Знания-Онтологии-Теории (ЗОНТ-09)

Алгоритмы пофонемного распознавания казахской речи в амплитудно-временно́м пространстве

Карабалаева М.Х., Шарипбаев А.А.

Евразийский университет им.Л.Н.Гумилева, ул. Мунайтпасова, д.5, г. Астана, 0100008, Казахстан.

mkarabal@gmail.com, sharalt@mail.ru

Аннотация. *Область исследования — автоматическое распознавание устной речи. В докладе предложен подход к исследованию речевого сигнала в его временно́м представлении, в рамках данного подхода описаны некоторые эффективные алгоритмы фонемной сегментации.*

Ключевые слова: естественные языки, распознавание речи, фонемная сегментация

1 Введение

Автоматическое распознавание устной речи — традиционная задача искусственного интеллекта. Ею начали с энтузиазмом заниматься еще на заре возникновения информатики как науки так же, как и задачей автоматического перевода с одного языка на другой. Однако по прошествии нескольких десятилетий результаты многочисленных исследовательских групп и в той и в другой области остаются довольно скромными. Две ключевые задачи распознавания речи — достижение стопроцентной точности на ограниченном наборе команд хотя бы для одного дикторского голоса и независимое от диктора распознавание произвольной слитной речи с приемлемым качеством — не решены, несмотря на полувековую историю их разработки.

Главная особенность речевого сигнала в том, что он очень сильно варьируется по многим параметрам: длительность, темп, высота голоса, искажения, вносимые большой изменчивостью голосового тракта человека, различными эмоциональными состояниями диктора, сильным различием голосов разных людей. Два временны́х представления одного и того же фрагмента речи даже для одного и того же человека, записанные в разное время, не будут совпадать. Необходимо искать такие параметры речевого сигнала, которые, с одной стороны, полностью бы его описывали (т.е. позволяли бы отличить один звук речи от другого), и с другой стороны, нивелировали бы указанные выше вариации речи. Затем эти параметры должны сравниваться с образцами, причем это должно быть не простое сравнение на совпадение, а поиск наибольшего соответствия. Это вынуждает искать нужную форму расстояния в найденном параметрическом пространстве.

Таким образом, процедура распознавания речи должна основываться на использовании подходящей системы параметров (признаков) и выполняться с помощью разумных алгоритмов. Особенностью изложенного ниже подхода к пофонемному распознаванию речи является преимущественное рассмотрение речевого сигнала в его временно́м, а не частотном представлении.

Термин «пофонемное распознавание» означает, что в качестве распознаваемых единиц речи мы используем не предложения, не слова, не морфемы, а фонемы — т.е. звуки речи, или элементы фонетического строя языка распознавания (в нашем случае, казахского).

Основоположник теории казахского языка Ахмет Байтурсынов в начале XX века выделил 28 исконных звуков казахского языка, из них 9 гласных (***а, о, ұ, ы, е, ә, ө, у, і***) и 19 согласных (***б, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, у, ш***) [1]. В 1929 году перед разработкой латинизированного алфавита в состав согласных звуков казахского языка был также включен

звук ***х***. В настоящее время в заимствованных из других языков словах также используются согласные звуки ***в*** и ***ф***, которые не нарушают фонетических и фонологических закономерностей казахского языка.

При решении задачи фонемной сегментации казахских слов, разумно разбить звуки казахской речи на несколько классов: гласные (***а, о, ұ, ы, е, ә, ө, у, і***), голосовые согласные (***б, в, г, ғ, д, ж, з, й, л, м, н, ң, р, у***) и глухие согласные (***к, қ, п, с, т, ф, х, ш***). В классе глухих согласных, в свою очередь, можно выделить подклассы шипящеобразных (***с, ф, х, ш***) и паузообразных (***к, қ, п, т***). При этом некоторые звуки, в частности, ***ғ*** и ***қ*** могут проявлять себя по-разному (как элементы разных классов) в зависимости от фонетического контекста.

2 Алгоритмы фонемной сегментации

2.1 Представление речевого сигнала в амплитудно-временно́м пространстве

Определившись с распознаваемыми единицами речи, обратимся к системе признаков, которая позволила бы нам различать классы фонем между собой с удовлетворительной точностью и скоростью.

Звуковой (в частности, речевой) сигнал, оцифрованный звукозаписывающим устройством, представляет собой массив отсчетов (сэмплов). Если пренебречь погрешностью квантования и зависимостью получаемого цифрового сигнала от характеристик микрофона и звуковой карты, то можно рассматривать речевой фрагмент как дискретную функцию амплитуды сигнала от времени.

При этом даже по внешнему виду графика этой функции можно сделать некоторые предположения о произнесенных звуках. Например, как видно на рис. 1, участки графика, соответствующие звукам, произнесенным без участия голоса (глухие согласные), заметно отличаются от участков с голосовыми звуками (гласные, звонкие согласные).

Рис.1. Визуализация слова «каша».

И если функция ведет себя по-разному на участках с разными фонемами, то можно попробовать поискать некоторые отличительные признаки ее «поведения», которые поддавались бы измерению и, будучи замеренными, позволяли бы отличить один класс фонем от другого путем сравнения с пороговыми значениями.

Примером подобного признака может служить величина $V = \sum_{i=1}^{n} \left| x_{i+1} - x_i \right|$ — численный аналог полной вариации функции для дискретного случая. Здесь n — количество отсчетов на участке сигнала, x_i —значение i-го отсчета.

Несмотря на нестабильность, чрезвычайную вариативность временно́го представления речевого сигнала, подобный подход к пофонемному распознаванию оказался плодотворным и позволил разработать комплекс эффективных алгоритмов распознавания речи, в частности, алгоритмов фонемной сегментации. Некоторые из них мы опишем подробнее.

2.2 Обнаружение глухих согласных

Как обнаружить в речевом фрагменте согласные, которые произносятся без участия голоса?

Назовем точками постоянства такие моменты времени, для которых в следующий момент величина сигнала остается неизменной:

$x_i = x_{i+1} \Rightarrow i$ — точка постоянства.

Остальные моменты времени назовем точками непостоянства.

Для паузы и паузообразных звуков характерно большое количество точек постоянства.

Обработаем сигнал цифровым полосовым фильтром с полосой пропускания 100-200 Гц. При этом шипящие, свистящие и аффрикаты превратятся в паузообразные (рис. 2).

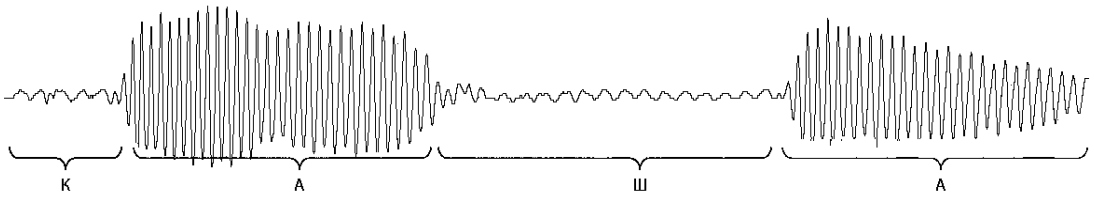

Рис.2. Визуализация слова «каша» после фильтрации.

Разобьем речевой фрагмент на одинаковые окна по n отсчетов (значение n выбирается экспериментально, в зависимости от частоты дискретизации сигнала). Теперь участки с глухими согласными можно выделить, оценив количество точек постоянства в каждом окне. Например, можно вычислить разницу между количеством точек постоянства и количеством точек непостоянства для каждого окна. Если в нескольких идущих подряд окнах эта разница превышает некий заданный порог, то такой участок соответствует глухой согласной (рис. 3) [2].

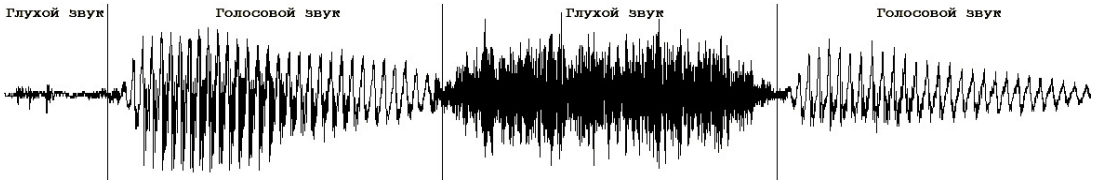

Рис.3. Обнаружение участков с глухими согласными в слове «каша».

2.3 Классификация шипящих и пауз

На выделенном участке с глухими согласными, как отделить шипящие звуки от паузообразных?

Определим для данного участка сигнала функцию V — вариацию с переменным верхним пределом: $V(1) = 0$, $V(n) = \sum_{i=1}^{n} |x_{i+1} - x_i|$. Эта функция будет возрастать для шипящих быстрее, для паузообразных медленнее. Чтобы использовать этот факт для классификации шипящих и пауз, определим также вспомогательную функцию W, возрастающую вместе с V, однако «сбрасываемую» на 0 по достижении некоторого фиксированного значения A.

Положим $W(n) = V(n)$ при $1 \le n \le N_1$,

где N_1 — максимальное число такое, что $W(N_1) \le A$.

Положим $W(N_1+1) = 0$,

$$W(n) = \sum_{i=N_1+1}^{n} |x_{i+1} - x_i| \quad \text{при } N_1+1 \le n \le N_2,$$

где N_2 — максимальное число такое, что $W(N_2) \le A$.

Продолжим ряд чисел N_1, N_2, …, пока не закончится выделенный участок с глухими согласными. При этом график функции W примет характерный вид (рис. 4).

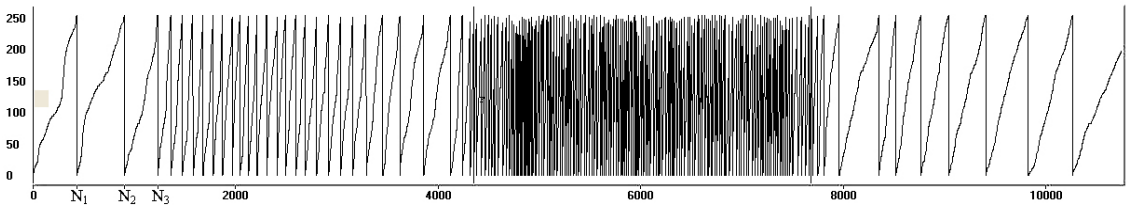

Рис.4. График функции W для слова «каша».

Теперь мы можем отделить шипящие от пауз, оценив расстояния между точками N_k: для шипящих они будут короче, для паузообразных заметно длиннее. Построим массив расстояний N_1, $N_2 - N_1$, $N_3 - N_2$, … Если на участке с глухими согласными элементы массива превышают некий заданный порог, то этот отрезок сигнала соответствует паузообразному звуку, иначе — шипящему (рис. 5) [2].

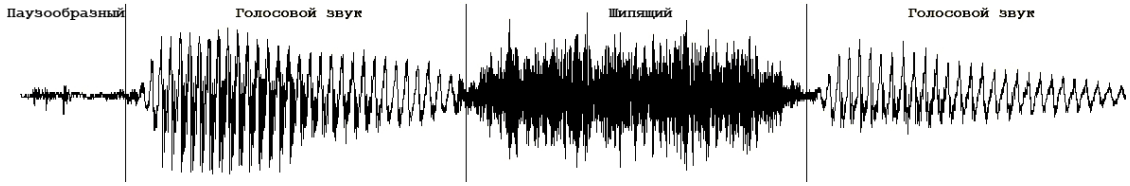

Рис.5. Классификация шипящих и пауз в слове «каша».

2.4 Классификация гласных и голосовых согласных

На выделенном участке с голосовыми звуками, как отделить гласные от голосовых согласных?

Разобьем выделенный участок сигнала на окна по *n* отсчетов (значение *n* зависит от частоты дискретизации сигнала). Пусть у нас получилось всего *m* окон. В каждом окне вычислим полную вариацию $V = \sum_{i=1}^{n} |x_{i+1} - x_i|$, полученные числовые значения запишем в массив из *m* элементов.

Теперь возьмем первые *k* окон (окна с номерами *1*, *2*, …, *k*) и найдем среднее арифметическое величин полной вариации в этих окнах (значение *k* выбирается экспериментально, в зависимости от *n* и исходя из соображений о длительности звучания отдельной фонемы). Вычисленное значение примем за пороговое. Те окна, в которых величина полной вариации превышает порог, пометим символом «В» («выше порога»), остальные окна — символом «Н» («ниже порога»). Получится столбец из *k* символов.

Затем сдвинемся на одно окно вправо и опять возьмем *k* окон (окна с номерами *2*, *3*, …, *k+1*). Повторим для них описанную обработку, в результате чего получится еще один столбец из *k* символов, сдвинутый относительно первого столбца на 1 символ вниз.

Будем повторять всю процедуру, каждый раз сдвигаясь на одно окно, пока не закончится выделенный участок. В конце получим трапецеидальную таблицу, заполненную символами «В» и «Н» (рис. 6).

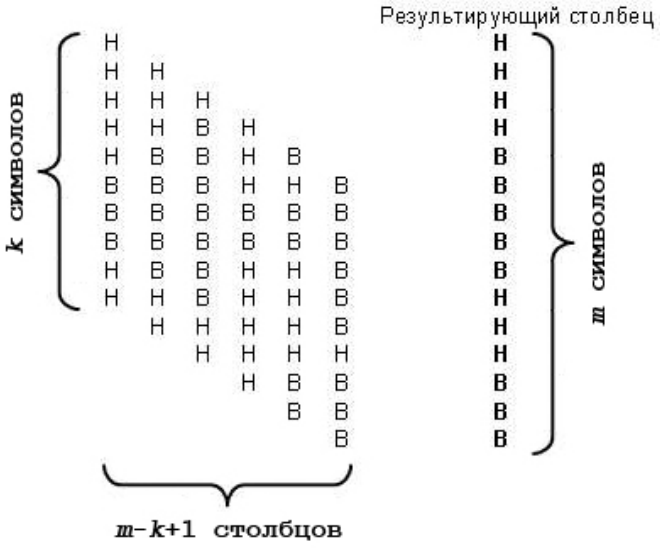

Рис.6. Результат пооконной «В-Н»–обработки.

Теперь перейдем от столбцов к строкам. Проанализируем строки одну за другой и примем окончательное решение о том, какие окна соответствуют гласным, а какие — голосовым согласным.

Если строка начинается и заканчивается одним и тем же символом, то припишем ей этот символ как итоговое значение. Если строка начинается одним символом, а заканчивается другим, то припишем ей тот символ, который чаще встречается в строке. В итоге получится результирующий столбец из m символов (рис. 6). Окна, помеченные символом «В», соответствуют гласным, а окна, помеченные символом «Н», – голосовым согласным (рис. 7).

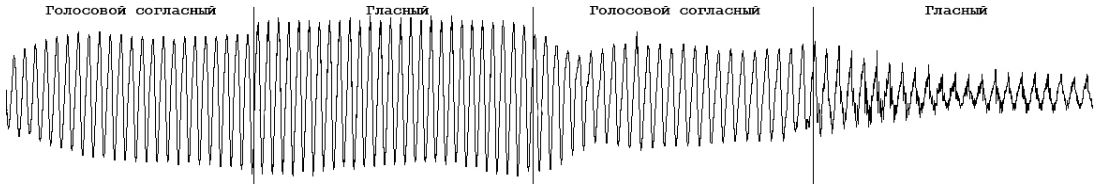

Рис.7. Классификация гласных и голосовых согласных в слове «мина».

Если требуется произвести фонемную сегментацию произвольного слова, то, в общем случае, сначала в речевом фрагменте выделяются участки с глухими согласными (звуками, произнесенными без участия голосовых связок); затем выделенные участки разбиваются на классы шипящих и паузообразных звуков; и наконец, находящиеся между ними участки с голосовыми звуками разбиваются на классы гласных и голосовых согласных (рис. 8) [2].

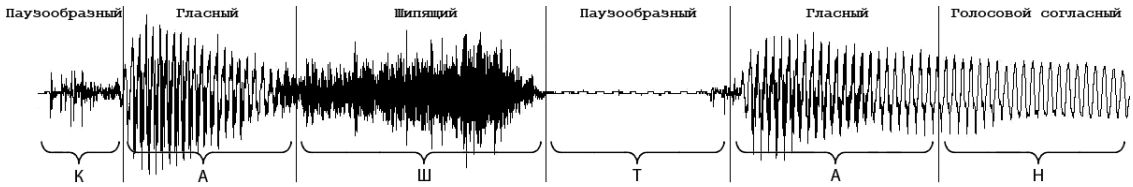

Рис.8. Сегментация слова «каштан».

3 Заключение

Описанные алгоритмы демонстрируют высокую точность и стабильность фонемной сегментации при корректном определении пороговых значений. В целом же, изложенный подход к исследованию речевого сигнала в его временно**м** представлении позволяет разрабатывать и оптимизировать аналогичные алгоритмы, предназначенные для решения множества других разнообразных задач, возникающих в процессе распознавания речи.

Литература

[1] Байтұрсынов А.: Тіл тағылымы. *Алматы: Ана тілі. – 1992. – С.448.*

[2] Шелепов В.Ю., Ниценко А.В.: Амплитудная сегментация речевого сигнала, использующая фильтрацию и известный фонетический состав *// Искусственный интеллект. – 2003. – № 3. – С. 421-426.*