УДК 681.51.011
Людовик Т.В., Сажок Н.Н.

ИСПОЛЬЗОВАНИЕ РЕЧЕВЫХ БАЗ ДАННЫХ БОЛЬШОГО ОБЪЕМА ПРИ СИНТЕЗЕ РЕЧИ В СИСТЕМАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ВВЕДЕНИЕ

Системы искусственного интеллекта предусматривают как управление голосом, так и синтез речи для обратной связи с пользователем. Синтезаторы речи основываются либо на использовании моделей речевого тракта человека [1][2], либо на конкатенации (склеивании) необходимых сегментов речи, хранящихся в речевых БД [3][4].

Конкатенативный метод синтеза речи состоит в хранении, выборе и сглаженной конкатенации (склеивании) предварительно записанных сегментов речи, которой может предшествовать модификация просодических характеристик, а именно длительности и частоты основного тона (ЧОТ).

Конкатенативные синтезаторы зависят от используемых речевых БД. Чем больше объем речевой БД, то есть, чем полнее представлена в ней звуковая, интонационная, темпоральная вариативность речи, тем естественнее звучащую синтезированную речь можно получить. При этом большую роль играет формулирование критериев автоматического выбора нужного речевого сегмента из множества всех хранящихся в БД.

Используемые в данной работе критерии выбора элементов из БД опираются на синхронную с ЧОТ сегментацию речевых отрезков, хранящихся в БД, и сегментно-просодическое описание этих речевых отрезков [5].

В настоящее время разработаны две БД с мужскими голосами и разрабатываются две БД с женскими голосами.

Далее приводится общая архитектура конкатенативного синтезатора украинской речи и более подробно описывается взаимодействие модулей синтезатора с речевой БД.

ОБЩАЯ АРХИТЕКТУРА КОНКАТЕНАТИВНОГО СИНТЕЗАТОРА УКРАИНСКОЙ РЕЧИ ПО ТЕКСТУ

На Рис. 1 представлена блок-схема конкатенативного синтезатора украинской речи по тексту. Синтезатор состоит из четырех модулей:

- модуля речевых БД;
- лингвистического процессора;
- модуля выбора оптимальных элементов из БД;
- акустического процессора.

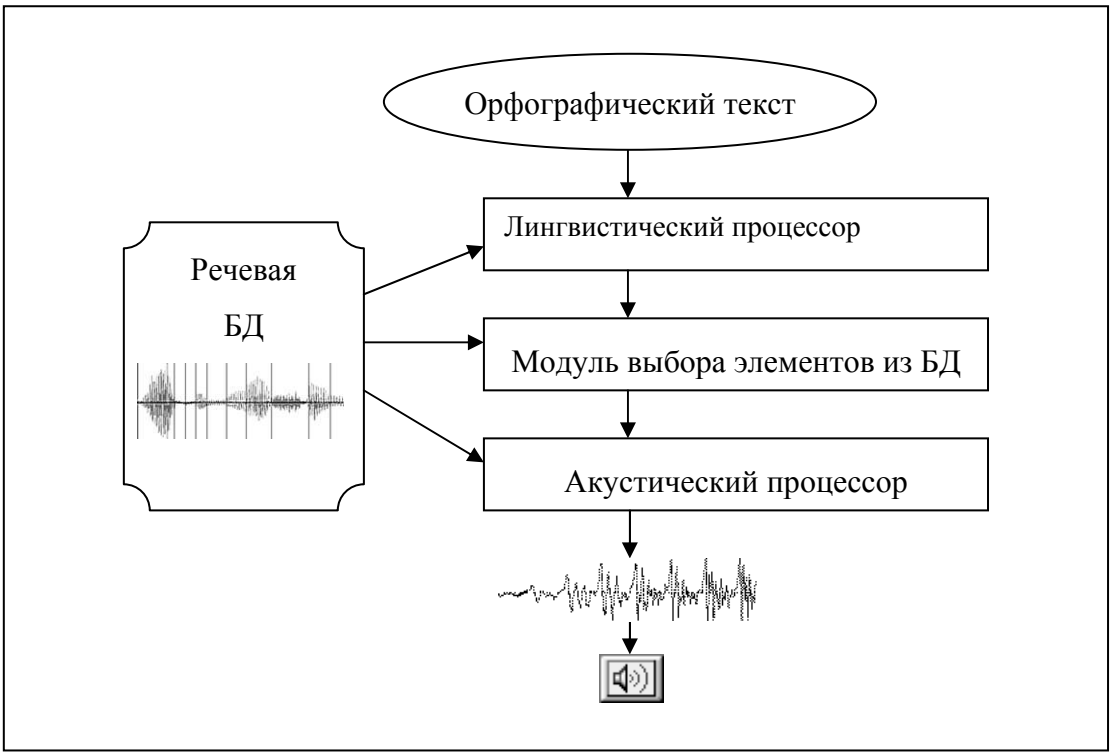

Рис. 1. Блок-схема конкатенативного синтезатора украинской речи за текстом

РЕЧЕВЫЕ БАЗЫ ДАННЫХ

Речевые БД являются существенной частью конкатенативного синтезатора. Элементами БД являются акустические прототипы речевых звуков (фонем-трифонов [5], аллофонов), то есть реализаций украинских фонем в различных фонетических контекстах.

Как для накопления и хранения БД, так и для конкатенации выбранных элементов используется представление речевых сигналов во временной области.

Для накопления речевых БД привлекались дикторы-непрофессионалы, прочитавшие вслух набор текстов и изолированных фраз. В настоящее время объем первой БЗ составляет около 12000 аллофонов, второй - около 3000. В БД могут храниться сегменты, соответствующие одному и тому же аллофону в одном о том же фонетическом, но в разных просодических контекстах (например, с разной интонацией). Аллофоны хранятся в БД в том порядке, в котором они находились в произнесенных диктором фразах. Эта информация учитывается при синтезе речи на этапе приминения критериев выбора аллофонов из БД.

Каждый аллофон БД отсегментирован на квазипериоды ЧОТ [5], что позволяет хранить не только сегментную (о контексте), но и подробную просодическую информацию.

При решении задачи разбиения речевого сигнала на квазипериоды использована следующая модель квазипериодичности и непериодичности сигнала во временном пространстве.

Сегмент сигнала $f_{n_{s-1}+j},\ j=0:(T_s-1), T_s=n_s-n_{s-1}$ является s-м одно-квазипериодическим сегментом длиной T_s, если он аппроксимируется в достаточной мере соседними, (s–1)-м или (s+1)-м сегментами. Причем, аппроксимированные величины $f^{-}_{n_{s-1}+j}$ та $f^{+}_{n_{s-1}+j}$ берутся с неизвестным амплитудным множителем α_s^- или α_s^+ соответственно:

$$f^{-}_{n_{s-1}+j}=\begin{cases}\alpha_s^- f_{n_{s-2}+j}, & j=0:(\min(T_s,T_{s-1})-1);\\ 0, & j=\min(T_s-1):(T_s-1),\end{cases}$$

$$f^{+}_{n_{s-1}+j}=\begin{cases}\alpha_s^+ f_{n_s+j}, & j=0:(\min(T_s,T_{s+1})-1);\\ 0, & j=\min(T_s,T_{s+1}):(T_s-1).\end{cases}$$

Введены априорные ограничения для множителей α, длительности текущего квазипериода T_s и его приращение $\Delta_s=T_s-T_{s-1}$:

$$\alpha_s:0<\alpha_{\min}\le\alpha_s\le\alpha_{\max},\ T_{\min}\le T_s\le T_{\max},\ |\Delta_s|\le\Delta_{\max}.$$

Построена элементарная мера квазипериодичности для s-го гипотетического одноквазипериодического сегмента $f_{n_{s-1}+j}=f_{n_s-T_s+j},\ j=0:(T_s-1)$, который проверяется на квазипериодичность путем сравнения со всеми возможными предыдущими сегментами $f_{n_s-2T_s+\Delta_s+j},\ j=0:(T_s-\Delta_s^--1),\ |\Delta^-|\le\Delta_{\max}$, и со всеми возможными последующими сегментами $f_{n_s+j},\ j=0:(T_s-\Delta^+-1),\ |\Delta^+|\le\Delta_{\max}$, и только наилучший результат сравнения ассоциируется со значением меры квазипериодичности $d((n_s,T_s),\Delta_s)$, а его аргумент Δ записывается в таблицу оптимальних управлений $\Delta(n,T)$.

Таким образом, поставленная задача сводится к задаче нахождения оптимальной траектории на графе, которая решается алгоритмически при помощи процедуры динамического программирования (ДП). При этом определяются начальные условия, элементарные меры квазипериодичности в узлах графа ДП, интегральный критерий и процедура восстановления траектории на графе ДП по таблице оптимальных управлений $\Delta(n,T)$.

Были исследованы несколько модификаций данного алгоритма, которые состояли в некоторых упрощениях самого алгоритма и фильтрации входящего сигнала.

В речевых БД отражены также особенности произношения разных дикторов. Описание каждого элемента БД состоит из: (1) идентификатора; (2) трехкомпонентного имени аллофона (имена предыдущей, текущей, последующей фонемы); (3) длительности в мс; (4) последовательности меток-границ квазипериодов ЧОТ.

ЛИНГВИСТИЧЕСКИЙ ПРОЦЕССОР

Лингвистический процессор, в свою очередь, состоит из четырех блоков: (1) блока нормализации текста; (2) акцентно-интонационного блока; (3) фонетического транскриптора; (4) блока вычисления просодических характеристик.

Текст, который предстоит озвучить, поступает на вход блока нормализации в орфографическом виде. Нормализация текста состоит в замене неалфавитных символов (например, цифр) и аббревиатур словами украинского языка. Акцентно-интонационный блок отвечает за расстановку ударений в словах и членение текста на интонационные группы. Фонетический транскриптор осуществляет преобразования буква -> аллофон фонемы.

Наконец, блок вычисления просодических характеристик для каждого аллофона фонемы задает значения длительности и ЧОТ, а затем преобразует фонетическую транскрипцию текста в фонетико-просодическую, в которой для гласных и звонких согласных аллофонов указаны длительность и значения ЧОТ, а для глухих согласных – только их длительность.

Выбор одной из имеющихся речевых БД влечет за собой настройку фонетического транскриптора и блока вычисления просодических характеристик на произношение, характерное для диктора-донора БД.

МОДУЛЬ ВЫБОРА ЭЛЕМЕНТОВ ИЗ БД

Данный модуль является центральным для описываемого метода синтеза речи. Алгоритм выбора элементов из БД использует:

- фонетико-просодическую транскрипцию входного текста;
- фонетико-просодические описания элементов речевой БД;
- таблицы фонемного сходства;
- фонетико-акустические критерии выбора.

Главным критерием выбора является контекстная идентичность искомого элемента (элемента транскрипции) и элемента-кандидата из БД. Имеются в виду правый и левый соседи элемента транскрипции. Вначале осуществляется поиск в БД аллофонов с идентичными правым и левым контекстами. Если такие элементы найдены, то отбор продолжается отдельно для гласных и звонких согласных и отдельно для глухих согласных элементов-аллофонов.

Для отбора гласных и звонких согласных используются просодические критерии. В частности, учитывается разность между средними длинами периода ЧОТ искомого и элемента-кандидата, соотношение количества периодов ЧОТ искомого элемента с

количеством периодов ЧОТ элемента-кандидата. Также учитывается непосредственное соседство выбранного на предыдущем шаге элемента и текущего кандидата в БД.

Следует отметить, что при наличии просодических критериев будет произведен выбор аллофона с приемлемой интонацией, что уменьшит необходимость изменения просодических характеристик акустическим процессором и улучшит качество синтезированной речи.

Для отбора глухих согласных используются такие критерии как разница в длительности и непосредственная близость выбранного на предыдущем шаге элемента и текущего кандидата в БД.

Если БД не содержит элементов в нужном сегментном контексте, то используются таблицы фонемного сходства, позволяющие подбирать для искомого элемента в качестве кандидатов элементы БД с отличным, но в фонетико-акустическом смысле подобным контекстом. При этом учитывается, что для гласных важнее левый контекст, а для согласных – правый.

АКУСТИЧЕСКИЙ ПРОЦЕССОР

Акустический процессор генерирует и озвучивает речевой файл, соответствующий входному тексту. Генерация речевого файла состоит в конкатенации выбранных оптимальных элементов речевой БД. При этом имеются две возможности: конкатенация элементов с их просодической модификацией в соответствии с вычисленными значениями длительности и ЧОТ и без просодической модификации, то есть с реальными длительностью и траекторией основного тона.

Просодическая модификация элемента БД состоит в повторении или пропуске определенных периодов ЧОТ, а также в укорочении или удлинении периодов ЧОТ с применением моделей линейного предсказания [5].

В основе этого метода лежит модель линейного предсказания сигнала, позволяющая по определенному количеству предыдущих отсчетов сигнала спрогнозировать последующие с довольно высокой точностью аппроксимации:

$$\tilde{f}_n = -\sum_{s=1}^{s=m} a_s f_{n-s} + \varepsilon_n ,$$

где $\tilde{f}_n$ — отсчеты прогнозируемого сигнала, f_n — отсчеты наблюдаемого сигнала, a_s, $s = 1{:}m$ — параметры предсказания, число которых m выбирается в пределах от 10 до 20, ε_n — погрешность прогнозирования. Параметры предсказания оцениваются на интервале анализа, равном одному или двум квазипериодам путем, например, минимизации суммы квадратов погрешности прогнозирования.

Просодическая модификация неизбежно вносит искажения в реальные речевые сегменты, что сказывается на естественности синтезированной речи. Важную роль в данном случае играет объем речевой БД: среди большого множества сегментно идентичных аллофонов с большей вероятностью будет обнаружен просодически подходящий аллофон.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Был проведен ряд экспериментов по восприятию синтезированной речи, в частности, исследовалось, насколько естественность синтезированной речи зависит от объема речевой БД и от просодической модификации последовательности элементов речевых БД, соответствующей входному тексту. Два текста, отличающиеся по стилю (художественная проза – 125 слов и радионовости – 116 слов) были озвучены синтезатором с использованием двух речевых БД: малого объема (712 аллофонов) и большого (5873 аллофонов), причем для второй БД оба текста были озвучены в двух вариантах: с просодической модификацией и без нее. Таким образом, для каждого из текстов были синтезированы три комбинации:

A) речевая БД малого объема, конкатенация с просодической модификацией;

B) речевая БД большого объема, конкатенация с просодической модификацией;

C) речевая БД большого объема, конкатенация без просодической модификации.

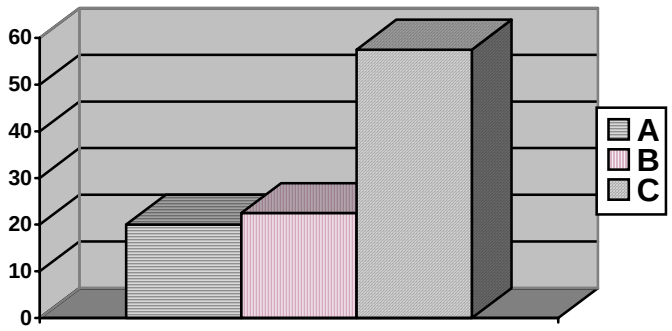

Рис. 2. Оценка натуральности звучания. Процент экспертов-аудиторов, отдавших предпочтение соответствующему варианту **A**, **B** или **C** синтезатора

В эксперименте принимали участие 20 аудиторов - носителей украинского языка из разных регионов Украины. Результаты эксперимента представлены на рис.2. Они свидетельствует о том, что наиболее естественной воспринимается речь, синтезированная с использованием БД большого объема без просодической модификации, наименее естественной – речь, синтезированная с использованием БД малого объема. Предпочтение варианта **C** можно объяснить тем, что слушатели склонны толерантно относится к перепадам спектральных и просодических характеристик при сохранении реального качества конкатенированных речевых сегментов.

ЛИТЕРАТУРА

1. Allen, J., Hunnicutt, M., Klatt, D. From text to speech. Cambridge University Press, 1987.
2. Лобанов Б.М., Минкевич В.В., Панченко Б.В. Устройство речевого вывода информации "Фонемофон-3" для ЭВМ // УСиМ, 1982, N 2. - С. 33 - 37.
3. O.F. Krivnova. Automatic synthesis of Russian speech // Proceedings of the XIV International Congress of Phonetic Sciences, Vol.1, pp. 507–510, San Francisco, 1999.
4. Hunt, A., and Black, A., "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," IEEE-ICASSP-96, Atlanta,Vol. 1. pp. 373-376, 1996.
5. Т. Вінцюк, Т. Людовик, М. Сажок, Р. Селюх. Автоматичний озвучувач українських текстів на основі фонемно-трифонної моделі з використанням природного мовного сигналу// Праці 6-ї Всеукраїнської міжнародної конференції “Оброблення сигналів і зображень та розпізнавання образів” – УкрОбраз’2002, Київ, 2002, с. 79–84.

Людовик Т.В., Сажок Н.Н.

КОНКАТЕНАТИВНЫЙ СИНТЕЗ РЕЧИ С ИСПОЛЬЗОВАНИЕМ РЕЧЕВЫХ БАЗ ДАННЫХ БОЛЬШОГО ОБЪЕМА

Синтезатор украинской речи предназначен для озвучивания произвольных орфографических текстов на украинском языке. В статье описывается структура синтезатора и используемый вариант конкатенативного метода синтеза речи, основанный на хранении, выборе и сглаженном склеивании предварительно записанных сегментов речи. Используются речевые БД большого объема, записанные разными дикторами. Фонетическая и просодическая информация, содержащаяся в БД, используется для поиска необходимых элементов БД.

Людовик Т.В., Сажок М.М.

КОНКАТЕНАТИВНИЙ СИНТЕЗ МОВЛЕННЯ З ВИКОРИСТАННЯМ МОВЛЕННЄВИХ БАЗ ДАНИХ ВЕЛИКОГО ОБСЯГУ

Синтезатор українського мовлення призначений для озвучування довільних орфографічних текстів українською мовою. У статті описується структура синтезатора та досліджена модифікація конкатенативного синтезу мовлення, що ґрунтується на збереженні, виборі та згладженому склаюванні попередньо записаних сегментів мовлення. Використовуються мовленнєві БД великого обсягу, записані з голосу різних дикторів. Фонетична і просодична інформація, що міститься у БД, використовується для пошуку необхідних елементів БД.

Lyudovyk T.V., Sazhok M.M.

CONCATENATIVE SPEECH SYNTHESIS USING LARGE SPEECH DATABASES

Ukrainian speech synthesizer is developed for pronouncing arbitrary orthographic texts in Ukrainian. The article deals with a general structure of the synthesizer and with concatenative speech synthesis method based on keeping, selection and concatenation of prerecorded speech segments. Large speech databases recorded by different donor speakers are used. Phonetic and prosodic information contained in a speech database is used for searching necessary database elements.