Синтез речи с моделированием особенностей произношения на основе анализа индивидуальных речевых баз данных большого объема

Т. В. Людовик

tetyana_lyudovyk@uasoiro.org.ua

1. Введение

В последнее время для синтеза речи по произвольному тексту широко используется конкатенативный метод, состоящий в соединении (склеивании) отрезков речевых сигналов, хранящихся в речевых базах данных (БД). Речевые БД различаются как размерами элементов (аллофоны, дифоны, полуфоны и т.д.), так и их количеством. Широко известны синтезаторы русской речи, основанные на конкатенации аллофонов [1, 2]; менее известны синтезаторы украинской речи, в которых предпочтение отдается дифонам (разработчики – Я.Козак, А.Черный).

В отличие от упомянутых синтезаторов, объем БД которых не превышает 2000 элементов, описываемый в данной статье синтезатор [3, 4] работает с базами данных большого объема (более 10000 элементов). Чем больше объем речевой БД, тем полнее представлена в ней звуковая, темпоральная и интонационная вариативность речи конкретного диктора-донора. Как следствие, синтезированная речь более «узнаваема» и в целом звучит более естественно. Кроме того, чем больше речевая БД, тем меньше искажения речевых отрезков возникающие при необходимом изменении длительности и частоты основного тона (ЧОТ) элементов БД. Иными словами, чем больше речевая БД, тем больше вероятность того, что в ней найдется элемент в необходимом контексте, с необходимой длительностью и контуром ЧОТ. Речевые БД не являются универсальными, они всегда индивидуальны и отражают особенности произношения конкретных дикторов.

Вопросы, связанные с разработкой речевых БД, а также с реализацией конкатенации, рассматриваются в [3]. Задача данной работы – анализ и моделирование индивидуальных особенностей произношения на сегментном и просодическом уровнях. Новизна предлагаемого подхода состоит в использовании фонетико-просодической аннотации речевой БД не только при выборе элементов конкатенации в процессе синтеза речи, но и на этапе настройки лингвистического процессора. Описываемый подход реализован в экспериментальной системе синтеза украинской речи.

2. Общая структура синтезатора речи

Экспериментальный синтезатор украинской речи по тексту состоит из традиционных для большинства конкатенативных синтезаторов модулей [1, 2, 5]: речевой БД, лингвистического и акустического процессоров, а также модуля выбора элементов конкатенации из БД (unit selection) [5 - 9], обязательного для синтезаторов, работающих с речевыми БД большого объема.

Центральную роль играет речевая БД. Информация, хранящаяся в ней (подробная аннотация акустических элементов, соответствующих контекстным аллофонам), используется всеми остальными модулями синтезатора. Предварительная настройка затрагивает все подмодули лингвистического процессора: блоки нормализации текста, просодической и аллофонной разметки, вычисления длительности и контура ЧОТ аллофонов (см. рис. 1).

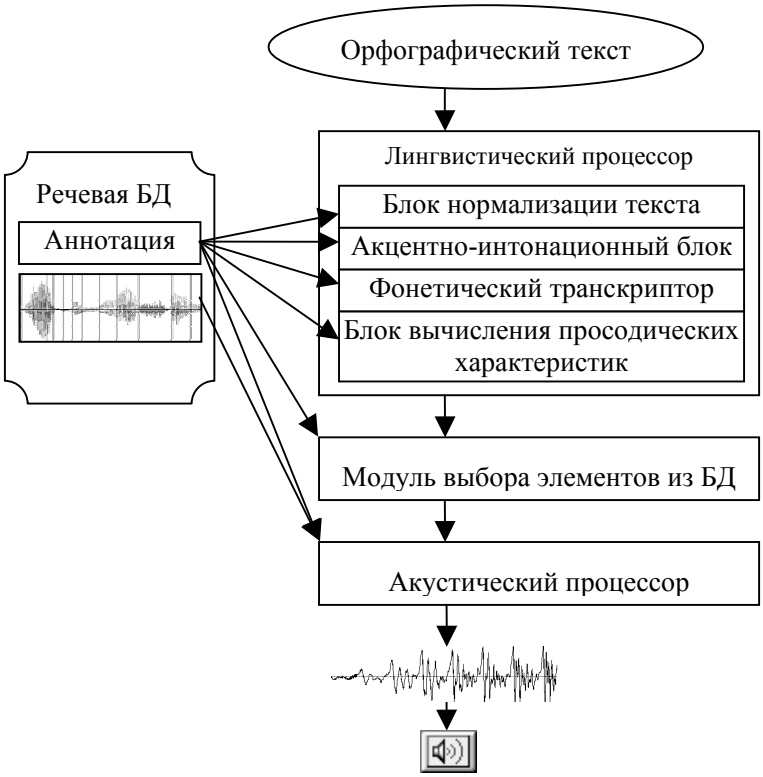

Рис. 1. Общая структура синтезатора речи.

Аннотация речевой БД используется также как главный источник информации непосредственно в процессе синтеза речи для выбора элементов конкатенации. Хотя алгоритм выбора элементов не зависит от конкретной речевой БД, используемые им критерии выбора формируются с учетом категорий аннотации (правый и левый контексты аллофона, его длительность и ЧОТ).

Отрезки речевых сигналов, хранящиеся в речевой БД, используются акустическим процессором для порождения результирующей речевой волны, соответствующей входному тексту [3].

3. Речевые базы данных

Качество синтезированной речи зависит от объема и представительности речевой БД, т.е. от того, насколько полно в ней представлено разнообразие значений сегментных и просодических характеристик. Не менее важными являются такие факторы как «качественный» диктор, качественные исходные акустические записи, качественное наполнение БД и качественная детальная аннотация. В данной работе особое внимание уделяется двум последним факторам.

3.1. Накопление и сегментация речевых баз данных

Для накопления речевых БД привлекались дикторы-непрофессионалы. Ими были прочтены вслух наборы тектов и изолированных фраз. Затем соответствующие акустическим файлам орфографические тексты обрабатывались программой автоматического транскрибирования с последующей коррекцией результатов вручную. Использовался следующий алфавит фонетических символов: 12 гласных (6 ударных и 6 безударных), 45 согласных (22 тведдых и 25 мягких) и пауза.

В соответствии с транскрипцией производилась сегментация акустического материала, и в базу данных записывались отрезки сигналов, соответствующие аллофонам, представленные во временной области. Затем каждый элемент БД, имеющий “по определению” квазипериодическую природу (гласные и звонкие согласные), подвергался автоматической сегментации на квазипериоды основного тона с последующей коррекцией вручную. Разработана методика, следуя которой человек, не являющийся экспертом в области речевых технологий, в состоянии создавать новые качественные речевые БД. В сотрудничестве с коллегами нами был проведен эксперимент, в ходе которого все операции, связанные с накоплением и сегментацией речевой БД, проводились автоматически. При этом сегментация речевых файлов в соответствии с фонетической транскрипцией (alignment) осуществлялась средствами HTK. В результате синтезированная речь (женский голос) получилась разборчивой, однако ошибки, связанные с неправильными границами между аллофонами, существенно ухудшили ее качество.

В речевых БД накопленные аллофоны хранятся в том порядке, в котором они находились в исходных речевых файлах.

3.2. Аннотация речевых баз данных

Аннотация, получаемая автоматически, является важным источником информации для всех модулей синтезатора речи. Во-первых, в ней отражены особенности произношения конкретного диктора, что важно для лингвистического процессора, задача которого – предсказать, как произнес бы некий входной текст именно данный диктор. Во-вторых, именно на аннотацию речевой БД опирается модуль выбора элементов

конкатенации из БД, получив «предсказание» от лингвистического процессора. Когда в базе данных насчитывается несколько десятков реализаций требуемого аллофона в одинаковом фонетическом контексте, выбор производится с учетом длительности и интонации. Следовательно, аннотация должна содержать детальное сегментное и просодическое описание элементов БД. Ниже приводится фрагмент аннотации одной из речевых БД. Каждая строка читается следующим образом: порядковый номер элемента БД, имя аллофона (трифона) с указанием левого и правого сегментных контекстов, длительность аллофона в мс, а также для квазипериодических аллофонов: количество квазипериодов ОТ, среднее значение длины квазипериода ОТ, длина начального, срединного и конечного квазипериодов ОТ.

```
2725   а - л - И    64.08   10   6.39   5.90   6.62   6.17
2726   л - И - с    80.32   15   5.35   6.08   5.31   5.03
2727   И - с - а   115.51
2728   с - а - г    68.21   11   6.17   5.53   6.17   6.94
2729   а - г - о    83.95   11   7.62   7.39   7.53   7.71
2730   г - о - л    55.42   7    7.89   7.53   7.98   8.30
2731   о - л - о    40.18   5    8.03   7.89   8.07   8.12
2732   л - о - в    62.54   8    7.80   6.98   8.89   8.16
2733   о - в - А    61.63   8    7.66   8.07   7.62   7.30
2734   в - А - #   159.41   27   5.90   7.17   5.90   4.90
```

В дальнейшем, при усовершенствовании синтезатора планируется использовать информацию о длине всех квазипериодов ОТ, что позволит оперировать контурами ОТ, отражающими малейшие нюансы микропросодики. Поскольку в известных нам синтезаторах значения ЧОТ определяются автоматически без ручной коррекции, то можно сказать, что мы используем наиболее точное описание просодии на акустическом уровне. Помимо использования для синтеза речи, эти данные могут служить материалом для экспериментально-фонетических исследований.

В настоящее время аннотация не отражает членения речевых отрезков на слоги и слова. Членение на интонационные группы с указанием их типов и главноударной гласной (несущей фразовое/логическое ударение) делается экспертом.

3.3. Отражение особенностей произношения в индивидуальных речевых базах данных

Речевые БД, в частности, их аннотации, отражают особенности индивидуального произношения дикторов-доноров. В них отражены также диалектные особенности и особенности стиля произношения. Эти особенности проявляются на всех уровнях:

- на уровне нормализации текстов при чтении (ср. «эт коммерческое» /«собака» (рус.) в адресах электронной почты);
- на уровне просодической разметки (например, преобладание длинных/коротких интонационных групп, ударение на разных слогах слов, допускающих эту вариативность);

- на уровне аллофонной разметки (например, наличие/отсутствие определенных типов фонетической ассимиляции, различная степень редукции);
- на уровне ритмики и темпа речи (например, различная длительность пауз);
- на уровне интонации (например, предпочтение определенных типов интонационных групп, различная стратегия интонационного акцентирования (повышения/понижения тона) предъядерных акцентных групп).

4. Предварительная настройка лингвистического процессора на индивидуальную речевую базу данных

Задача лингвистического процессора в аллофонном конкатенативном синтезаторе речи состоит в преобразовании входного орфографического текста в фонетико-просодическую транскрипцию, т.е. последовательность аллофонов с указанием их длительности и контура ЧОТ. Чем точнее фонетико-просодическая транскрипция будет отражать речь диктора-донора, тем качественнее и естественнее будет синтезированная речь.

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

Ниже приведен фрагмент фонетико-просодической транскрипции, поступающей из лингвистического процессора в модуль выбора элементов из речевой БД. Для глухих согласных и паузы указана их длительность, а для гласных и звонких согласных – последовательность длин квазипериодов основного тона (в мс).

```
#-в-е   7.4   7.4   7.4   7.5   7.5   7.5   7.5   7.5   7.6   7.6
в-е-л   7.6   7.6   7.6   7.6   7.6   7.7   7.7
е-л-И   7.7   7.7   7.6   7.6   7.6   7.5   7.5   7.5   7.4   7.4
л-И-к   7.4   7.4   7.3   7.3   7.2   7.2   7.0   6.9   6.8   6.7   6.7   6.7
И-к-а   90
к-а-р   6.7   6.8   7.0   7.1   7.3   7.4   7.6   7.8   7.9
```

Лингвистический процессор использует два типа данных. Независимые от диктора данные являются общими для всех носителей данного языка. К ним относятся алфавит фонетических символов и набор признаков аллофонов. Зависимыми от диктора данными являются: 1) словари, используемые блоком нормализации текстов; 2) максимальная длина (в фонетических словах) интонационной группы, используемая акцентно-интонационным транскриптором; 3) правила фонетической ассимиляции и редукции, используемые фонетическим транскриптором; 4) средние длительности аллофонов, коэффициенты изменения длительности, используемые блоком вычисления просодических характеристик; 5) инвентари интонационных контуров, также используемые блоком вычисления просодических характеристик.

Рассмотрим некоторые зависимые от диктора данные на примере двух речевых БД (мужские голоса): максимальная длина интонационной группы – соответственно 7 и 8 фонетических слов, средние длительности ударных гласных [А] – 117 мс и 95 мс, [О] – 102 мс и 84 мс, [Е] – 104 мс и 84 мс, [У] – 94 мс и 78 мс. На рис.

2 приведено графическое изображение интонационных контуров незавершенности, наблюдаемых у этих же дикторов. В обоих случаях интонационная группа состоит из трех акцентных групп.

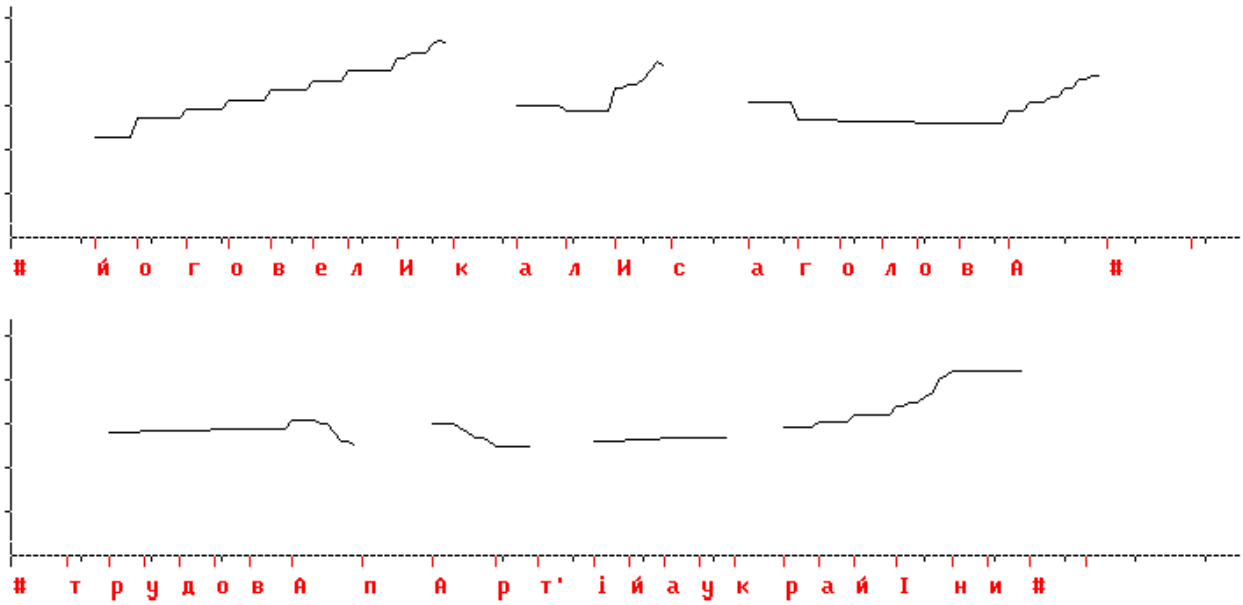

Рис. 2. Графическое изображение интонационных контуров незавершенности, наблюдаемых у разных дикторов.

В настоящее время полностью автоматически на речевые БД настраивается подмодуль, связанный с вычислением длительности аллофонов. Вычисление длительности аллофонов исходит из того, что каждая фонема имеет базовую (среднюю) длительность и набор коэффициентов изменения длительности. Значения базовых длительностей и коэффициентов длительности у каждого диктора свои. Они могут быть разными и у одного диктора при различных стилях произношения.

5. Выбор элементов из речевой базы данных в процессе синтеза речи

Алгоритм выбора аллофонов из речевой БД основывается на фонетических и просодических критериях. В нем используются:

- фонетико-просодическая транскрипция входного орфографического текста, поступающая из лингвистического процессора;
- фонетико-просодическое описание (аннотация) элементов речевой БД;
- таблицы фонемного сходства.

Главным критерием выбора в настоящее время является, в соответствии с фонемно-трифонной моделью [10], контекстная идентичность искомого элемента (элемента фонетико-просодической транскрипции) и элемента-кандидата из БД. Для каждого элемента фонетико-просодической транскрипции вначале

осуществляется поиск в БД элементов с идентичными правым и левым контекстами. Если такие элементы найдены, то продолжение отбора осуществляется отдельно для элементов, имеющих квазипериодическую природу (гласные и звонкие согласные) и не имеющих ее (глухие согласные и пауза). Для отбора гласных и звонких согласных используются просодические критерии. В частности, учитывается разность между средними длинами квазипериода основного тона искомого и аллофона-кандидата, соотношение количества квазипериодов искомого элемента с количеством квазипериодов элемента-кандидата. Для отбора глухих согласных в качестве критерия используется разность длительностей. Еще одним критерием выбора является непосредственное соседство в речевой БД аллофона-кандидата и аллофона, выбранного из БД на предыдущем шаге. Если речевая БД не содержит элементов с нужными левым и правым сегментными контекстами, то используются таблицы фонемного сходства для поиска элементов-кандидатов с наиболее близкими контекстами. При этом учитывается, что для гласных важнее левый контекст, а для согласных – правый.

Следует отметить, что наличие просодических критериев позволяет выбирать аллофоны с приемлемой интонацией и длительностью, что улучшает качество синтезированной речи. Для речевых БД среднего объема уменьшается необходимость изменения просодических характеристик акустическим процессором. Для речевых БД большого объема необходимость изменения длительности и ЧОТ может вообще отпадать, поскольку велика вероятность, что будут найдены просодически подходящие элементы.

Модуль выбора элементов из речевой БД преобразует фонетико-просодическую транскрипцию входного текста в акустико-фонетико-просодическую, отличающуюся от фонетико-просодической указанием на порядковые номера элементов в БД, а также на то, какие квазипериоды в каком количестве (для увеличения или сокращения длительности аллофона) и с какой длиной (для изменения контура основного тона) следует взять для конкатенации. Ниже приведен фрагмент акустико-фонетико-просодической транскрипции:

```
4020  #-в-е  (1 7.4) (2 7.4) (3 7.4) (4 7.5) (5 7.5) (7 7.5) (8 7.5) (9 7.5) (10 7.6) (11 7.6)
5031  в-е-л  (1 7.6) (2 7.6) (3 7.6) (4 7.6) (6 7.6) (7 7.7) (8 7.7)
3888  е-л-И  (1 7.7) (2 7.7) (2 7.6) (3 7.6) (4 7.6) (4 7.5) (5 7.5) (6 7.5) (7 7.4) (7 7.4)
4884  л-И-к  (1 7.4) (2 7.4) (3 7.3) (4 7.3) (5 7.2) (6 7.2) (7 7.0) (7 6.9) (8 6.8) (9 6.7) (10 6.7) (11 6.7)
2723  И-к-а  0  90
785  к-а-р  (1 6.7) (2 6.8) (3 7.0) (5 7.1) (6 7.3) (7 7.4) (8 7.6) (10 7.8) (11 7.9)
```

6. Выводы

В данной работе сделан шаг на пути полностью автоматического «создания» голосов для конкатенативного синтеза речи. Последовательная реализация подхода, основанного на анализе большого объема речевых данных, позволяет легко и быстро настраивать синтезатор речи на голоса и манеру произношения разных людей, на разные стили произношения, а также, в известной степени, на разные языки.

Помимо практического использования в конкатенативном синтезаторе речи полученные данные о сегментных и просодических характеристиках речи разных дикторов являются полезным материалом для экспериментально-фонетических исследований.

Что касается естественности и узнаваемости синтезированной речи, то важными определяющими факторами (не считая качества исходного речевого материала) являются:

- объем речевой БД и ее представительность, отражающая сегментную и просодическую вариативность;
- информационная емкость и качество аннотации речевой БД;
- учет лингвистическим процессором индивидуальности произношения;
- алгоритм выбора элементов из речевой БД в процессе синтеза речи.

7. Литература

1. 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.

2. Лобанов Б.М., Карневская Е.Б., Левковская Т.В. Синтезатор речи по тексту как компьютерное средство «клонирования» персонального голоса // Тр. Международной конференции Диалог-2001 / М., 2001. С. 265-272.

3. Людовик Т.В., Сажок Н.Н. Использование речевых баз данных большого объема при синтезе речи в системах искусственного интеллекта // Проблемы управления и информатики, 2003, №6. С. 82-87.

4. Lyudovyk, T., Sazhok, M. Unit Selection Speech Synthesis Using Phonetic-Prosodic Description of Speech Databases // Proceedings of the 9-th International Conference Speech and Computer SPECOM'2004, St.Petersburg, Russia, (to appear).

5. M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T Next-Gen TTS System // The Proceedings of the Joint Meeting of ASA, EAA, and DAGA, Berlin, Germany, 1999, pp.18-24.

6. Hunt, A., Black, A. W. Unit selection in a concatenative speech synthesis system using a large speech database // Proceedings of the IEEE International Conference on Acoustics and Speech Processing, Munchen, Germany, 1996, Vol. 1, pp. 373-376.

7. Coorman, G., Fackrell, J., Rutten, P., Van Coile, B. Segment Selection in the L&H RealSpeak Laboratory TTS System, // Proceedings of the International Conference on Spoken Language Processing, Vol. 2, 395-398, Beijing, China, 2000.

8. Conkie, A. Robust unit selection system for speech synthesis // Proceedings of the Joint Meeting of ASA, EAA, and DAGA, 18-24, Berlin, Germany, 1999.

9. Beutnagel, M., Conkie, A., and Syrdal, A. Diphone synthesis using unit selection // Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, 185-190, Jenolan Caves, Australia, 1998.

10. Vintsiuk, T., Lyudovyk, T., Sazhok, M., Selyukh, R. Automated player of Ukrainian texts based on phoneme-threephone model with natural signal involved // Proceedings of the 6th All-Ukrainian Conference “Signal/Image Processing and Pattern Recognition, Kyiv, 2002, pp. 79–84 (in Ukrainian).