АРТИКУЛЯТОРНО-ОРИЕНТИРОВАННАЯ СИСТЕМА РАСПОЗНАВАНИЯ РЕЧИ

В.Н.Сорокин, А.Н.Ижнин, А.И.Цыплихин, Д.Н.Чепелев
Институт проблем передачи информации РАН

Аннотация

Несмотря на достигнутые успехи в распознавании больших словарей, метод скрытых марковских моделей не обеспечивает необходимую устойчивость к вариациям дикторских голосов, помех и искажений. Для борьбы со стационарными помехами и искажениями речевого сигнала необходимо использовать его динамические свойства, а остальные виды изменчивости могут быть нейтрализованы за счет кодовых свойств речи. Особенности динамики речи выделяются системой первичных детекторов спектрально-временных неоднородностей речевого сигнала. Сегментация речевого потока осуществляется детекторами артикуляторных событий, т.е. переходов из одного артикуляторного состояния в другое. Процесс распознавания реализован в виде процедуры списочного декодирования по критерию максимума апостериорной вероятности. Система распознавания обучалась на базе данных для изолированных, раздельных и слитных произнесений числительных, произносимых 47 дикторами, причем одновременно записывались сигналы от двух разных типов микрофонов. Использовалось 5 типов микрофонов и телефонных трубок, а также имитатор телефонного канала, что значительно уменьшает риск настройки на условия формирования базы данных. Система еще не завершена в разработке. Ее предварительные характеристики: 12% ошибок на базе данных, около 10% ошибок и 6% переспросов при распознавании в реальном времени, причем задержка решения - менее 70 мс.

1. Введение

Почти все известные системы распознавания основаны на использовании скрытых марковских моделей речевого сигнала с синхронным отсчетом векторов спектра или кепстра во времени. Привлекательность метода скрытых марковских моделей для разработчиков состоит в его формализованности, практически не требующей участия экспертов-речевиков, и исключающей длительные исследования свойств речевого сигнала. Несмотря на то, что этот метод имеет несомненное преимущество по сравнению с устаревшим методом динамического преобразования оси времени, и демонстрирует довольно высокую надежность распознавания, он практически достиг предела своих возможностей, все еще не обеспечивая необходимую устойчивость к вариациям дикторских голосов, помех и искажений. Это определяется тем, что метод скрытых марковских моделей может непосредственно компенсировать лишь изменчивость длительности сегментов речевого сигнала. Остальные же виды изменчивости, такие как

- реверберация помещений,
- разнообразные амплитудно-частотные характеристики каналов связи (включая разные типы микрофонов, расстояние до микрофона и его направленность),
- нестационарные шумы,
- стиль речи диктора,
- граничные эффекты в слитной речи,
- контекстная изменчивость элементов речи,

относятся в этом методе к случайным искажениям. Такая неадекватность модели речевого сигнала усугубляется неизбежной непредставительностью статистической выборки для обучения марковских цепей.

Метод скрытых марковских моделей неустойчив по отношению к шумам и типу микрофона, в результате чего в реальных условиях надежность распознавания падает до 40 - 60%. Но даже и наилучшие показатели этого метода для фиксированных условий обучения и распознавания, слегка превышающие 90% словесной надежности распознавания, не дают оснований для утверждения о решении проблемы распознавания речи, поскольку фраза из 5 слов будет распознана лишь примерно в 50% случаев. Ограниченность метода марковских цепей и необходимость в принципиально новых подходах хорошо осознается исследователями (Lippmann, 1997; Hermansky, 1998).

Очевидно, что для борьбы со стационарными помехами и искажениями речевого сигнала необходимо использовать его динамические свойства, а остальные виды изменчивости могут быть нейтрализованы за счет кодовых свойств речи. Структура речевого сигнала аналогична структуре случайного каскадного кода. Поэтому процесс распознавания должен быть реализован в виде процедуры декодирования. Предварительные исследования свойств последовательного декодера в применении к речи указывают на высокую эффективность такого подхода к распознаванию речи (Зигангиров, Сорокин, 1977; Сорокин, 1977; Sorokin, 2003).

В Институте проблем передачи информации РАН был разработан метод динамического анализа речевого сигнала, использующий детекторы артикуляторных событий, несколько напоминающие детекторы временных неоднородностей звуковых сигналов, найденные в слуховых системах живых существ. Первая версия системы распознавания была разработана в 1992 - 1994 гг. Она была испытана на стандартной базе данных Texas Instruments TI46, содержащей десять цифр и десять команд английского языка для 16 дикторов. Средняя надежность распознавания изолированных слов независимо от диктора составляла около 95%, что было очень хорошим результатом для того времени.

2. Потенциальные применения систем автоматического распознавания числительных

Распознавание числительных служит удобным полигоном для отладки новых принципов распознавания, поскольку в этой задаче практически отсутствует семантическая и синтаксическая избыточность, словарь вполне обозрим, и существует практическуая потербность в таких системах. Складывается определенный круг применения систем автоматического распознавания числительных, уже испытанных или намеченных к испытаниям. К их числу относится, например, автоматический набор телефонного номера, произнесенного голосом:

- распознавание номера в персональном компьютере, соединенном с телефонной сетью в условиях офиса;
- распознавание номера на АТС при передаче речевого сигнала по телефонной (кабельной) линии;
- распознавание номера в бортовом компьютере автомобиля;
- распознавание номера в мобильном телефоне;
- распознавание номера на АТС при передаче речевого сигнала по радиоканалу.
- справочная информации при произнесении определенного номера или кодового слова (текущий прогноз погоды, время, курс валют и т.д.);
- вызов скорой помощи, милиции, пожарных и газовой службы;
- передача цифровой информации на пейджер;
- операции с кредитными карточками и банковскими счетами.

Такой широкий круг применения предъявляет определенные требования к формированию обучающей выборки. Один из существенных недостатков современных систем распознавания состоит в том, они обучаются на однородной базе данных, т.е. созданной в одних и тех же условиях. Изменение условий эксплуатации зачастую приводит к полной потере эффективности.

3. База данных

Словарь базы включает в себя однозначные, двузначные и некоторые трехзначные количественные числительные от 0 до 990, произнесенные изолированно, раздельно (с небольшими паузами в составе семизначного номера), и слитно (в составе семизначного номера или более длинного произнесения). Было записано 47 дикторов (34 мужчины и 13 женщин) разного возраста. Стиль произнесения - разговорный, ненапряженный. Отношение сигнал/шум иногда падало ниже 10 дБ.

Речь записывалась в комнате размером примерно 5x3x3 м, с обычным уровнем шумов от вентиляторов персональных компьютеров, уличных слабых шумов при открытом окне или сильных внешних шумов (типа перфоратора) при закрытом окне. Речевой сигнал вводился в персональный компьютер одновременно с двух микрофонов через две звуковые карты типа Sound Blaster. Речевой сигнал квантовался на 16 бит с частотой отсчетов 16 кГц и пропускался через цифровой фильтр высоких частот Баттерворта 10-го порядка с частотой среза 70 Гц, а затем записывался в формате WAVE.

16 дикторов записывались в следующих условиях: один микрофон (направленный, электретный, тип - MC1000, с отношением сигнал/внутренний шум около 60 дБ) укреплялся вертикально на одежде диктора на груди на расстоянии примерно 25 - 30 см от рта (задержка в распространении сигнала составляет около 1 мс). Второй микрофон находился в стандартной телефонной трубке от телефона типа Panasonic KX-T2335, которую диктор держал на обычном расстоянии от рта.

Еще 16 дикторов записывались в следующих условиях: один микрофон (всенаправленный, электретный, тип - Böeder) укреплялся мониторе. Расстояние от диктора до микрофона варьировалось в диапазоне 75 - 125 см (задержка в распространении сигнала составляет около 3 мс). Второй микрофон находился в стандартной телефонной трубке другого типа, которую диктор держал на обычном расстоянии от рта.

Наряду с исходными записями, сигналы от этих 32 дикторов были пропущены через имитатор абонентской телефонной линии, и также включены в состав базы данных.

15 дикторов записывались в следующих условиях: один микрофон (всенаправленный, электретный, тип - Boeder) укреплялся мониторе. Расстояние от диктора до микрофона варьировалось в диапазоне 75 - 125 см (задержка в распространении сигнала составляет около 3 мс). Второй микрофон типа Shure VR230B располагался в головной гарнитуре.

Такой состав базы данных исключает настройку на уникальные условия, и тестирование системы распознавания на ней более реально характеризует ожидаемые параметры - надежность, ошибки и переспросы.

4. Техническое описание

Система автоматического распознавания речи предназначена для работы в режиме независимости от диктора для произвольного типа микрофона и каналов передачи и обработки речевого сигнала с отношением сигнал/шум не хуже 10 дБ. Расположение микрофона относительно рта диктора - произвольное (телефонная трубка, головная гарнитура, микрофон на груди, микрофон на мониторе компьютера на расстоянии не более 1 м в помещении с интегральным уровнем шумов не выше 60 дб). Минимальные аппаратные требования включают в себя персональный компьютер с тактовой частотой не ниже 750 МГц и объемом оперативной памяти от 64 МБ. Нужна также стандартная звуковая карта типа Sound Blaster, 16 бит. Система распознавания работает под операционной системой Windows 95 или Windows 98.

5. Структурная схема

На рисунке наглядно представлена блок-схема системы распознавания, и далее приводится описание каждого из этапов процесса распознавания.

Блок 1. Первичный анализ.

Речевой сигнал вводится в персональный компьютер через звуковую плату с квантованием на 16 бит и частотой отсчетов 16кГц. Устранение постоянной составляющей звуковых карт и наводок от электрических сетей выполняется с помощью цифрового полосового фильтра в диапазоне частот 70 - 6000 Гц, к выходному сигналу которого применяется преобразование Фурье в окне Лапласа на 256 точек с шагом в 5 мс. Сигналы после полосовых фильтров 70-450, 70-900 Гц используются для определения мгновенных частот и коэффициентов периодичности.

Блок 2. Подавление шумов.

Подавление шумов в широкополосном сигнале осуществляется с помощью нелинейного вычитания спектра шума, вычисленного на интервале 500 мс при отсутствии речевого сигнала, и последующем центральном ограничении сигнала, пропорциональном дисперсии шума. Для узкополосных сигналов применяется центральное ограничение во временной области, значение которого определяется также по уровню шумов на интервале в 500 мс.

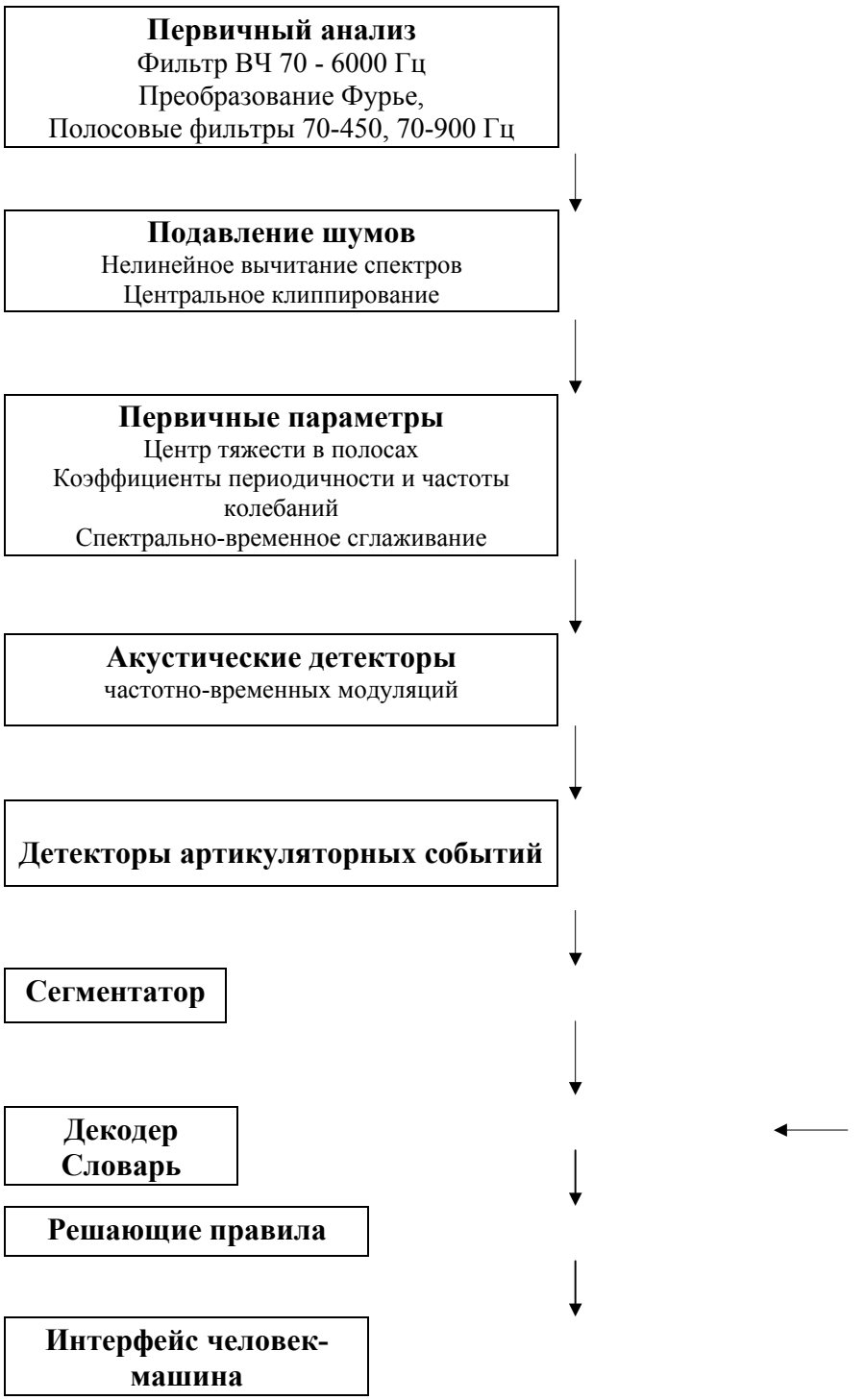
Блок 3. Первичные параметры.

В число первичных параметров входят частоты центра тяжести спектра в различных частотных полосах, коэффициенты периодичности и мгновенные частоты, найденные для узкополосных сигналов во временной области. Динамический спектр сигнала подвергается частотно-временной фильтрации, а затем формируются сигналы с различных частотных полосах, сглаженные с различными постоянными времени. Используются как ненормированные сигналы, так и нормированные к общей энергии.

Блок 4. Акустические детекторы.

Акустические детекторы частотных и амплитудных модуляций представляют собой логарифмы отношения сигналов в разных частотных полосах с разными постоянными времени сглаживания.

Структурная схема



Блок 5. Детекторы артикуляторных событий.

Детекторы артикуляторных событий формируются как многомерные векторы с размерностью 10 - 12, в состав которых входят как акустические детекторы, так и значения спектральных параметров в определенные моменты времени. После обучения по базе данных, артикуляторные детекторы выдают оценки апостериорной вероятности перехода из одного артикуляторного состояния в другое каждые 5 мс.

Блок 6. Словарь.

Словарь эталонов слов создается по фактическим вариантам произнесений для многих дикторов и контекстов по разметке речевых сигналов на установленный ассортимент распознаваемых сегментов речи.

Блок 7. Сегментатор.

Сегментатор отбирает среди множества срабатываний артикуляторных детекторов только те, которые находятся в определенном временном диапазоне относительно сегмента, указанного декодером.

Блок 8. Декодер.

Декодер движется по всем ветвям кодового дерева, в виде которого представлен словарь, и вычисляет значение апостериорной вероятности появления каждого слова, соответствующего вычисленным цепочкам сегментов речи.

Блок 9. Решающие правила.

Решающие правила определяют условия переспроса (отказа от распознавания) и выбор наилучшего (в смысле максимума апостериорной вероятности) слова среди множества кандидатов.

В системе распознавания встроен внутренний ограничитель отношения сигнал/шум, равный 10 дБ, причем шумы могут быть созданы как внешними акустическими источниками, так и шумами электронных цепей. Основными источниками электронных шумов являются наводки на соединительный кабель от микрофона к аналого-цифровому преобразователю и внутренние шумы АЦП. Если в процессе эксплуатации в относительно тихом помещении число переспросов (отказов от распознавания) по признаку отношения сигнал/шум слишком велико, то следует заменить АЦП. Переспросы могут появляться по признаку слишком низкой функции правдоподобия наилучшего решения или по признаку слишком близких значений функции правдоподобия между наилучшим решением и следующим по рангу решением. Увеличение порогов отказа приводит к снижению числа ошибок распознавания и увеличению числа отказов.

Блок 10. Интерфейс "человек-машина".

Определяет характер диалога и формирует исполняющие действия (например, набор телефонного номера или ввод цифровых данных в компьютер).

6. Результаты испытаний

Проводилось два типа испытаний. В одном из них к распознаванию предъявлялись изолированные числительные из базы данных. Результаты распознавания представлены в Табл. 1 и 2. Как видно, средняя надежность распознавания составляет более 88% при практическом отсутствии переспросов. Эти показатели уже близки к минимальным требованиям надежности для коммерческих систем.

Второй тип испытаний состоял в распознавании в реальном масштабе времени с привлечением дикторов, которые не принимали участие в формировании базы данных. Средняя задержка принятия решения на персональном компьютере с процессором 1.3 ГГц составляла около 70 мс при весьма ограниченной дисперсии. По этому параметру разработанная система распознавания вписывается в самые строгие требования. За счет повышения порога переспроса, средняя ошибка распознавания понижается примерно до 10% при 6% переспросов. Встречались, однако, и такие дикторы, у которых ошибка распознавания значительно превышала средние показатели.

Таблица 1. Матрица распознавания

	ноль	один	два	три	четыре	пять	шесть	семь	восемь	девять	NOThing
ноль	39	0	27	1	0	2	0	1	1	5	1
один	0	70	1	0	1	4	0	3	0	0	0
два	5	0	66	0	0	0	0	1	0	1	0
три	0	2	0	70	1	1	1	3	0	0	0
четыре	0	0	0	0	69	1	2	4	0	0	0

пять	2	0	0	2	0	70	1	0	0	2	1
шесть	0	0	0	0	1	1	69	5	0	0	0
семь	0	1	0	0	1	1	1	72	0	0	0
восемь	0	0	1	0	0	1	1	0	72	0	0
девять	0	0	0	1	0	3	2	1	0	68	0

Таблица 2. Средние характеристики распознавания

ноль	39/77 (50%)	Error/No answer: 37/1
один	70/79 (88%)	Error/No answer: 9/0
два	66/73 (90%)	Error/No answer: 7/0
три	70/78 (89%)	Error/No answer: 8/0
четыре	69/76 (90%)	Error/No answer: 7/0
пять	70/78 (89%)	Error/No answer: 7/1
шесть	69/76 (90%)	Error/No answer: 7/0
семь	72/76 (94%)	Error/No answer: 4/0
восемь	72/75 (96%)	Error/No answer: 3/0
девять	68/75 (90%)	Error/No answer: 7/0
Итого:	665/763 (87.1%)	Error/No answer: 96/2

7. Заключение

Вторая версия системы распознавания, использующей концепцию артикуляторных детекторов и кодовой структуры речи, демонстрирует уверенный прогресс. В нынешнем состоянии пока не ставится вопрос о коммерческом использовании, однако есть все основания полагать, что в ближайшем будущем будут достигнуты характеристики, удовлетворяющие большинство потенциальных пользователей в задачах распознавания цифр независимо от диктора.

Литература

1. К.Ш.Зигангиров, В.Н.Сорокин, (1977). "Применение последовательного декодирования к распознаванию слитной речи". Проблемы передачи информации, N 4, с. 81-88.
2. В.Н.Сорокин (1977). "Элементы кодовой структуры речи". В кн. Распознавание образов. Теория и приложения, Наука, М., с. 42-60.
3. Hermansky H. (1998), "Should recognizers have ears?", Speech Communication, v. 25, N 1-3, pp. 3-28.
4. Lippmann R.P. (1997), "Speech recognition by machines and humans", Speech Communication, v. 22, N 1, pp. 1-16.
5. Sorokin V.N. (2003), Some coding properties of speech, Speech Communication, (в печати).