (где I – количество аллофонов в базе) для его последующего использования на следующей фазе.

В ходе третьей, завершающей фазы, этапа обучения совместная последовательность векторов поступает на вход блока поиска параметров функции конверсии, производящего кластеризацию совместного пространства признаков Z. Для удобства последовательность параметров ИД обозначим $Prm^{src} = \vec{x} = (x_1, x_2, ..., x_N)$, а для ЦД $Prm^{trg} = \vec{y} = (y_1, y_2, ..., y_N)$, где N – количество векторов параметров размерность p для всех фреймов всех аллофонов всех фраз. Тогда совместный вектор параметров можно записать как $Z = (\{Prm^{src}, Prm^{trg}\}) = \vec{z} = [x^T y^T]^T$. Характеристики найденных классов являются параметрами функции конверсии огибающих, которые используются в процессе работы системы. Метод их определения основан на использовании аппарата множественных гауссовых смесей (МГС) для моделирования функции плотности распределения векторов спектральной огибающей [6]. Модель МГС позволяет выполнить мягкую классификацию, учитывая тот факт, что акустические классы в пространстве могут перекрываться.

Функция плотности вероятности в МГС задается как, взвешенная сумма многомерных функций распределения Гаусса:

$$p(\vec{z} \mid \alpha, \mu, \Sigma) = \sum_{q=1}^{Q} \alpha_q G(\vec{z} \mid \mu_q, \Sigma_q), \quad (6)$$

где z – совместный вектор параметров сигнала размерностью $2p \times N$, G – компонента смеси, представляющая собой $2p$-мерную функцию распределения Гаусса; α – веса компонент смеси, $\alpha \geq 0, \forall q = 1, ..., Q, \sum \alpha_q = 1$, $\mu_q = [\mu_q^{\vec{x}\vec{x}}\ \mu_q^{\vec{y}\vec{y}}]^{\mathrm{T}}$ – вектор математических ожиданий размерностью $2p$, $\sum_q = I \times [\sum_q^{\vec{x}\vec{x}}\ \sum_q^{\vec{y}\vec{x}}]^{\mathrm{T}} + I^T \times [\sum_q^{\vec{x}\vec{y}}\ \sum_q^{\vec{y}\vec{y}}]^{\mathrm{T}}$ – ковариационная матрица размерностью $2p \times 2p$. Перечисленные параметры в выражении (3) определяются с использованием известного итерационного *EM*-алгоритма [6]. Следовательно, модель МГС полностью определяется следующим набором параметров $\theta = \{a_q, \mu_q, \Sigma_q\}$, для q = 1, ... ,Q. Таким образом, совместное пространство параметров исходного и целевого диктора описывается с помощью Q гауссовых смесей, имеющих набор параметров θ. На поиске параметров МГС этап обучения системы МГСРТ можно считать завершенным. Данный этап требует его провидения единожды для каждого нового голоса, добавляемого в систему.

Структура МГСРТ в рабочем режиме представлена на рис. 4, *б*. В процессе функционирования в данном режиме на вход системы поступает орфографический текст $Torpho$, который обрабатывается лингвистическим и фонетическим процессорами $F(L(Torpho)) \rightarrow Tallo$, аналогично второй фазе этапа обучения. Кроме того, просодическим процессором на основе интонационной маркировки выполненной лингвистическим процессором, а также собственной базы данных и правил выполняется формирование просодического контура $Prsdy^{src} = (\{F_0^{src}, A_{F_0}^{src}, T_{F_0}^{src}\}_1, ...\{F_0^{src}, A_{F_0}^{src}, T_{F_0}^{src}\}_N)$, задающего энергетику, ритмику и мелодику синтезируемого текста. Далее, на основе аллофонного текста $Tallo$ из параметрической базы аллофонов исходного диктора, собирается последовательность векторов

Библиотека БГУИР

акустических параметров $Prm^{src} = \vec{x}$. Она поступает в блок конверсии огибающих, где согласно выбранной регрессионной функции конверсии с параметрами, определенными на этапе обучения, конвертируется в последовательность векторов целевого диктора, согласно выражению

$$y^* = F(x) = \sum_{q=1}^{Q} p_q(x)[\mu_q^y + \sum_q^{yx} \sum_q^{xx-1} (x - \mu_q^x)], \quad (7)$$

где y^* – сконвертированный вектор параметров спектральной огибающей, F – функция конверсии, $p_q(x)$ – апостериорная вероятность принадлежности элемента последовательности векторов $\vec{x}$ к классу q модели МГС. Преобразование контура параметров просодики $F_{Prsdy} : Prsdy^{src} \rightarrow Prsdy^{trg*}$ осуществляется в блоке конверсии ЧОТ согласно методике представленной авторами в статье [7]. Далее сконвертированные последовательности векторов параметров спектральной огибающей Prm^{trg*} и просодики $Prsdy^{trg*}$ поступают на блок синтеза речевого сигнала, который является альтернативой акустическому процессору из канонической схемы системы синтеза. Данный блок производит восстановление речевого сигнала из набора параметров на базе обратного преобразования Фурье, согласованного с частотой основного тона. На выходе блока мы получаем речевой сигнал озвучиваемой фразы с характеристиками голоса целевого диктора, тем самым решив поставленную задачу создания МГСРТ с текстонезависимым обучением.

Заключение

Рассмотрены подходы и принципы построения систем конверсии голоса и систем синтеза речи. Приводится обоснование выбора архитектуры системы, а также предлагается возможный вариант структурной схемы МГСРТ с описанием принципа их функционирования. Предлагаемая схема, благодаря синергетическому эффекту от интеграции двух видов систем, позволяет в полной мере использовать полезные свойства обеих и решает задачу создания МГСРТ с улучшенными показателями узнаваемости диктора и удобства использования системы.

ARCHITECTURE OF THE MULTIVOICE TEXT-TO-SPEECH SYSTEM

V.A. ZAKHARYEU, A.A. PETROVSKY

Abstract

Architecture of the multimodal text to speech synthesis system based on the voice conversion framework was proposed. Such system could be tuned to the specific speaker without any costs losses on the training phase and based on one speaker base, having in TTS system. Structural scheme for this type of the speech synthesizer, with the description of the functionality of the main blocks were presented. Their specific characteristics are synergy approach to the architecture and text-independent mode in the training phase.

Список литературы

1. *Лобанов Б.М.* Компьютерный синтез и клонирование речи. Минск, 2008.
2. *Sundermann D.* // ICASSP. 2006. P. 81–84.
3. *Duxans B.* // PUC. 2006. P. 171–175.
4. Анализаторы речевых и звуковых сигналов: методы, алгоритмы и практика. // Под ред. А.А. Петровского. Минск, 2009
5. *Bourlard H.* Introduction to Hidden Markov Models. Lauseane, 2010.
6. *Stylianou Y.* // Springer. 2007. P. 502–532.
7. *Захарьев В.А, Петровский А.А.* // Докл. БГУИР. 2013. № 1 (71). С. 39–45.

Библиотека БГУИР