АСПЕКТЫ ПРАКТИЧЕСКОЙ РЕАЛИЗАЦИИ МУЛЬТИГОЛОСОВОГО СИНТЕЗАТОРА РЕЧИ ПО ТЕКСТУ

В.А. Захарьев, А.А. Петровский
Кафедра электронных вычислительных средств
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: {zahariev, palex}@bsuir.by

В докладе предлагается рассмотреть возможность решения задачи построения мультиголосового синтезатора речи с использованием технологии конверсии голоса, на основе системы синтеза речи по тексту «Мультифон», разработанной в лаборатории компьютерного синтеза и распознавания речи Объединённого института проблем информатики НАН Беларуси [1], а также результаты анализа ряда практических аспектов такой реализации.

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 (МГ ССРТ) – требует особого подхода и внимания, поскольку в существующих ССРТ перенастройка системы на нового диктора требует больших материальных и временных затрат от разработчиков системы.

I. Архитектура МГ 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 (ЦД). Объектами конверсии голоса, как технологии обработки сигналов, являются стабильные во времени свойства говорящего, проявляющиеся в речевом сигнале через изменение его акустических параметров [2–5].

Вполне очевидной является попытка 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[6,7].Однако, такой 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. Желательно чтобы система конверсии голоса имела доступ к параметрам модели синтеза и могла работать с ними напрямую. Хотя блок выбора компиляционных единиц оптимизирован для получения как можно более естественной синтетической речи без существенных неоднородностей, тот факт, что полученный в результате речевой сигнал должен быть преобразован с помощью определённой функции конверсии голоса, должен быть также приняты во внимание. Это достигается путём интеграции модуля конверсии голоса в блок акустического процессора и рационального разделения задач конверсии просодики и параметров голоса между двумя видами систем. Архитектура мультиголосовой ССРТ представлена на рис. 1, в которой были учтены отмеченные выше замечания.

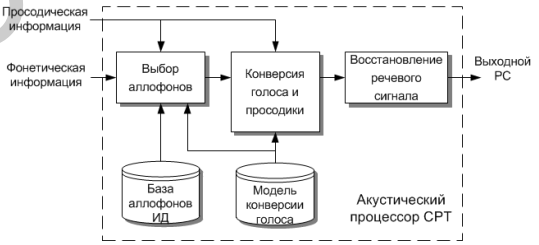

Рис. 1 – Архитектура МГ ССРТ на основе интеграции систем

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.

II. Аспекты реализации МГ ССРТ

В ходе дальнейшей работы была произведена имплементация предлагаемой архитектуры в виде разработанных подробных алгоритмов и текстов исходных кодов, а также созданного на их основе программное обеспечение (ПО) системы мультиголосового синтеза речи по тексту.

Были выявлены следующие основные аспекты и технические нюансы построения данных систем, требующие дополнительных разъёснений для увеличения технологичности создания подобных систем в дальнейшем.

Необходимо отметить, за основу программного ядра синтеза при построении МГ ССРТ был взят существующий синтезатора речи по тексту “Мультифон”, разработанной в лаборатории компьютерного синтеза и распознавания речи ОИПИ НАН РБ[1]. И поскольку функция мультиголосового синтеза в виде системы конверсии голоса, согласно разработанной интегрированной архитектуре, внедряется только на уровне акустического процессора, то авторами статьи совместно с разработчиками синтезатора “Мультифон”, было принято решение, не выполнять заново полную имплементацию СРТ, а воспользоваться как можно большим количеством уже существующих компонентов синтезатор. Так как все остальные блоки синтезатора, такие как лингвистический процессор, а также просодический и фонетический процессоры, не требовали кардинальных изменений по сравнению с реализации не мультиголосового системы синтеза. Это позволило сэкономить ресурсы на разработку новой системы синтеза речи по тексту и основные усилия направить на реализацию акустического процессора с функцией конверсии голоса, однако требовало необходимости поиска соответствующих механизмов для того, чтобы не нарушить функционирование существующих подсистем.

Вторым технических нюансом стал факт разной степени уровня программной реализации и проработанности объединяемых типов систем СРТ и КГ.Система КГ применялась в виде ПО, носящего научно-исследовательский и инструментальный характер, чем конкретное прикладное решение уровня коммерческой реализации, что было связано с необходимостью частого внесения изменений в систему. Добавления и реализации новых моделей и алгоритмов, проверки гипотез, изменения внутренних структур данных и формата ввода-вывода информации в систему. Эта проблема была решена путём использования технологии упрощенного генерирования оболочек и интерфейсов SWIG (Simplified Wrapper and Interface Generator) [8]. SWIG является инструментальным ПО для связывания программ и библиотек написанных на C/C++ со скриптовыми языками, такими как Perl, Python, Кгин, PHP, а также языками с промежуточным представлением кода такими как и др. Технология предоставляет удобный программный интерфейс доступа и связи программных компонентов с минимальными усилиями: в файлы заголовка программы добавляется небольшое количество указаний, по которым SWIG генерирует исходный код для интеграции компонентов на C/C++ и нужного языка. Поскольку непосредственной реализации SWIG для языка Matlab не существует, то в качестве скриптового языка посредника был выбран Python, котором легко интегрируется с Matlab. Диаграмма компонентов системы мультиголосового синтезатора речи по тексту представлена на рис. 2.

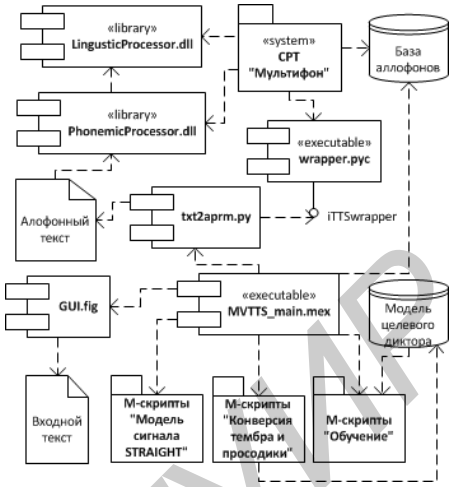

Рис. 2 – Диаграмма компонентов МГ ССРТ

Заключение

В докладе рассмотрены аспекты практической реализации мультиголосовых систем синтеза речи по тексту. Предложена новая архитектура на базе интеграции систем синтеза речи и конверсии голоса на уровне видоизменённого акустического процессора системы синтеза. Рассмотрены нюансы технической реализации предлагаемой архитектуры в контексте применения существующей системы «Мультифон» и специально разработанного для неё программного модуля конверсии голоса. Полное или частичное выполнение выработанных рекомендаций позволяет повысить технологичность предлагаемого решения как законченного прикладного программного продукта.

1. Лобанов, Б.М. Компьютерный синтез и клонирование речи / Б.М. Лобанов, Л.И. Цирульник. – Минск: Белорусская наука, 2008. –344 с.
2. Stylinau, Y. Voice tracnsformation: A survey / Y. Stylinau. // ICASSP. –2009. –P. 3585–3588.
3. Voice conversion: a crtitical survey [Electronic resource] /A. F. Machado, M. Queiroz // Open-acess article. –2010. –Access mode: http://www.ime.usp.br/ mqz/SMC2010Voice.–Date of access: 13.05.2013.
4. Toda, T. Spectral conversion based maximum liklehood estimation considering global variance of converted parameter / T. Toda, A. Black, K. Tokuda. // ICASSP. –2005. –P. 9–12.
5. Stylinau, Y. Continuous probabilistic transform for voice conversion. / Y. Stylianou // IEEE TSAP. – 1998. –№ 6 –P. 131–142.
6. Анализаторы речевых и звуковых сигналов /под ред. д.т.н. профессора Петровского А.А. –Минск: Бестпринт, 2009. –456 с.
7. Erro, D. Weighted frequency warping for voice conversion / D. Erro, A. Moreno // Audio, Speech, and Language Processing, IEEE Transactions on – 2010. – Vol. 18. –P. 543–550.
8. Simplified Wrapper and Interface Generator [[Electronic resource] –Access mode: http://www.swig.org –Date of access: 01.09.2014.