СИСТЕМЫ СИНТЕЗА РЕЧИ

УДК 81'322.6

А. И. Соломенник, П. Г. Чистиков, С. В. Рыбин,
А. О. Таланов, Н. А. Томашенко

АВТОМАТИЗАЦИЯ ПРОЦЕДУРЫ ПОДГОТОВКИ НОВОГО ГОЛОСА ДЛЯ СИСТЕМЫ СИНТЕЗА РУССКОЙ РЕЧИ

Предложены методика и средства автоматизации процедуры создания голоса заданного диктора для работы в системе синтеза речи VitalVoice. Реализованный алгоритм автоматизированной подготовки голоса включает несколько этапов: выбор текстового материала, запись речи с оперативным контролем параметров записи, создание размеченной звуковой базы, настройка параметров подбора элементов.

Ключевые слова: *синтез речи, создание голоса, автоматическая разметка, дифон, корпус текстов.*

Введение. Технология синтеза речи по тексту давно интересует исследователей всего мира. Существуют разные способы получения речевого сигнала: синтез по правилам (формантный синтез), артикуляторный синтез, компилятивный синтез, синтез на основе статистических моделей (HMM-синтез). Синтез методом Unit Selection (выбора элементов, US) [1], подготовка нового голоса для которого составляет предмет настоящей статьи, является одним из видов компилятивного синтеза. Суть его состоит в том, что синтезированная речь компилируется не из базы специально записанных элементов (аллофонов, дифонов, трифонов, полуфонов, слогов и т. п.), каждый из которых представлен единственным вариантом, а из произнесенных предложений естественного языка, и для каждого элемента из множества выбирается наиболее подходящий вариант. Данный метод позволяет достичь очень высокой естественности синтезированной речи. Однако качественный синтез возможен только на основе полного, сбалансированного и корректно размеченного речевого корпуса. С целью разметки речевой базы для метода US в ООО „ЦРТ“ была разработана специальная многоуровневая система [2]. Добавление нового голоса является нетривиальной задачей для любой системы компилятивного синтеза, так как требует записи новой звуковой базы, из которой подбираются элементы, составляющие синтезированную речь. В особенности это актуально для синтеза методом US, поскольку звуковая база для качественного синтеза голоса должна быть достаточно велика (до нескольких часов звучащей речи) [3]. Именно поэтому важно максимально автоматизировать процесс добавления голоса.

В 2010 г. в ООО „ЦРТ“ разработана специальная подсистема [4], на основе которой позднее было создано приложение VoiceConstructor — программа, позволяющая создавать голоса для системы синтеза русской речи VitalVoice [5]. Программа состоит из модулей подготовки текстов, записи фонограмм и формирования звуковой базы голоса.

В **модуле подготовки текстов**, разработанном специально для русского языка, создаются фонетически сбалансированные корпуса текстов заданного размера. Самый простой

способ получить все необходимые для синтеза элементы — записать большую базу данных речи (десятки часов). Но просто наличия большого объема записанной речи недостаточно, корпус должен быть сбалансированным и по возможности полным, т.е. содержать все необходимые единицы во всех возможных контекстах с различными возможными характеристиками, такими как акустические параметры, частота основного тона, длительность, позиция в слоге и т.п. Но так как для создания базы данных нужна сегментация, которая обычно требует по крайней мере некоторой ручной коррекции после автоматической сегментации, размер базы данных влияет на время, необходимое для подготовки ее к использованию. Кроме того, большие базы данных неудобны для хранения и поиска в них. Таким образом, должен соблюдаться баланс между размером и репрезентативностью данных.

Существует целый ряд исследований по автоматическому созданию текстовых корпусов для различных языков [6, 7]. Для русского языка в работе [8] описывается схожий алгоритм. Главное преимущество метода, рассмотренного в работе [9], состоит в том, что он обеспечивает удобство создания текстовой базы, давая возможность не просто выборки предложений из большого корпуса текстов, но позволяет выбрать тип звуковой единицы корпуса, заранее создать и редактировать необходимые корпусы текстов. Модуль автоматической подготовки текстового корпуса был создан на основе программы анализа статистики фонетических единиц [10].

Работа с системой начинается с указания параметров создаваемой базы данных. Пользователь должен выбрать тип основной единицы: дифон или аллофон, установить среднюю скорость речи и желаемый размер базы. Программа показывает текущие размеры базы данных, текста и корпуса. Программа работает с четырьмя корпусами текстов: базовым, включающим в себя наборы частотных и специфических фраз (алфавит, числа, аббревиатуры и т. п.); пользовательским, в который можно загрузить тексты, необходимые для использования в системе синтеза (например, для чтения объявлений в торговом центре имеет смысл ввести примерные тексты объявлений, которые будут подаваться на синтез); фонетическим, который формируется путем выбора предложений из исходного корпуса так, чтобы максимально включить в тексты необходимые для синтеза звуковые единицы (дифоны или трифоны), если их не хватает в базовом и пользовательском корпусах: исходным корпусом, из которого набираются предложения для фонетического корпуса.

Алгоритм генерации фонетического корпуса включает в себя следующие этапы. Во-первых, система транскрибирует все необходимые тексты. Затем вычисляется необходимый объем фонетического корпуса с учетом данных об общем желаемом размере корпуса и размере основного и пользовательского корпусов (если таковые имеются). Предложения выбираются из исходного корпуса в зависимости от количества отсутствующих в создаваемом корпусе единиц, которые они содержат, предложения с максимальным количеством отсутствующих единиц берутся в первую очередь. Если два предложения содержат одинаковое число таких единиц, предложение с менее частотными дифонами будет взято в первую очередь. Также учитывается длина предложения (предпочтение отдается более коротким). Для редких дифонов процедура выбора такая же, она запускается, когда все дифоны исходного корпуса уже присутствуют в основном и пользовательском корпусах. Подбор предложений заканчивается, когда текст достигнет желаемого размера, причем в тот момент, когда в корпус уже добавлены все отсутствующие дифоны, выдается соответствующее предупреждение. Далее на запись подаются предложения из первых трех корпусов.

Модуль записи фонограмм. На этом шаге производится запись звуковых файлов для выбранных текстовых корпусов. Каждое предложение записывается в отдельный файл. Перед проведением сеанса записи требуется измерить шум канала (в паузе). Превышение заданного значения отношения сигнал/шум отмечается индикатором, предупреждающим о том, что следует изменить условия записи, иначе качество создаваемого голоса может оказаться неудов-

летворительным. Аналогичные индикаторы имеются для уровня и для энергии записываемого речевого сигнала. Процесс записи фонограммы контролируется в режиме реального времени с помощью двух графиков: траектории частоты основного тона и осциллограммы сигнала. Траектория частоты основного тона измеряется автокорреляционным методом. Диктор читает предложение за предложением из текущего списка. В любой момент любое предложение можно перезаписать и продолжить запись.

Модуль формирования звуковой базы голоса. На этом шаге производится разметка звуковых файлов, для того чтобы при синтезе из базы голоса выбирались нужные элементы. Метки хранятся в отдельных текстовых файлах, просмотр и корректировка размеченных файлов производятся в звуковом редакторе WaveAssistant. Для формирования базы голоса необходимо получить разбивку на периоды частоты основного тона (ЧОТ) и аллофонную сегментацию.

Для выполнения разметки по ЧОТ в программе WaveAssistant реализован автокорреляционный метод расчета основного тона с предварительной фильтрацией и постобработкой с целью уточнения положения меток основного тона (ОТ). Низкочастотная фильтрация используется для снижения ошибки определения ОТ путем удаления из сигнала составляющих с частотой выше 500 Гц. Высокочастотная предварительная фильтрация используется для определения участков, на которых нет ОТ (невокализованные звуки). Постобработка положения меток позволяет удалять „слишком частые“ или „слишком редкие“ метки, уточнять положение меток в сложных случаях, когда метки смещаются в ту или другую сторону.

Аллофонная сегментация выполняется автоматически с помощью модулей системы распознавания речи (ASR) с использованием HMM (скрытых марковских моделей). Сегментация проводится на основе выравнивания (force alignment) транскрипции и звукового сигнала, она состоит из трех этапов: обучение акустических моделей; сегментация и автоматическая корректировка границ аллофонов. На первом этапе строятся акустические модели монофонов, так как именно монофоны наилучшим образом подходят для данной задачи. Качество сегментации улучшается, если для каждого диктора имеется достаточное количество данных, чтобы обучить индивидуальные модели. Если данных для построения индивидуальных акустических моделей недостаточно, при сегментации используются либо общие акустические модели, построенные по большой базе (более 50 дикторов), либо строятся модели с использованием данных тех дикторов, голоса которых по своим акустическим характеристикам близки к целевому голосу. На шаге сегментации получаются два варианта —„идеальная“ сегментация, которая в точности соответствует заданной транскрипции, и „реальная“ — отличающаяся от первой более точным акустическим соответствием с фонограммой. Оба варианта сегментации в дальнейшем используются при синтезе речи. Заключительный этап автоматической сегментации заключается в автоматической корректировке полученных на предыдущем этапе границ аллофонов на основе дополнительной информации (разметка ЧОТ и правила, составленные на основе статистического анализа систематических неточностей).

Затем выполняется фильтрация звука с целью выравнивания материала по энергии и уменьшения возможной реверберации на глухих участках согласных. Во время сборки базы получаемая для диктора статистическая информация по длительности и амплитуде аллофонов записывается и затем используется при настройке параметров подбора элементов. В зависимости от пола и возраста диктора уточняются настройки различных параметров элементов. Затем пользователю предлагается запустить инсталляцию голоса, по ее завершении новый голос появляется в списке установленных голосов.

Заключение. Рассмотренная методика автоматизированного создания голоса была опробована на речевом материале различного объема (от нескольких минут до 10 часов речи). Она позволила получить практически важные результаты: при минимальной ручной корректировке разметки достигнута почти полная разборчивость речи и практически стопроцентная

узнаваемость исходного диктора даже на базах необходимого объема (от получаса звучащей речи). Реализованный модуль выбора текстового корпуса позволил при том же объеме базы получить большую аллофонную вариативность, что также позволило улучшить получаемую синтезированную речь.

СПИСОК ЛИТЕРАТУРЫ

1. *Black A. W., Hunt A. J.* Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database // Proc. of ICASSP 96. Atlanta, Georgia, 1996. Vol. 1. P. 373—376.
2. *Продан А. И., Корольков Е. А., Опарин И. В., Таланов А. О.* Особенности использования многоуровневой разметки звукового корпуса Unit Selection в системе гибридного синтеза „Живой голос“ // Матер. Междунар. конф. „Диалог“. 2009. С. 415—419.
3. *Black A. W.* Perfect Synthesis for all of the people all of the time // Keynote. IEEE TTS Workshop. Santa Monica, CA, 2002. P. 146—170.
4. *Продан А. И., Таланов А. О., Чистиков П. Г.* Система подготовки нового голоса для системы синтеза „Живой голос“ // Матер. Междунар. конф. „Диалог“. 2010. С. 394—399.
5. *Oparin I., Talanov A.* Outline of a New Hybrid Russian TTS System // Proc. of the 12th Intern. Conf. on Speech and Computer. SPECOM 2007. Moscow, Russia, 2007. P. 603—608.
6. *Chevelu J., Barbot N., Boeffard O., Delhay A.* Comparing set-covering strategies for optimal corpus design // Proc. of the 6th Intern. Language Resources and Evaluation. 2008. P. 2951—2956.
7. *van Santen J. P. H., Buchsbaum A. L.* Methods for optimal text selection // Proc. of Eurospeech. Rhodes, Greece, 1997. P. 553—556.
8. *Кривнова О. Ф., Захаров Л. М., Строкин Г. С.* Подбор текстового материала и статистический инструментарий для создания речевых корпусов // Сб. тр. XI сессии Российского акустического общества. Т. 3. Акустика речи. Медицинская и биологическая акустика. М.: ГЕОС, 2001. С. 87—92.
9. *Solomennik A. I., Chistikov P. G.* Automatic generation of text corpora for creating voice databases in a Russian text-to-speech system // Матер. Междунар. конф. „Диалог“. 2012. С. 607—615.
10. *Смирнова Н. С., Чистиков П. Г.* Программа анализа фонетических статистик в текстах на русском языке и ее использование для решения прикладных задач в области речевых технологий // Матер. Междунар. конф. „Диалог“. 2011. С. 632—643.

Сведения об авторах

Анна Ивановна Соломенник — ООО „Речевые технологии“, Минск; научный сотрудник; E-mail: solomennik-a@speechpro.com

Павел Геннадьевич Чистиков — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: chistikov@speechpro.com

Сергей Витальевич Рыбин — канд. физ.-мат. наук; ООО „ЦРТ“, Санкт-Петербург; ведущий программист; Санкт-Петербургский национальный исследовательский университет информационных технологий, кафедра речевых информационных систем; доцент; E-mail: rybin@speechpro.com

Андрей Олегович Таланов — канд. техн. наук; ООО „ЦРТ“, Санкт-Петербург; руководитель отдела синтеза речи; E-mail: andre@speechpro.com

Наталья Александровна Томашенко — ООО „ЦРТ“, Санкт-Петербург; младший научный сотрудник; E-mail: tomashenko-n@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.