ОБ ОДНОМ ПОДХОДЕ К ВОПРОСУ ФОНЕТИЧЕСКОЙ ИДЕНТИФИКАЦИИ ГРУППЫ ЩЕЛЕВЫХ СОГЛАСНЫХ И АФФРИКАТ РУССКОГО ЯЗЫКА

М.Ф. БОНДАРЕНКО, А.Н. ГАВРАШЕНКО

Кафедра вычислительной техники
Институт радиоэлектроники
Харьков, Украина, СССР, 310141

АННОТАЦИЯ

В докладе рассматривается группа щелевых согласных (С,З,Ш,Ж,Ф,Х') и аффрикат (Ц,Ч) русского языка, предлагаются методы сегментации и фонетической идентификации этой категории звуков, надёжно работающие как при обработке изолированных слов, так и слитной речи. Предлагаемые алгоритмы используют результаты временного и спектрального анализов речи, а также результаты анализа тонкой структуры речевых сигналов, дискретизированных с учётом эффекта сглаживания в слухе /1/.

При разработке надёжных и эффективных методов распознавания отдельных фонем в речевом потоке важное значение приобретает решение задачи сегментации речи на составляющие её звуки. От успешности её решения зависит успех в решении задачи надёжного распознавания фонем и всего распознаваемого сообщения в целом.

Для надёжного выделения рассматриваемых в докладе фонем, из изолированно произносимых слов или слитной речи, предлагается метод сегментации речевого сигнала на участки, фонетически соответствующие шумовым согласным. В практике распознавания речи задача сегментации решается различными методами. Наибольшее распространение получили спектральные методы и методы анализа клиппированного сигнала. Однако, использование для сегментации процедур спектрального анализа, требующих больших вычислительных затрат или специальных устройств, не всегда целесообразно в практических системах.

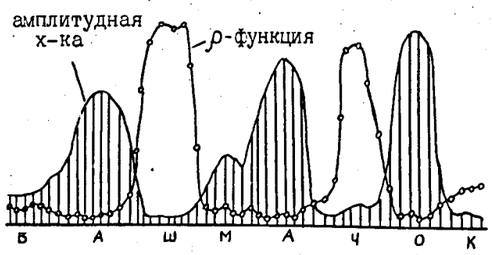
Для решения задачи сегментации речевого сигнала на участки, соответствующие шумовым звукам, предлагается следующий алгоритм сегментации, обладающий высокой помехоустойчивостью к воздействию аддитивных и локальных помех. Разработанный алгоритм ориентирован на использование в условиях с высоким уровнем окружающих акустических шумов, вплоть до 80 дБ.

При разработке алгоритма сегментации была использована графическая форма представления речевых сигналов. Для этого необходимо выполнить кратковременный анализ интенсивности сигнала и функции нулевых пересечений. Такие графические изображения, названные условно "динамическими портретами, отражают динамику исследуемых параметров речевого сигнала во времени. Отличительной особенностью рассматриваемой в докладе категории звуков является ярко выраженный шумовой характер этих фонем, что отличает их от всех остальных звуков русского языка. Для выделения шумной части звука анализ функции нулевых пересечений выполнялся на интервале скользящего с шагом 5 мс временного окна в 25 мс. Количество всех пересечений нуля в таком окне определяет одну точку графика. Метод перекрывающегося окна был использован с целью устранения присутствующих в речевом сигнале неоднородностей. Для устранения некоторой "лохматости" картинки "динамический портрет" сглаживается методом скользящего усреднения с интервалом сглаживания, равным 7 точкам графика. Полученное таким образом изображение достаточно наглядно показывает места расположения шумовых согласных в потоке речи. Для определения границ таких участков используется метод сечений, который в общей сложности предполагает построение двух сечений. Одно из них проводится на некотором усреднённом уровне, уровень проведения второго зависит от индивидуальных особенностей диктора. Срез по сечениям и определяет границы шумовых участков в речевом сигнале.

Выделением из речевого потока шумовых участков завершается первый шаг работы алгоритма сегментации. Известно, что аффрикаты являются более сложными по своему составу звуками по сравнению со щелевыми согласными ввиду присутствия кроме шумного участка ещё и участка смычки. Поэтому выделение из акустического сигнала шумового сегмента ещё не определяет всего согласного. Это обстоятельство, а также необходимость уточнения границ щелевых согласных, предусматривает переход ко второму шагу работы алгоритма сегментации, на котором строятся "динамические портреты" другого рода. Для этого используются ре-

зультаты кратковременного анализа не только функции нулевых пересечений, но и интенсивности речевого сигнала на интервалах непересекающихся 10 мс сегментов. Выбор интервала обусловлен особенностями тракта речеобразования человека. Результатом кратковременного анализа интенсивности акустического сигнала является набор значений абсолютных максимумов сигнала на каждом из 10 мс сегментов. Неоднородности, как правило присутствующие в изображении, сглаживаются процедурой скользящего усреднения с окном, равным 5 точкам графика. Построенный таким образом "динамический портрет" даёт возможность разграничить участки расположения в речевом сигнале гласных и согласных звуков по изменениям амплитудной характеристики.

В подавляющем большинстве случаев согласные звуки в речи располагаются в окружении гласных. Находясь между гласными, шумовые согласные довольно отчётливо определяют своё место на "динамическом портрете". При этом поиск их границ осуществляется на основе анализа характера изменения интенсивности окружающих гласных и ρ -функции на участке шумового согласного. В качестве примера можно рассмотреть слово "БАШМАЧОК". "Динамический портрет" на втором шаге работы алгоритма сегментации для этого слова представлен на рисунке 1.



"Динамический портрет" слова "БАШМАЧОК"

Рис. 1

Из изображения видно, что границы между шумовыми согласными и окружающими их гласными расположены на пересечении огибающей интенсивности и ρ -функции. Слуховой анализ выделенных участков речевого сигнала подтверждает правильность выделения границ.

Если в речевом потоке шумовые согласные расположены в окружении согласных звуков (кроме исследуемых шумовых), то в качестве границ принимаются границы шумовой части фонемы, чего вполне достаточно для щелевых согласных, содержащих в своём составе только шумный участок. Выделение же участка смычки у аффрикат осуществляется на следующих этапах анализа.

Таким образом, использование описанного алгоритма сегментации позволяет выделять из потока речи шумовые согласные независимо от контекста при условии, что две шумовые согласные не могут располагаться в речи рядом. Выделяемый алгоритмом звук "Щ" исключается из рассмотрения по признаку характерной длительности шумового участка, свойственной только этой фонеме.

Для осуществления предварительной фонетической идентификации выделенных из речи шумовых участков был разработан алгоритм, позволивший на базе небольшого количества признаков осуществить грубую маркировку исследуемых согласных.

Как показал анализ множества осциллограмм и "динамических портретов" словосочетаний с шумовыми согласными в различных контекстах и положениях в слове, даже стационарная часть звуков характеризуется большой вариабельностью и неоднородностью характеристик. В связи с этим из всего количества присутствующих на стационарном участке элементарных сегментов необходимо выбрать группу таких рядом расположенных сегментов, которые характеризуются наибольшей однородностью и стабильностью в рамках описанного ниже критерия. Оценка фонетического качества шумовых согласных на таких акустически однородных участках является наиболее устойчивой. Как показали эксперименты, шумовые фонемы можно надёжно идентифицировать на интервале трёх рядом расположенных акустически однородных 10 мс сегментах. При этом не требуется больших вычислительных затрат на определение характерных сегментных признаков.

Объединение сегментов в акустически однородные области производилось по следующему критерию. Была введена мера близости между тремя соседними сегментами стационарной части шумовых согласных. При этом на каждом сегменте вычислялись значения двух признаков: коэффициент монотонности, характеризуемый общим количеством участков монотонного изменения сигнала и количество мгновенных значений речевого сигнала, превышающих некоторый уровень С. Три рядом стоящих сегмента объединяются в акустически однородную область по минимуму среднеквадратического отклонения значений признаков на этих сегментах от их среднего уровня. На выделенном таким образом участке звука будет вестись его дальнейшее распознавание.

Для предварительной фонетической идентификации выделенных из речи шумовых согласных сформируем систему признаков, в эту систему были включены ряд широко используемых (энергия сигнала, число переходов сигнала через нуль, нормализованный коэффициент автокорреляции, максимальная амплитуда сигнала на сегменте и другие), а также такие признаки, как коэффициент монотонности и количество значений сигнала на сегменте, превышающих некоторый порог.

Как показали проведенные эксперименты, для целей предварительной классификации шумовых согласных на группы важную информацию несут такие признаки, как коэффициент монотонности огибающей речевого сигнала (F), количество нулевых интервалов меньших 50 мкс (L) и общее количество мгновенных значений сигнала, расположенных выше некоторого уровня $D_{пор}$ (G). Значения признаков вычисляются на интервале 10 мс сегментов. Так, признаки F и L позволили разделить все исследуемые шумовые согласные на классы

$$\left.\begin{array}{l}(\omega^c, \omega^3, \omega^4)\in\Omega_1\\ (\omega^ш, \omega^*, \omega^ф, \omega^ц, \omega^х)\in\Omega_2\end{array}\right\} \quad (1)$$

Признак G позволил выделить из класса Ω_2 фонемы "Ш", "Ж", "Ч". В результате получили систему из трёх классов

$$\left.\begin{array}{l}(\omega^c, \omega^3, \omega^4)\in\Omega_1\\ (\omega^ш, \omega^*, \omega^ц)\in\Omega_2\\ (\omega^ф, \omega^х)\in\Omega_3\end{array}\right\} \quad (2)$$

Аффрикаты "Ц" и "Ч" надёжно выделяются из классов Ω_1 и Ω_2 по значению признака G, вычисленному на 10 мс сегментах всего звука и n сегментах предшествующих ему ($n = 9$). Значение порога $D_{пор}$ для каждой из аффрикат "Ц" и "Ч" выбирается с учётом различий в их интенсивности. Распределение значений признака G на анализируемом участке речевого сигнала надёжно характеризует смычку, свойственную этим звукам. Для остальных фонем из классов Ω_1 и Ω_2 распределение признака G имеет совсем иной характер.

Таким образом, по результатам предварительной фонетической идентификации удалось надёжно разделить шумовые согласные на следующие группы

$$\left.\begin{array}{l}(\omega^c, \omega^3)\in\Omega_1;\quad (\omega^ш, \omega^*)\in\Omega_2;\\ (\omega^ф, \omega^х)\in\Omega_3;\ (\omega^4)\in\Omega_4;\ (\omega^ц)\in\Omega_5\end{array}\right\} \quad (3)$$

Методы временного анализа речевого сигнала позволяют с высокой надёжностью выполнять сегментацию и предварительную фонетическую идентификацию шумовых согласных в потоке речи. Но информации, получаемой на этом уровне обработки недостаточно для разделения фонемных пар внутри выделенных групп. Для решения этой задачи необходимо осуществить анализ более тонкой структуры анализируемых речевых сигналов. С этой целью были использованы средства спектрального анализа, позволяющие исследовать микроструктуру фонетических пар звонкий-глухой и определить их отличительные признаки.

Известно /2/, что в образовании звонких шумовых согласных принимает участие фонация. Поэтому такие звуки характеризуются наличием в их спектре составляющих основного тона голоса (ОТГ). Из ранних работ Варшавского /3/ следует, что частота 400 Гц может быть принята как абсолютная верхняя граница наличия основного тона (ОТ) в естественно произнесенном речевом сигнале. С учётом этого для анализа был использован полосовой фильтр с диапазоном 80–400 Гц для выделения ОТ. Спектральному анализу будем подвергать интервал протяжённостью 40 мс. Этот интервал включал акустически однородную область и один из примыкающих к ней сегментов, менее других отличающийся от неё по своим характеристикам. Таким образом, на анализируемом участке будет содержаться не менее четырёх периодов ОТ.

Прежде чем речевой сигнал поступит на спектральный анализатор он подвергался нормированию по динамическому диапазону. Правило, позволяющее отделять звонкие фонемы "З" и "Ж" от их глухих аналогов "С" и "Ш", использует значения двух признаков – энергия сигнала в канале 80–400 Гц и признак периодичности. Известно /4/, что процесс восприятия ОТГ человеком представляет собой определение периодичности огибающих во всех или некоторых каналах слухового анализатора.

При дальнейшем анализе за меру периодичности принимается степень близости различных замеров периодов ОТ. На участках звонких согласных соседние замеры ОТ очень близки. На глухих согласных – замеры периодов ОТ случайны и сильно отличаются друг от друга.

Для определения значений периодов ОТ был использован анализ микроструктуры шумовых фонем в частотном канале 80–400 Гц. С этой целью речевой сигнал дискретизировался с учётом эффекта сглаживания в слухе /1/. Речевой сигнал, представленный в бинарном виде, является весьма удобным объектом исследования прежде всего в том смысле, что единственным информативным элементом в данном случае является только временной интервал между смежными единицами. Эксперименты показали, что всё многообразие комбинаций нулей и единиц, определяющих структуру временных отношений в пространстве выборки, можно свести к конечному и при том довольно небольшому числу основных или базовых блоков по критерию регулярности последовательности двоичных элементов. Таким образом, были выявлены регулярные стационарные блоки единиц Y (длительностью K), регулярные блоки нулей N (длительностью m) и нерегулярные нестационарные блоки X (произвольные комбинации двоичных эле-

ментов).

Для упрощения анализа двоичного представления сигнала бинарный код подвергался двойному медианному сглаживанию с различными окнами (3 и 5 шагов).

Исследование наиболее общих закономерностей структурной организации периодов ОТГ позволили выявить стабильные комбинации блоков

$$\left.\begin{array}{l} Y\ N_1\ Y'(X)\ N_2 \\ Y\ X_1 N_1\ Y'(X_2) N_2 \\ Y\ X_1 N_1\ X_2 \\ Y\ X\ N \end{array}\right\} \qquad (4).$$

Проведенные исследования показали, что период ОТ определяется по наличию в бинарном коде конкретной комбинации (4) и допустимых численных значений входящих в неё блоков. Для определённой реализации звонких шумовых согласных структурная организация периодов ОТ имеет довольно стабильный характер. Для глухих же согласных рядом стоящие периоды могут иметь любую из организаций (4).

Таким образом, по описанному методу на интервале анализа предъявляемого для распознавания звука определяется четыре периода ОТ, которые сравниваются между собой по некоторому критерию. Если различия между периодами ОТ меньше некоторого порога R, то это звонкий звук, если больше - то глухой.

Предложенный метод оценки ОТ обладает тем преимуществом, что позволяет идентифицировать глухой шумовой согласный не только по соотношению длительностей нескольких периодов ОТ, но и по специфике структурных комбинаций базовых блоков, составляющих ряд последовательных периодов ОТ. Большая изменчивость таких структур на интервале анализа является характерной для глухих шумовых согласных, что не свойственно звонким фонемам.

Для разрешения возникающих неопределённостей используется значение энергии сигнала в той же полосе частот 80-400 Гц.

Шумовые согласные "Ф" и "Х" из класса Ω_3 (3) обладают перекрывающимися в широком диапазоне спектрально-временными характеристиками и поэтому различаются между собой с очень низкой надёжностью (около 64%).

В заключение следует отметить, что предложенные в данном докладе методы сегментации и фонетической идентификации шумовых согласных обладают высокой помехоустойчивостью по отношению к аддитивным и локальным помехам. По результатам контрольной проверки для 9 дикторов (6 мужчин и 3 женщины) обеспечивается надёжность распознавания 98,2%.

Список литературы

[1] Абрамов О.М., Дрюченко А.Я., Усенко С.А., Шабанов-Кушнаренко Ю.П. Эффект сглаживания в слухе. - В сб. Проблемы бионики, Харьков, 1977, вып.19, с.31-37.

[2] Сапожков М.А., Михайлов В.Г. Вокодерная связь. - М.: Радио и связь, 1983. - 248с.

[3] Варшавский Л.А., Литвак И.М. Исследование формантного состава и некоторых других физических характеристик звуков русской речи. - Проблемы физиологической акустики, М.-Л.: изд-во АН СССР, 1955, вып.3,с.5-17.

[4] Фант Г. Акустическая теория речеобразования. - М.: Наука, 1964.-284с.

Se 37.5.4