КОНВЕРСИЯ ГОЛОСА ДЛЯ СИСТЕМ МУЛЬТИГОЛОСОВОГО СИНТЕЗА 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 рассмотрены вопросы применения технологии конверсии голоса для построения мультиголосовых систем синтез речи по тексту (МГСРТ). Описана разработанная интегрированная архитектура для построения МГСРТ. Обоснован выбор модели представления и параметризации речевого сигнала на базе STRAIGHT. Предложена функция конверсии голоса на основе МГС и множественной регрессионной функции отображения, позволяющая улучшить качественные показатели работы системы МГСРТ.

Введение

На данном этапе развития систем синтеза речи по тексту (СРТ) 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 систем мультиголосового синтеза речи по тексту (МГСРТ), требует особого внимания, поскольку пользователями зачастую становиться востребована функция перенастройка системы под себя, добавление массива новых голосов или даже настройка на голос произвольного диктора. В настоящее время реализация такой функции персонализации связана с большими материальными и временными затратами. В докладе решение проблемы создания МГСРТ предлагается рассмотреть с помощью другой перспективной речевой технологии получившей название конверсии голоса [1].

I. Архитектура МГ ССРТ

Конверсия голоса (КГ) —– это технологии обработки речевого сигнала (РС), позволяющей реализовать процесс трансформации параметров голоса, характеризующих речь исходного диктора (ИД), в параметры целевого (ЦД) без изменения смысла сообщения [2]. Объектами конверсии голоса в первую очередь выступают тембральная (спектральная огибающая) и просодическая (контур частоты основного тона – ЧОТ) характеристики диктора. Для решения задачи построения МГСРТ с помощью технологии КГ был сформулирован и предложен подход на основе интегрированной архитектуры системы СРТ и КГ. Предлагается встроить функциональные блоки конверсии голоса в состав компиляционного СРТ на уровне акустического процессора системы. В качестве исходно используется информация о голосе диктора, хранящаяся в БД акустических фрагментов речевой волны синтезатора в параметризированном виде. Предлагаемый подход позволяет достичь большей степени связности между двумя типами систем. Архитектура МГСРТ представлена на рис. 1, в которой были учтены отмеченные выше замечания.

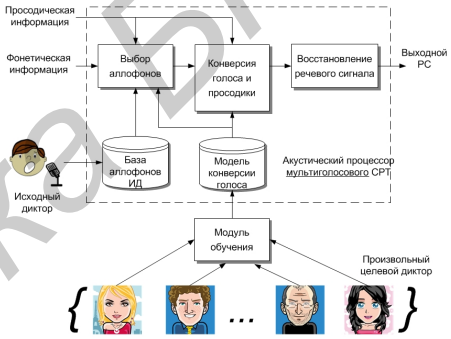

Рис. 1 – Архитектура МГСРТ на основе интеграции систем СРТ и КГ

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. Модель представления речевого сигнала

В ходе проведённого обзор литературных источников и сравнительного анализа актуальных методов речевого сигнала [3], было установлено что наиболее перспективной для построения МГСРТ является модель STRAIGHT [4]. Она позволяет разложить сигнал на три компоненты: контур ЧОТ, периодическую, апериодическую и независимо манипулировать ими.Вести анализ РС синхронизировано с частотой основного тона, в области мгновенных параметров сигнала. Затем на основе усреднённой оценки меж-

ду двумя мгновенными значениями спектральной огибающей получить сглаженное частотно-временное представление огибающей спектра.

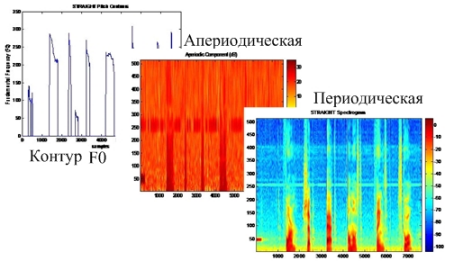

Рис. 2 – Компоненты сигнала на основе модели STRAIGHT

STRAIGHT-спектрограмма определяется согласно выражению:

$$P_T(\omega) = \frac{1}{N}\sum_{k=0}^{N-1} |S(\omega, \tau + \frac{kT_0}{N})|^2, \qquad (1)$$

где $P_T(\omega)$ – сглаженный стабильный во времени спектр мощности сигнала,при условии что центры локализации временных окон разделены на $\frac{T_0}{N}$, $N \in Z$ – количество окон для расчёта, $|S(\omega,\tau)|^2$ – мгновенный Фурье-спектр для момента времени τ, $\tau + \frac{kT_0}{N}$ – смещение по времени для анализа спектра сигнала в момент $|S(\omega, \tau + \frac{T_0}{2})|^2|$. Такое представление (1) позволяет точно определить значения как периодической, так и для апериодической компоненты сигнала, наиболее подходящей для трансформации спектральных огибающих. В дальнейшем оно даёт возможность наиболее просто и без внесения дополнительных ошибок выполнять конверсию каждой из дикторозависимых характеристик сигнала (огибающей спектра для периодической и апериодической компонент, а также параметров просодики в виде контура ЧОТ). Метод также позволяет манипулировать параметрами источника возбуждения, выполнять восстановление сигнала из параметрической области и обеспечивать синтез сигнала с малыми вносимыми искажениями.

III. Конверсия на основе МГС

На данный момент самой распространённой моделью является статистическая модель на основе множественных гауссовых смесей (МГС), которая в оригинале была предложена [5], а её доработанные варианты представлены [6,7]. Системы на базе данной модели имеют удовлетворительные результаты в плане характеристики сходства между преобразованным и целевым речевыми сигналами. Обучение производится на основе набора параллельных пар векторов параметров ИД и ЦД,для которых строиться совместная модель МГС. Веткора математических ожиданий и ковариационные матрицы предоставляемы МГС используются в качестве параметров функции конверсии, которая представлена выражением:

$$F(\mathbf{x}) = \sum_{i=1}^{M} p_i(\mathbf{x})[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(\mathbf{x} - \mu_i^x)], \quad (2)$$

$$p_i(\mathbf{x}) = \frac{\alpha_i N(\mathbf{x}, \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{M} \alpha_j N(\mathbf{x}, \mu_j^x, \Sigma_j^{xx})}.$$

где $\mathbf{x}$ – вектор параметреров исходного диктора,M – количество компонент смеси, μ_i^x и μ_i^y – вектора мат. ожиданий i-ой компоненты смеси, Σ_i^{xx} – ковариационная матрица исходного диктора i-ой компоненты, Σ_i^{yx} – кросковариационная матрица для векторов исходного и целевого диктора i-ой компоненты, $p_i(\mathbf{x})$ –апостореорная вероятность принадлежности вектора $\mathbf{x}$ i-ой компоненте, N – многомерное Гауссово распределение с параметрами преведёнными выше. Функция конверсии вида (2) позволяет учесть статистические зависимости и принадлежность к определённому акустическому классу не только элементов вектора параметров сигнала текущего фрейма, но также и соседних с ним фреймов.

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. Более подробно с основными можно в статье [8].

1. Shikano, K. Speaker adaptation through vector quantization / K. Shikano, K. Lee, R. Reddy // ICASSP. – 1986. –Vol. 11. –P. 231–237.
2. Stylinau, Y. Voice tracnsformation: A survey / Y. Stylinau. // ICASSP. –2009. –P. 3585–3588.
3. Анализаторы речевых и звуковых сигналов /под ред. д.т.н. профессора Петровского А.А. –Минск: Бестпринт, 2009. –456 с.
4. Kawahara H. et al. TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation // ICASSP. – 2008. –IEEE. –P. 3933–3936.
5. 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.
6. Stylinau, Y. Continuous probabilistic transform for voice conversion. / Y. Stylianou // IEEE TSAP. – 1998. –№ 6 –P. 131–142.
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. Захарьев, В.А. Система синтеза речи по тексту с возможностью настройки на голос целевого диктора / В.А. Захарьев, А.А. Петровский, Б.М. Лобанов // Труды СПИИРАН. – Спб: СПИИРАН. – 2014. – №1(32). – С. 82-98.