АВТОМАТИЧЕСКИЙ ПОИСК КЛЮЧЕВЫХ СЛОВ В НЕПРЕРЫВНОМ ПОТОКЕ РЕЧИ НА ОСНОВЕ ТЕХНОЛОГИИ "РАСПОЗНАВАНИЕ ЧЕРЕЗ СИНТЕЗ"

AUTOMATIC KEYWORD SPOTTING IN CONTINUOUS SPEECH USING RECOGNITION-BY-SYNTHESIS TECHNIQUE

В.В. Киселёв *(kiselev-v@speechpro.com)***,**

А.О. Таланов *(andre@speechpro.com)***,**

И.Б. Тампель,

М.Ю. Татарникова,

Ю.Ю. Хохлов
ООО "Центр речевых технологий", Санкт-Петербург

Проблема поиска ключевых слов в непрерывной речи актуальна в системах безопасности, телефонных сервисах, а так же в ряде прикладных задач. В статье описана система, разработанная на основе динамического программирования и синтезатора речи. Однопроходный метод позволяет достичь высоких процентов поиска, с относительно небольшой ошибкой ложных срабатываний.

Введение

Метод поиска ключевых слов (*KWS: Keyword Spotting*) является одним из эффективных способов автоматического поиска необходимого фрагмента фонограмм в огромных звуковых базах или звуковых потоках. Хотя концепция *KWS* не нова, применимость её в коммерческих целях для систем реального времени возможна лишь с недавних пор. Практическое применение таких «поисковиков» может быть в системах национальной безопасности, телефонных сервисах, системах контроля качества, системах речевых фильтров (например, фильтр нецензурной лексики) и в ряде других.

Существует несколько подходов к методам поиска ключевых слов [1]:

- *KWS основанные на распознавании слов;*
- *KWS основанные на распознавании последовательности фонем;*
- *KWS основанные на распознавании слитной речи;*

При этом, методы сопоставления речевых сигналов может быть как метод динамического программирования (впервые предложенный Bridle [2] в качестве метода поиска ключевых слов), так и статистические метод скрытых Марковских моделей [3]. В данном докладе представлен метод, основанный на распознавании изолированных слов с использованием динамического программирования.

Концепция *KSW* имеет ряд особенностей по сравнению с методами, применяемыми в областях распознавания речи систем управления голосом. При использовании технологии поиска приходится иметь дело со слитной речью двух или более дикторов. Разговор может идти на произвольные темы, которые могут иметь неограниченный словарный состав (при этом язык диктора может меняться). В этом случае для поиска по шаблонному методу необходимо иметь эталоны произвольного характера.

Главной особенностью разработанной системы является возможность ввода ключевого слова с клавиатуры. Традиционные же системы предлагают сначала надиктовать и подготовить массив звуковых данных, которые в дальнейшем будут служить эталонами ключевых слов. Неудобство такого подхода заключается в том, что зачастую невозможно собрать достаточное количество звуковых данных, а если речь идёт о поиске в спонтанной речи произвольного слова или фразы, сделать такое в принципе не представляется возможным. Вместо этого, впервые для русского языка, авторы разработанной системы предложили использовать синтезатор речи.

Общие описание системы

Разработанная система поиска состоит из нескольких главных блоков (рис. 1): подготовка первичных акустических моделей с помощью синтезатора, процесс обучения с формированием базы эталонов и сравнение эталонов с входным акустическим потоком.

В компании "*Центр речевых технологий*" вот уже несколько лет успешно развивается система синтеза русской речи [4,5]. Качество звучания оценивается как высокое и приближается к натуральному. Множество

синтезированных голосов позволяет сформировать достаточное количество синтезированных сигналов для формирования массива акустических данных. Ключевые слова в текстовом виде, введённые пользователем, поступают на вход синтезатору.

Для каждого ключевого слова (или фразы) синтезатор автоматически определяет место ударения и формирует основные просодические модели. Далее, для каждой модели автоматически формируется набор параметров громкости, длительности и основного тона. В целом, для каждого ключевого слова генерируются 15 различных сигналов, синтезированных каждым голосом. Набор синтезированных голосов и модификация каждого из них позволяет сформировать достаточный набор звуковых данных для успешного обучения ключевого слова.

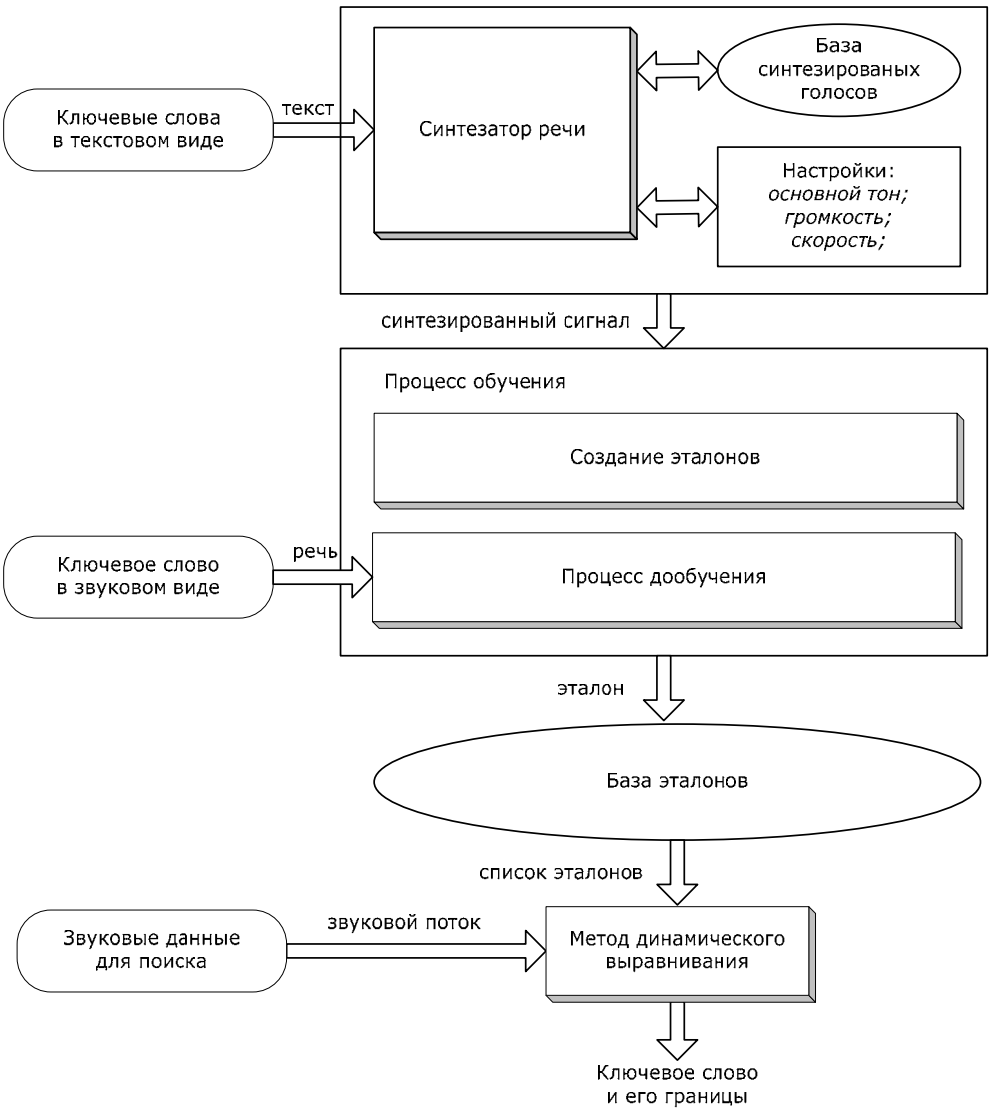

Рис. 1 Общая схема системы

Для описания эталонов ключевых слов и преобразования анализируемого речевого сигнала в системе использован алгоритм получения первичных признаков, основанный на использовании гребенки БИХ фильтров второго порядка:

$$y_n = x_n + C_i y_{n-1} - r^2 y_{n-2}$$

где C_i - коэффициент фильтра с номером i, r – радиус полюса (один для всех фильтров гребенки).

Центральные частоты фильтров рассчитываются по формуле: $f_i = F \cdot ar\cos(C_i/(2 \cdot r))/(2 \cdot \pi)$,

где F – частота квантования.

Получение данного набора признаков реализуется следующим образом. Входной речевой сигнал оцифровывается с частотой дискретизации 11025 кГц или 8 кГц. Положение окна анализа определяется по максимуму огибающей в диапазоне интервалов времени, характерных для периодов основного тона, соответствующих частотам основного тона от 80 Гц до 300 Гц. Сигнал, выделенный в окне, дифференцируется, и

к нему применяются описанные фильтры.

Выход каждого фильтра домножается на коэффициент усиления, обеспечивающий соответствие общей огибающей характеристикам слуховой системы человека. Далее выполняется сглаживание во времени спектральных компонент фильтром низких частот второго порядка и получение вектора признаков для каждого окна анализа путем попарного объединения соседних спектральных компонент. Таким образом, получается первичное описание эталонов ключевых слов и входного анализируемого сигнала набором векторов признаков размерности 8: энергия и 7 спектральных компонент.

Подмодуль дообучения организован в качестве дополнительной опции и предназначен для добавления к эталону ключевых слов звукового сигнала (например, заранее подготовленный сигнал, вводимый через микрофон или телефонную линию). Таким образом, система накапливает эталоны ключевых слов в специальной базе и предоставляет возможность быстро и эффективно извлекать их для поиска.

Сравнение фрагментов входного речевого сигнала и эталонов ключевых слов.

Сравнение входного потока речевого сигнала и эталонов ключевых слов выполняется известным методом нелинейного временного выравнивания (динамического программирования, ДП) с использованием взвешенной Евклидовой метрики.

Входной поток речевого сигнала преобразуется в последовательность векторов признаков и просматривается окном длиной равной максимальной длине эталона ключевого слова с шагом, соответствующим средней длине звука русской речи при быстром темпе произнесения. Выделяемое окно анализируется на значение “речь/не речь” на основе анализа превышения спектральных компонент порога шумового окружения и сравнения максимума огибающей с пороговым значением.

Далее вычисляется расстояние от неизвестного речевого фрагмента до эталонов ключевых слов. При использовании метода ДП для вычисления расстояния, определяется путь, соответствующий минимальному накопленному расстоянию при прохождении между начальными и конечными точками неизвестного и эталонного образа.

Пусть сравниваются два образа, которые описываются последовательностью векторов:

$X = \{\bar{x}_0, \bar{x}_1, ..., \bar{x}_i, ..., \bar{x}_M\}$ и $Y = \{\bar{y}_0, \bar{y}_1, ..., \bar{y}_i, ..., \bar{y}_M\}$

Различие между векторами двух образов определяется последовательностью состояний C_k и обозначается:

$$F() = C_0, C_1, ..., C_k, ..., C_K$$

где:

C_0 - начальное состояние,

C_K - конечное состояние,

$F()$ - функция временного выравнивания, которая проектирует временную область одного образа на временную область другого образа.

Метод ДП заключается в том, что ищется такая функция F, при которой путь из состояния C(0) в состояние C(K), будет оптимальным, и при этом будет найдено минимальное накопленное расстояние между двумя образами. При построении оптимального пути, на каждом шаге алгоритма используется основная формула ДП:

$$D(\bar{x}_i, \bar{y}_j) = \min \begin{matrix} D(\bar{x}_i, \bar{y}_{j-1}) + d(\bar{x}_i, \bar{y}_j), \\ D(\bar{x}_{i-1}, \bar{y}_{i-1}) + d(\bar{x}_i, \bar{y}_j), \\ D(\bar{x}_{i-1}, \bar{y}_j) + d(\bar{x}_i, \bar{y}_j), \end{matrix}$$

где $0 <= i <= M$, $0 <= j <= N$, и

$$d(\bar{x}, \bar{y}) = \sum_{k=1}^{N\_SEC} w \cdot (x_k - y_k)^2$$

взвешенная Евклидова метрика, где *x* и *y* вектора, принадлежащие сравниваемым образам, *N_SEC* – размерность векторов признаков.

Для определения расстояния в конечной точке состояний необходимо вычислить матрицу расстояний между последовательностями векторов. Это требует больших вычислительных мощностей и объема памяти. На практике обычно последовательно вычисляют две строки матрицы, используя для вычисления следующей строки значения предыдущей.

Полученные расстояния от текущего окна до группы эталонов ключевых слов анализируются. Ищется ближайшее расстояние и сравнивается с порогом.

Пороговое значение, используемое в решающем правиле можно изменять, тем самым регулируя процент ошибок. Первое приближение порогового значения получается в результате анализа результатов распознавания тестовой выборки. При увеличении порогового значения увеличивается количество ложных срабатываний и уменьшается вероятность того, что правильное слово не будет распознано.

Результаты исследования

Для оценки любой *KWS* системы необходимо использовать несколько мер, а именно:

Правильное определение (Correct spotting) – процент правильно найденных ключевых слов;

Ложный отказ (False Rejection) – процент ложного отказа ключевых слов;

Ложное срабатывание (False Alarm) – процент принятия ложных слов в качестве ключевых;

Разработанная система KWS тестировалась в два этапа: 1. тестирование микрофонной записи; 2. тестирование в телефонном канале. Рассмотрим последовательно эти этапы.

Тестирование микрофонной записи проводилось по файлам, которые содержали технический текст, начитанный дикторами мужчинами через микрофон и оцифрованный через стандартную звуковую карту. Поиск осуществлялся по трем ключевым словам: «Незабудка», «Фонограмм», «Каналов ДМА». Было поставлено два эксперимента поиска ключевых слов:

1. По одному эталону на ключевое слово для дикторов подбиралось соответствующее значение основного тона. Была выставлена соответствующая анализируемому сигналу громкость эталона.
   Результаты:
   Для коэффициента ложной тревоги 9.3%, правильность распознавания составила – 83%, ошибка – 17%;
2. Для каждого ключевого слова создавался набор эталонов с различными параметрами уровня громкости, темпа и основного тона. Результаты:
   Для коэффициента ложной тревоги 17.5%, правильность распознавания составила – 100%, ошибка – 0%.

Под коэффициентом ложной тревоги понималось значение отношения количества слов, ложно распознанных как искомая ключевая команда, к общему количеству слов, произнесенных в анализируемом сегменте, за исключением общего количества встречаемости искомого ключевого слова в анализируемом сегменте:

$$FA = \frac{N_{FA}}{N_{total} - N_{KW}}$$

где:

FA - коэффициент ложной тревоги;

N_{FA} - количество ложных срабатываний;

N_{total} - общее количество слов, произнесенных в анализируемом сегменте;

N_{KW} - общее количество встречаемости ключевой команды в анализируемом сегменте;

При тестировании телефонного сигнала анализировались звуковые файлы, приблизительно по 7 мин., с записанным разговором на заданную тему. Ключевые слова составлялись отдельно для каждого диктора (таблица 1).

Номер диктора	*Ключевые слова*
01	«Ленинграде» «Центре Речевых технологий»
02	«Сталин» «Профессор» «Синий цвет» «Первая демонстрация»
03	«День космонавтики» «Эйзенхауэр» «Аризоне» «Рузвельт»

Таблица 1 Список ключевых слов

Для каждого ключевого слова создавался набор эталонов с различными параметрами уровня громкости, темпа и основного тона.

Результаты:

Правильность распознавания составила 78.5% (ошибка 21.5%) для коэффициента ложной тревоги 60%. Большое количество ложных срабатываний обусловлено тем, что телефонный сигнал сильно отличается от полученных синтезированных сигналов (в этом случае нужна специальная дополнительная нормировка). Также телефонный разговор характеризуется спонтанной речью, в отличие от микрофонного, где речь воспроизводиться в стиле диктанта.

Преимущества разработанной системы и дальнейшая работа

При разработке *KWS* систем реального времени необходимо учитывать ряд факторов, которые являются наиболее важными при практическом применении. Перечислим основные их них:

Фактор *продолжительность работы*, является одним их критических для конечных пользователей. Надёжная *KWS* система должна работать 24 часа в сутки и уметь обслуживать параллельно несколько каналов. Время обработки одного звукового потока не должно быть больше времени его звучания.

К фактору *надёжности*, можно отнести такие свойства *KWS* системы как устойчивость, дикторонезависимость, неограниченный набор ключевых слов.

Гибкость или возможность настройки для *KWS* систем необходимы для редактирования ключевых слов, настроек различных порогов, процентов перекрытия ключевых слов и ряд других собственных параметров.

Поддержка многоязычности. При разговоре одного или нескольких дикторов возможна смена языка говорящего. Поэтому достаточно важным является поддержка поиска ключевых слов на нескольких языках, в рамках одной и той же *KWS* системы.

Разработанная система удовлетворяет практически всем перечисленным требованиям. Так, продолжительность работы системы не ограничена временными рамками и показывает надёжную, стабильную работу. Количество входных ключевых слов так же не ограниченно, однако разработчики рекомендуют использовать список, состоящий менее из 100 ключевых слов. Набор параметров, предоставляемых для изменений, делает систему гибкой и адаптируемой.

В качестве дальнейшего развития предполагается добавлять синтезированные голоса, а так же реализовать систему мультиязычного синтеза, с поддержкой множества языков. Одним из вариантов практического применения системы поиска является ее применение в телекоммуникационных комплексах, поэтому для более надёжного поиска в таких комплексах необходима специальная нормировка синтезированного сигнала. Проведённые исследования показали, что результаты поиска сильно зависят от входных параметров синтезированного сигнала. Лучшие результаты были получены с помощью ручной корректировки входных параметров (как было показано выше), таких как громкость, тон, тембр. Градация этих параметров на различные уровни (например, для тона: низкий, средний, высокий) позволили развить ещё одно направление – автоматическую подстройку синтезатора под входной акустический поток. Предполагается, что алгоритм будет анализировать входной сигнал для извлечения наиболее оптимальных характеристик.

Однопроходный метод позволяет реализовать достаточно быстрый поиск. Так для одного ключевого слова время поиска составляет менее 5 % от длины записи.

Программная реализация и внедрение

Разработанная система поиска ключевых слов реализована в виде отдельного SDK, включающая динамические библиотеки, заголовочные файлы, примеры использования, демонстрационную программу и документацию. По техническим характеристикам для успешной работы SDK необходимо 100 Mb свободного дискового пространства на жестком диске, не менее 128 Mb оперативной памяти и процессор с частотой не ниже 1кГц.

Демонстрационная программа (Рис. 2) выполнена в виде Win 32 приложения, и предназначена для демонстрации основных возможностей SDK. Пользователю предлагается набрать ключевые слова и выбрать источник поиска. Результаты поиска формируются в таблице, с указанием границ ключевых слов и процента достоверности (высчитывается относительно порога). Графический, интуитивно понятный интерфейс делает систему простой в и использовании и тестировании работы SDK.

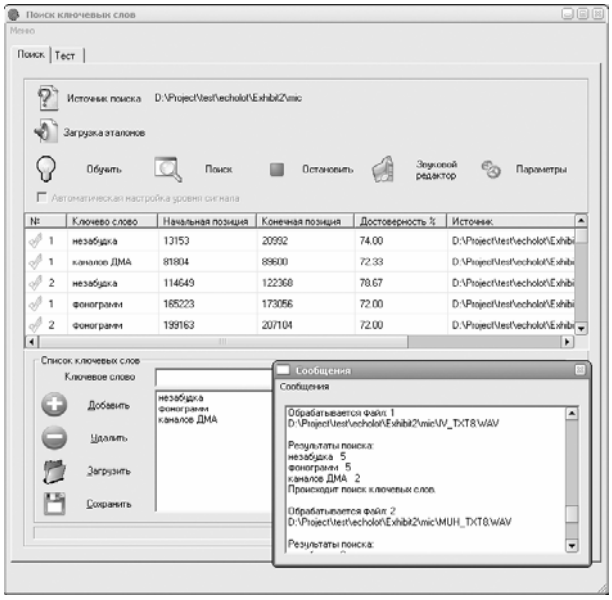

Рис. 2 Демонстрационная программа

Реализованное программное обеспечение (SDK) было внедрено в коммерческую систему многоканальной записи, регистрации и архивирования звуковых сигналов, Незабудка II [4].

Список литературы:

1. Szoke I., Schwarz P., Matejka P., Burget L., Karafiat M., Fapso M. and Cernocky J. *"Comparison of Keyword Spotting Approaches for Informal Continuous Speech", //* Proceedings of InterSpeech 2005, *September 4-8 2005 Lisbon, Portugal p. 633 – 636*

2. J. S. Bridle. *An efficient elastic template method for detecting given words in running speech. // In* Brit. Acoust. Soc. Meeting*, pages 1-4, 1973.*

3. R. Rohlicek, W. Russell, S. Roukos, and H. Gish, *"Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting," // 1989* IEEE ICASSP*, pp. 627-630.*

4. http://www.speechpro.ru/

5. Вольская Н., Коваль А., Коваль С., Опарин И., Погарева Е., Скрелин П., Смирнова Н., Таланов А. *"Синтезатор русской речи по тексту нового поколения" // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог'2005" (Звенигород, 1-6 июня, 2005 г.)/ Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П. Селегея. - М.:Наука, 2005., стр. 234-237.*