Особенности использования многоуровневой разметки звукового корпуса unit selection в системе гибридного синтеза «Живой голос»[1]

Multi-tier markup of speech corpus for hybrid russian tts system «Vitalvoice»

Продан А. И. (prodan@speechpro.com),
Корольков Е. А. (korolkov@speechpro.com),
Опарин И. В. (ilya@speechpro.com),
Таланов А. О. (andre@speechpro.com)

ООО «Центр речевых технологий», Санкт-Петербург, Россия

Рассматривается система многоуровневой разметки звукового корпуса и её использование в системе гибридного синтеза «Живой голос» ООО «Центр речевых технологий» (ЦРТ). Система включает в себя взаимосвязанные уровни разметки, создаваемые и используемые как независимо друг от друга, так и в комплексе.

1. Введение

Система многоуровневой разметки речевого корпуса фонограмм используется при подготовке речевых данных для синтезатора речи по тексту «Живой голос» ЦРТ [1, 2].

Существует несколько подходов к организации автоматического синтеза речи по тексту. К основным можно отнести синтез по правилам (формантный синтез), артикуляторный синтез, компилятивный синтез, синтез на основании статистических моделей (HMM-синтез). На данный момент наилучшие результаты достигаются с использованием технологии Unit Selection. Данная технология позволяет достичь максимальной естественности синтезированной речи. В рамках работы по созданию новой системы синтеза русской речи, осуществляемой ЦРТ, создан синтезатор на основе использования технологии Unit Selection, совмещенной с аллофонным синтезом. Гибридный характер системы позволяет осуществлять масштабирование всей системы синтеза в зависимости от доступных ресурсов. Полноценный синтез Unit Selection, обеспечивающий наилучшее качество синтезированной речи, предполагается использовать на стационарных компьютерах; для мобильных решений предложен компромисс между качеством звучания и используемыми ресурсами памяти при помощи технологии аллофонного синтеза.

Одна из основных особенностей системы синтеза «Живой голос» — совмещение положительных сторон двух подходов — Unit Selection и компилятивного аллофонного синтеза. Таким образом, для синтеза каждым голосом подготавливаются два звуковых корпуса: аллофонный, в котором хранятся аллофоны во всех возможных контекстах, и репрезентативный речевой корпус для выбора звуковых единиц методом Unit Selection. В статье рассматривается система многоуровневой разметки корпуса для Unit Selection.

Характерной особенностью синтеза методом Unit Selection является его критическая зависимость от состава и полноты речевого корпуса. Качественный синтез возможен только на основе полного, сбалансированного и корректно размеченного речевого корпуса.

С ростом объема корпуса достигается темповая и интонационная вариативность речи диктора. Иными словами, чем больше корпус, тем больше вероятность того, что в нем найдется элемент в необходимом контексте с необходимой длительностью и контуром частоты основного тона (ЧОТ). Как следствие, меньше искажения от вынужденной модификации сигнала, а значит — выше естественность синтезируемой речи.

В целом, использование корректно размеченного, сбалансированного корпуса, есть необходимое условие для достижения высокого качества синте-

[1] Работа выполнена в рамках федеральной целевой программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2012 годы»

зируемой речи. Известно, что качество синтеза Unit Selection не является постоянной величиной и зависит от состава синтезируемого текста. Такое свойство заложено в самой технологии. Действительно, когда выходной сигнал синтезатора составлен из оригинальных (немодифицированных) крупных фрагментов непрерывной речи, то качество речи практически совпадает с естественной. С другой стороны, когда требуется синтезировать речь по тексту, фрагменты которого представлены в корпусе лишь отдельными аллофонами, то в этом случае качество синтеза решающим образом определяется составом уровней разметки и точностью установки границ на этих уровнях. Ошибки в составе и границах обычно приводят к тому, что никакие дальнейшие усилия, связанные с модификацией речевого сигнала не в состоянии сделать синтезированную речь близкой к естественной.

2. Уровни разметки речевого корпуса Unit Selection

Основным принципом разметки корпуса является возможность учета всей информации, которая может потребоваться для синтеза. На разных уровнях разметки присутствуют как обозначения сегментных единиц: аллофонов, слогов, слов, пауз и их характеристик, так и информация более высокого уровня — об интонационном оформлении синтагм и отдельных слов, отметки об эмоциональной составляющей, и выделение неречевых явлений: смеха, кашля, заполненных пауз хезитации и т. п.

Такая система показала свою эффективность при подборе наиболее подходящих звуковых единиц, т.е. с наименьшими величинами стоимости замены и стоимости связи [3–6]: при необходимости по такой разметке из речевого корпуса можно извлечь любые единицы с заданными характеристиками.

Разметка корпуса производится в два этапа. На первом — разметка выполняется вручную с использованием специализированного звукового редактора Wave Assistant, каждый уровень хранится в отдельном текстовом файле. На рис. 1 представлен пример окна с сигналом в программе Wave Assistant, с принятыми уровнями разметки:

На втором этапе, разметка выполняется автоматически, при этом часть корпуса, размеченная вручную, используется для обучения акустических моделей системы автоматической разметки.

Далее рассмотрим каждый уровень и возможности его использования в системе синтеза более подробно.

2.1. Уровень разметки периодов основного тона

Первый и самый низкий уровень разметки — уровень периодов основного тона (ОТ). На нём каждый период основного тона обозначен метками, благодаря которым точно известна частота основного тона аллофона и скорость её изменения. Любая из меток может иметь специальный идентификатор, с помощью которого выделяются особые свойства периода. Отмечаются периоды, которые по каким-либо причинам (щелчок, какой-то посторонний короткий шум) лучше не использовать. Характер-

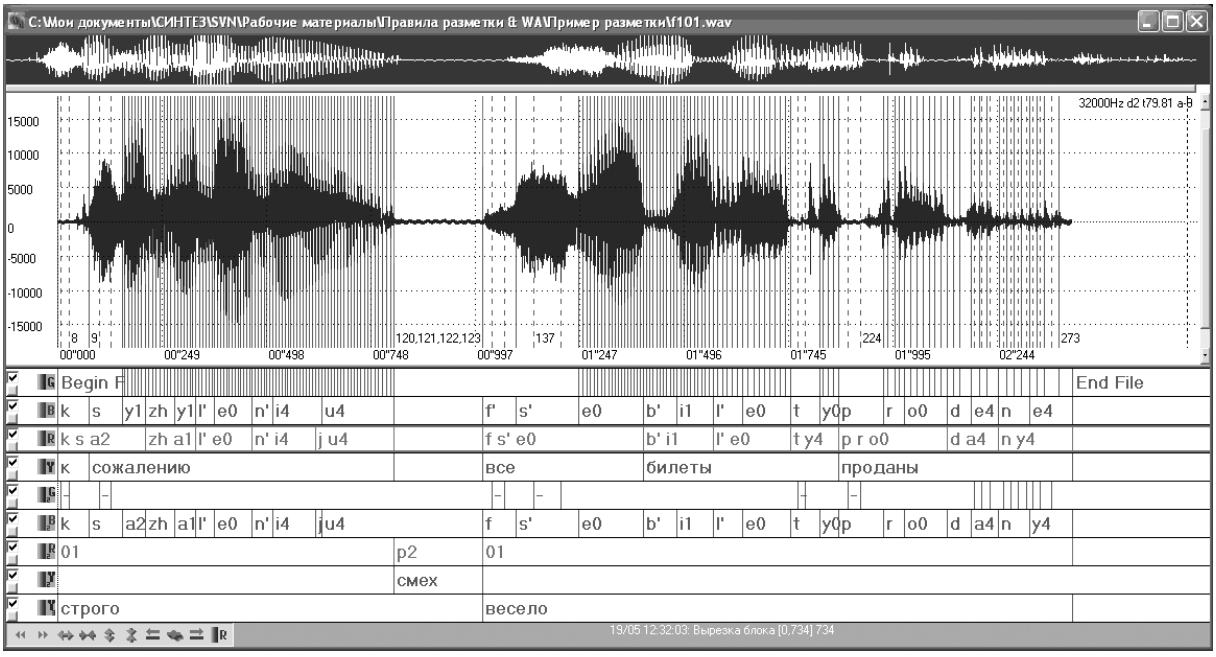

Рис. 1. Пример окна с сигналом, размеченным на всех используемых уровнях разметки

ные для звука периоды, которые при модификации речевого сигнала нельзя удалять или повторять, так же отмечаются специальным идентификатором.

Первый период после паузы и последний перед паузой также отмечаются особыми знаками, эти периоды захватывают часть паузы, благодаря чему звук при синтезе не обрывается, а естественным образом начинается или заканчивается.

На этом же уровне отмечается начало участка оглушения у звонких согласных, чтобы фрагмент после метки не принимался за обычный период ОТ, а так же, чтобы при выборе аллофона учитывалось, что согласный частично оглушён.

2.2. Уровень меток, использующихся для модификации речевого сигнала

Несмотря на то, что технология Unit Selection подразумевает выбор из звукового корпуса максимально длинных и хорошо стыкующихся между собой цепочек аллофонов без модификации, всё же в некоторых случаях минимальная модификация исходного речевого сигнала становится необходимой. Например, если длительность аллофона значительно (больше заданного порога) превышает предсказанную, подсчитанную с учётом средней длительности аллофона в корпусе, или наоборот значительно меньше её, включается модификация по длительности и этот аллофон соответственно укорачивается или удлиняется. Для этого необходимы специальные пометки, обозначающие относительно стационарные участки аллофона, которые можно сократить или наоборот частично повторить. Модификация производится и в том случае, если аллофон по частоте основного тона значительно отличается от соседних в выбранной цепочке. Обычно, если используется аллофон из аллофонного корпуса, то модификация по частоте ОТ нужна практически всегда.

Метки возможной модификации речевого сигнала используются для того, чтобы обозначить те зоны, в которых можно модифицировать исходный сигнал по длительности или частоте ОТ. Такие метки ставятся на двух уровнях: для модификации вокализованных звуков используются описанные выше метки на периодах основного тона и метки оглушения на уровне меток основного тона, а для глухих согласных или оглушённых частей звонких согласных используются дополнительные метки возможной модификации по длительности на специальном уровне.

Кроме того, на этом уровне есть возможность отметить аллофоны, которые по каким-либо причинам лучше не использовать для синтеза или те, которые лучше не брать по отдельности, а только в контексте их непосредственных соседей (например, сильно оглушённые звонкие согласные). Также на этом уровне устанавливаются другие специальные отметки, характеризующие способ сочетания элементов.

2.3. Уровни реальной и идеальной транскрипции

С целью совместить в себе возможность выбора длительных последовательностей аллофонов — при совпадении целых словосочетаний или даже фраз — и точный выбор необходимых более коротких цепочек в один-два аллофона для транскрипции используются сразу два уровня: уровень реальной и уровень идеальной транскрипции. На уровне реальной транскрипции устанавливаются реальные границы аллофонов и обозначаются именно те аллофоны, которые произнёс диктор. На уровне же идеальной транскрипции находятся аллофоны идеальной транскрипции в таком составе и порядке, в каком их генерирует автоматический транскриптор синтезатора. Метки идеальной транскрипции ставятся в соответствие меткам реальной. Если в действительности звук не реализован, то он отмечается только в идеальной транскрипции.

При значительном несовпадении — не соответствующее предсказанному место ударения, оговорка диктора — исправления вносятся в идеальную транскрипцию так, чтобы каждому аллофону на уровне идеальной транскрипции соответствовали только его возможные варианты произнесения на уровне реальной (и пропуск).

Кроме того, на уровне реальной транскрипции отмечаются такие явления, как назализованный нейтральный гласный в начале или конце фразы.

Поиск требуемых аллофонов производится по идеальной транскрипции, это позволяет найти максимально длинные последовательности аллофонов для заданного текста, причём отличия между реальной и идеальной транскрипцией учитываются в разных случаях с большим или меньшим весом. В том случае, когда требуется взять отдельный аллофон, то сразу идёт проверка по уровню реальной транскрипции.

2.4. Уровень слов

На уровне разметки слов устанавливаются метки на границах слов, идентификатором метки является само слово в орфографической записи. Данный уровень разметки создается автоматически с использованием текста, прочитанного диктором. Уровень используется для подбора аллофона по одному из параметров при расчете стоимости замены.

2.5. Уровень слогов

Уровень разметки слогов также генерируется автоматически. Слова делятся на открытые слоги. Уровень разметки на слоги используется при расчете стоимости замены. Учитывается положение нужного аллофона в слоге, количество слогов от начала синтагмы и число слогов до конца синтагмы.

2.6. Уровень интонации и пауз

В системе синтеза «Живой голос» используется список основных мелодических типов и их вариантов в звучащем тексте, созданный в СПбГУ на кафедре фонетики и методики преподавания иностранных языков [7]. В качестве основы взята расширенная классификация типов интонации Е.А.Брызгуновой [8]. Всего выделяется 13 типов интонационных конструкций с различными подтипами в зависимости от различных мелодических типов, места синтагматического ударения и т. д.

Случаи логического и эмфатического ударения, а также переноса синтагматического отмечены специальными знаками на уровне слов, что добавляет точности интонационной разметке.

В модели различается шесть типов пауз, в зависимости от завершённости или незавершённости предшествующей синтагмы, знаков препинания в исходном тексте и т.п.

Таким образом, информацию об интонационном оформлении можно получить как для отдельного аллофона, так и для всей синтагмы нужного типа в целом. Это полезно при выборе конкретного варианта интонационного оформления синтагмы: как именно он был произнесён диктором и в то же время даёт возможность получить «усреднённый» по всему корпусу вариант реализации того или иного подтипа интонационной модели. Это придаёт синтезированной речи большую естественность за счёт разнообразия её интонационного оформления.

2.7. Уровни разметки эмоциональной окрашенности

Для разметки эмоциональных модальностей и различных неречевых явлений, которые могут понадобиться для повышения естественности синтезированной речи, предназначены ещё два уровня системы разметки. На первом отмечаются явления смеха, кашля, причмокивания и т.п. На втором отмечаются эмоциональные модальности. Локализованные эмоции выделяются метками внутри синтагмы, если эмоция нелокализованная, то она задается для всей синтагмы целиком.

3. Система проверки разметки звуковой базы

Звуковой корпус размечается как вручную, так и автоматически. Лучшим компромиссом является автоматическая разметка, подстроенная под определённого диктора, то есть обученная на части материала, размеченной вручную [9]. Но ни ручная, ни автоматическая разметка никогда не дают стопроцентной точности и правильности. Неизбежны опечатки, неточно установленные границы, просто случаи, где необходимо указать вручную какие-либо особенности произнесения (для автоматической разметки) или какие-либо другие нестандартные ситуации.

Множество подобных ошибок можно найти автоматически. В ЦРТ специально для этого была разработана программа «MarkupChecker». При помощи неё проверяются на допустимость названия меток на разных уровнях и соответствия между ними. Программа не только даёт указания на явные ошибки, но также предупреждает о местах, где по каким-либо причинам ошибка является вероятной.

В данный момент автоматическая проверка ведётся по следующим параметрам:

- Проверка соответствия идеальной транскрипции в корпусе и транскрипции, полученной на выходе автоматического транскриптора, используемого в составе синтезатора.
- Проверяется наличие необходимых уровней разметки для звукового файла.
- На каждом уровне разметки производится проверка на допустимость присутствующих там обозначений (по списку).
- Производится проверка на соответствие меток начала слов и меток начала аллофонов, меток начала синтагм и меток начала слов, меток уровней идеальной и реальной транскрипции.
- Производится проверка на наличие зон возможной модификации по длительности для глухих согласных и оглушённых участков звонких.
- Производится проверка на наличие разметки по периодам основного тона для гласных и звонких согласных.
- Производится проверка на резкие изменения по длине периодов основного тона (слишком длинные или короткие периоды по сравнению с соседними).
- Производится проверка на наличие отметок пауз на уровнях слов и интонации (по соответствию меткам конца аллофона).
- Производится проверка на наличие метки в начале фразы.
- Производится проверка на наличие в слове нескольких ударных гласных и правильной расстановки степеней редукции по отношению к ударению.
- Производится проверка на допустимую разницу идеальной и реальной транскрипции (по подгружаемой таблице вариативности).
- Удаляются лишние пробелы в идентификаторах меток.
- Выдаются предупреждения о слишком больших зонах оглушения звонких звуков.

Как видно из приведённого выше списка, бóльшая часть ошибок выделяется именно благодаря сопоставлению различных уровней разметки.

4. Выводы

Предложенный способ разметки звукового корпуса для системы гибридного синтеза с использованием технологии Unit Selection является исчерпывающим, поскольку для каждой звуковой единицы корпуса и её соседей будь то период основного тона, аллофон, синтагма или фраза, обеспечивается доступ к информации на всех уровнях в комплексе. Данный способ разметки в сочетании с разработанным набором параметров для расчета минимальной стоимости замены и связи звуковых элементов позволяет получить высокое качество синтезируемой речи.

Литература

1. *Корольков Е.А., Главатских И. А., Таланов А. О., Киселев В. В., Опарин И. В.* Синтез естественной русской речи при помощи метода Unit Selection // Материалы XXXVI Международной филологической конференции.
2. *Oparin I., Talanov A.* Outline of a New Hybrid Russian TTS System // Proc. of the 12th International conference on Speech and Computer, SPECOM 2007, Moscow, Russia, 2007. Pp. 603–608.
3. *Black A. W., Hunt A. J.* Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database // In Proceedings of ICASSP 96. Atlanta, Georgia, 1996. Vol. 1, pp. 373–376.
4. *Vepa J.* Join Cost for Unit Selection Speech Synthesis // University of Edinburgh, 2004.
5. *Clark R. A. G., Richmond K., King S.* Multisyn: Open-domain unit selection for the Festival speech synthesis system // Speech Communication, 2007. Vol. 49, issue 4, pp. 317–330.
6. *Vepa J., King S.* Subjective evaluation of join cost functions used in unit selection speech synthesis // In Proceedings of the International Conference on Speech and Language Processing 2004. Jeju, Korea, 2004. Pp. 1181–1184.
7. *Вольская Н. Б., Скрелин П. А.* Моделирование интонации для синтеза речи по тексту // Уфа: 1998.
8. *Брызгунова Е. А.* Интонация // Русская грамматика. М.: 1980.
9. *Tatham M, Morton K.* Developments in Speech Synthesis // John Wiley & Sons Ltd, 2005.