АНАЛИЗ ПАРАМЕТРОВ РЕЧЕВОГО СИГНАЛА СОЗДАЮЩИХ ВОСПРИЯТИЕ ЭЛЕМЕНТАРНЫХ ЗВУКОВ РЕЧИ

ANALYSIS OF THE VOICE PARAMETER SIGNAL TO GIVE A SPEECH PERCEPTION OF ELEMENTARY SOUNDS

Э.Г. Кнеллер *(keg@istrasoft.ru)*
Консорциум "Российские Речевые Технологии", "ИстраСофт"

Рассматривается новый подход к первичной обработке сигнала, выделению и измерению его параметров, непосредственно отвечающих за восприятие звуков речи естественных языков.

В настоящее время системы распознавания речи главным образом основаны на формально-математическом аппарате скрытых Марковских моделей. В них используется недостаточная первичная обработка сигнала для выделения признаков звуков. Этот подход обеспечивает минимально приемлемую надежность распознавания - около 90% только в строго фиксированных условиях, - но отличается неустойчивостью по отношению к помехам внешней среды и каналу связи, что существенно сужает область его применения.

Фирмой «ИстраСофт» (www.istrasoft.ru) разработан новый подход к первичной обработке сигнала, позволяющий выделить и измерить его параметры, непосредственно отвечающие за ощущение того или иного звука речи соответствующего языка. Он основан на математической модели улитки, как первичного анализатора акустофонетической информации, и классификации измеренных параметров речевого сообщения для получения полной транскрибации речи. Полная транскрибация (Rich transcriptions) - это процесс преобразования звуковых сигналов к полно аннотируемому текстовому представлению (слова + метаданные). В отличие от речевых технологий, разработанных другими компаниями, наша технология позволяет выделить и измерить в речевом сигнале фонемы, независимо от особенностей голоса говорящего, что дает хорошие результаты даже при высоком уровне фонового шума.

Под фонемами мы понимаем только ту часть речевого сигнала, которая создает ощущения элементарного звука речи естественного языка.

В процессе исследований были определены и измерены первичные характеристики и границы зон параметров речевого сигнала, передающих (создающих) ощущения звуков речи соответствующего языка. Как известно ощущение звуков речи можно создать, генерируя их как, естественными системами (речь человека или, например, попугая), так и искусственными. При естественной или искусственной генерации речи в речевом сигнале изменяются физические параметры, которые, воздействуя на мембрану уха, возбуждают группы рецепторов. Изменения этих параметров во времени создают звуковые образы (траектории параметров во времени), воспринимающиеся как соответствующие звуки языка.

Поставленная задача заключалась в исследовании речевого сигнала, определении и визуализации параметров, создающих ощущения звуков (фонем), измерении и классифицировании этих параметров.

При исследовании были рассмотрены следующие аспекты:

- обработка сигнала
- слуховые модели
- артикуляторные модели
- модели произношения
- алгоритмы поиска
- обучающиеся алгоритмы

и возможные способы извлечения метаданных из речевого сигнала, включающих информацию о:

- акценте и эмоциях
- стилях говорения
- интонации речи (например, вопрос или утверждение)

Как известно, звуки речи человека генерируются, как правило, артикуляционным аппаратом. В общем, его математическую модель можно представить в виде возбуждающих генераторов тонового и белого шума и группы фильтров, модуляторов и ключей (рот, нос, язык, губы), обеспечивающих фильтрацию и формирование определенного ощущения звука. При генерации речи речевым аппаратом человека для получения различных типов звуков используются следующие физические принципы:

- генерация голосовой щелью периодических звуковых импульсов (сигналов), в этом случае получается "гласный" звук

- формирование артикуляционным аппаратом шумового сигнала (в этом случае голосовая щель отключена), получается “шипящий согласный”
- смешанные шипяще-тоновые звуки типа “З”и “Ж”, где одновременно присутствует шумовая составляющая, модулированная голосовой щелью, или типа “Р”, где модулируется тоновый сигнал
- перекрытие потока воздуха артикуляционными органами и последующим акустическом ударе, генерируется “взрывная согласная”
- отсутствие звука-“пауза”
- изменение параметров артикуляции в процессе генерации звука, создающее ощущение определенного звука (дифтонги, аффрикаты).
- относительное изменение основного тона, определяющее интонацию.

Наиболее известной характеристикой речевого сигнала является основной тон. Эта характеристика представляет собой обычную частотную модуляцию сигнала, параметры которой легко измеряются. Классифицируются относительное изменение частоты, и ее траектория во времени при произнесении слова или фразы. Относительное изменение частоты может достигать 15%, что в европейских языках передает эмоциональную составляющую речи, а в некоторых восточных - смысловую. Так, в русском языке различные траектории вызывают ощущение до 28 типов эмоций. Установлено, что период основного тона разных людей (мужчин и женщин, взрослых и детей) находится в диапазоне 50-250 Гц.

Согласно теории распознавания речи, основанной на формантной модели, по формантам можно определять звук. Наши исследования показали что, форманты это только один из способов генерации звука, служащий для получения соответствующих физических характеристик сигнала создающих ощущение звука. Так, например звук “А” у различных людей может состоять как из одной, так и из двух формант. Положение формант на частотной оси у голосов различных людей (мужчин и женщин, детей и взрослых) также прямо не коррелирует с соответствующими звуками. Кроме того, некоторые форманты присущи индивидууму, создают ощущение его индивидуальности (узнаваемости) и не влияют на формирование ощущения звуков речи. Такие форманты могут использоваться при идентификации личности.

Известно, что ухо преобразует акустический сигнал в спектральную область. Этот преобразователь имеет хорошо исследованные характеристики чувствительности по частоте и линейности преобразования в зависимости от энергетики в спектральных зонах, времени воздействия сигнала, его динамики и времени восстановления чувствительности после воздействия (эффект маскирования, спектральные зоны – барки и т.п.)

Поэтому алгоритмы, используемые для преобразования сигнала в спектральную (частотную) область должны иметь характеристики частотной чувствительности и линейности, близкие к таким при естественном преобразовании звукового сигнала ухом. Это требование является необходимым для правильной интерпретации значений этих характеристик при распознавании и синтезе речи.

Обычно для первичного частотного преобразования в спектральную область в существующих алгоритмах применяется Фурье преобразование. Его недостатком является сворачивание временного отрезка в точку, что не позволяет выделить и измерить динамические параметры взрывных звуков. Вайвлет – преобразования, наоборот, имеют большое разрешение во временной области, что не нужно при измерении других типов звуков и их комбинации.

Нами была разработана математическая модель спектрального преобразователя. Эта модель основана на выделении спектра гребенкой рекурсивных фильтров с настройкой параметров выделения в соответствии с характеристиками чувствительности, близкими к естественному преобразователю звукового сигнала, каким является ухо.

Проведенные нами исследования показали, что определяющими характеристиками, создающими ощущения звуков речи (фонем), вне зависимости от типа возбуждающего сигнала шумового, тонового или их комбинации, являются некоторые параметры (Рис. 1), в том числе динамические, речевого сигнала, воздействующие в течение значимых отрезков времени (5-20мс, 20-50мс. 50-100мс). Это:

- относительная энергетика спектрального воздействия в определенных зонах
- количество зон относительного спектрального воздействия
- ширина зоны относительного спектрального воздействия

и параметры, которые определяют ширину зоны воздействия:

- частота среза спектрального воздействия
- наклон частоты среза спектрального воздействия
- добротность воздействия сигнала

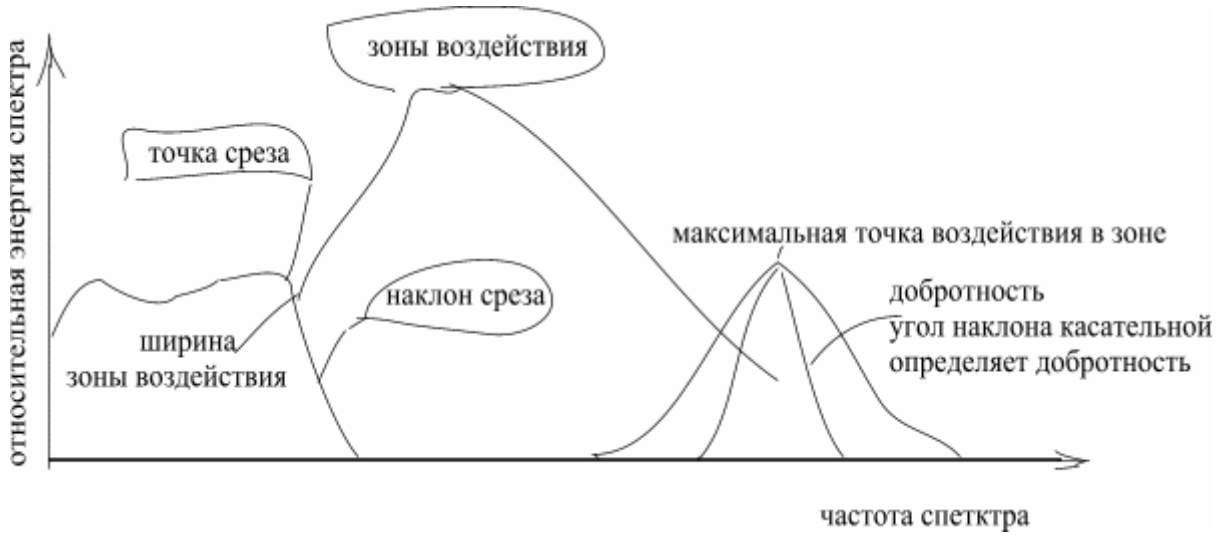

Рис. 1 Характеристики речевого сигнала, создающие ощущения звуков речи

Имеются несколько типов вышеперечисленных характеристик, описывающих изменения этих физических параметров во времени в зонах воздействия:

- статические
- динамические,
- взрывные.

Под статическими нами понимаются характеристики, параметры которых незначительно изменяются в течении 30-100 мс и более.

Под динамическими понимаются характеристики, параметры которых изменяются по определенным траекториям в течение 30-100 мс, причем траектории могут иметь разный знак, но величина производной одинакова (важна динамика, а не ее знак).

Под взрывными понимаются характеристики, параметры которых изменяются по определенным траекториям в течении 10-20 мс.

Зон воздействия может быть одна, две или три.

Разработана методика определения граничных параметров значащих характеристик сигнала, отвечающих за ощущения звука. Предложены алгоритмы измерения этих параметров. Критерием определения соответствия характеристики тому или иному звуку, правильности определения границ параметров, служила оценка звука, который генерировался на основе выделенных параметров, экспертами.

В настоящее время разработанные алгоритмы используются для визуализации фонем и получения оценки произношения звуков в известных языковых обучающих программах серии “Профессор Хиггинс”, выпускаемых фирмой “ИстраСофт”, для сжатия звуковых файлов с целью уменьшения их, а также для работы программы в сети Интернет.

Создан DLL модуль, на основе которого разрабатываются программы для

- передачи речи и музыки через Интернет
- голосовой почты
- компрессии звуковых баз данных
- компактной записи речи и музыки.

Написаны демонстрационные программы ‘Sound Squeezer” (сжатие музыки и речи),”SF6 player” (проигрывание сжатых файлов), “IstraSoft Voice Commander" (командное голосонезависимое распознавание).