Метод обнаружения и выделения звука [р] в речевом сигнале

Шелепов В.Ю.,
доктор физико-математических наук, профессор, главный научный сотрудник ГУ ИППИ

Ниценко А.В.,
научный сотрудник ГУ ИППИ

Аннотация

В статье представлен метод обнаружения и выделения дрожащего звука «р» в речи в твердом и мягком вариантах за счет использования некоторых особенностей этого звука. Разработан алгоритм, который позволяет локализовать звук [р] независимо от фонетического окружения, тем самым повышая точность автоматической сегментации речи, а также сократить число слов-кандидатов на распознавание.

Ключевые слова: сегментация речи, локализация звука [р], вариация, точка постоянства, полосовой фильтр.

Звук «р» в твердом и мягком вариантах обладает в русской речи высокой статистической значимостью. Так, работая с электронной версией словаря Зализняка [1], содержащей 93.507 слов, и отбирая те из них, которые содержат «р», получаем 45.665 слов. Отсюда следует, что умение обнаруживать и выделять этот звук может быть очень полезно при распознавании устной русской речи. Этой и близким проблемам посвящены работы [2–7], в частности, статья [2], в которой авторам принадлежит метод, использующий последовательное сглаживание. В настоящей статье предлагается новый подход к этой задаче, который с одной стороны проще, а с другой, как показывает опыт, дает лучшие результаты.

Иногда нам придется различать твердый звук [р] и мягкий звук [r]. Если же то, что говорится, относится к обоим этим случаям, будем использовать для них общее обозначение R.

Пусть $x_1,x_2,\ldots$ — последовательные отсчеты сигнала. Мы используем ниже окна по 256 отсчетов, для которых будет вычисляться вариация (численный аналог полной вариации):

$$V = \sum_{i=0}^{254} |x_{i+1} - x_i|, \qquad (1)$$

а также количество точек постоянства (мы называем момент дискретного времени точкой постоянства сигнала, если в следующий момент значение сигнала не меняется; в противном случае мы называем этот момент точкой непостоянства).

Известно, что «р» в конце слова может реализовываться как глухой звук. Мы не будем касаться этого случая, и будем всегда иметь в виду звонкое «р». Тогда можно утверждать, что этот звук образуется за счет одного или нескольких следующих один за другим (подобно барабанной дроби) ударов языка о нёбо при работающих голосовых связках. Будем называть их р-ударами. При каждом р-ударе происходит кратковременное перекрытие голосовой щели. Это отражается в записанном сигнале в виде одного или нескольких коротких паузообразных участков, которые разделены голосовыми элементами (см. рис. 1 и 2). На этом будет основана идентификация звука R.

Рис. 1. Сигнал для слова "народ" с одним р-ударом.

Рис. 2. Сигнал для слова "народ" с несколькими р-ударами.

Для лучшего различения паузообразных и голосовых элементов в звуке [р] предлагается обработать сигнал простейшим полосовым фильтром (см. [8]) с последующей нормализацией путем деления на амплитуду сигнала и умножения на 256. Для одного из авторов наиболее подходящим является фильтр с полосой пропускания 400–600 Гц. Результат представлен на рисунке 3.

Рис. 3. Сигнал рисунка 2 после обработки фильтром 400–600 Гц.

Как показывает опыт, выбор полосы пропускания зависит от диктора. Для того чтобы уменьшить эту зависимость, увеличим полосу пропускания до 300–900 Гц (см. рис. 4).

Рис. 4. Сигнал рисунка 2 после обработки фильтром 300–900 Гц.

Операции следующих далее пунктов 1, 2 осуществляются для профильтрованного сигнала, все остальные операции выполняются для исходного сигнала.

1. Сигнал разбивается на неперекрывающиеся окна по 256 отсчетов. На каждом из них вычисляется разность между количеством точек непостоянства и количеством точек постоянства. Пусть D — массив этих разностей. Устраняются единичные включения положительных элементов в массиве D с помощью замены их на -1. Участок сигнала, на котором значения в массиве D меньше 0, помечается как «N» (см. рисунки 5 и 6), если длина этого участка больше некоторой минимальной величины *m* (она выбрана равной 5 окнам по 256 отсчетов). N-участки не содержат R.

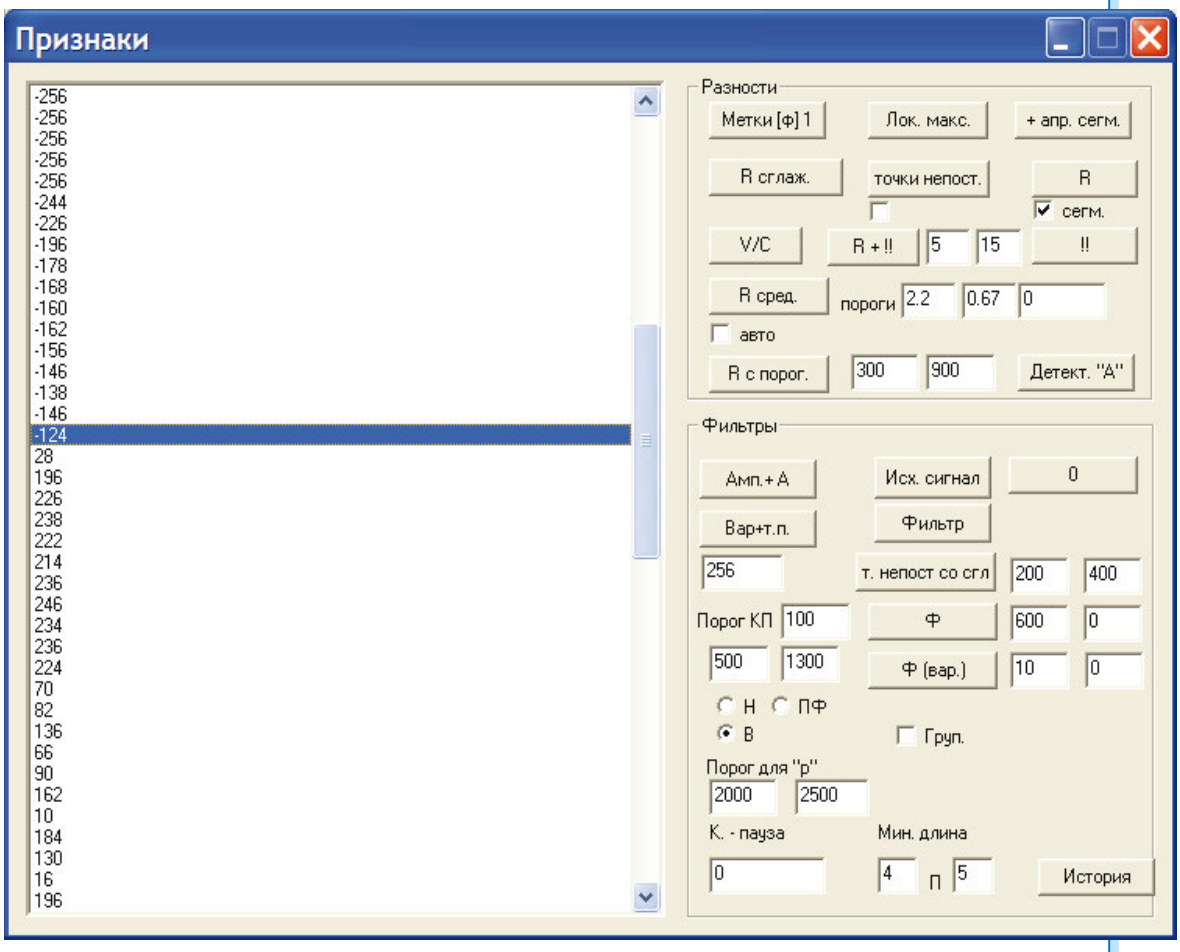

Рис. 5. Слева массив разностей точек непостоянства и постоянства для сигнала на рисунке 2. Курсор отмечает конец 1-го N-участка.

Рис. 6. Результат выделения «N»-участков в сигнале на рисунке 2.

2. На участках сигнала, не помеченных как «N», вычисляется массив значений вариации (1). На рисунке 7 он представлен для среднего участка рисунка 6.

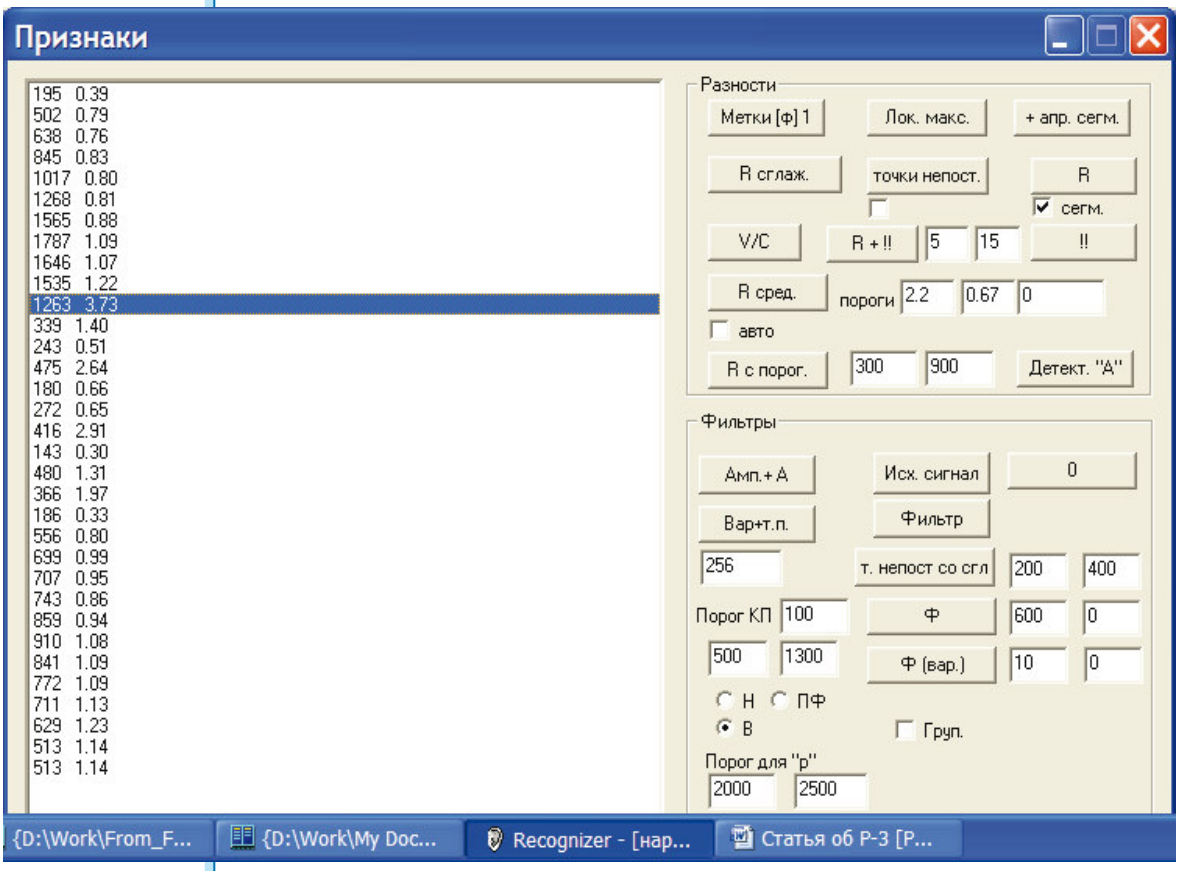

Рис. 7. Массив значений вариации для среднего участка рисунка 6 с отношениями предыдущего к последующему.

Пусть для некоторого элемента V[i] массива выполняется условие $V[i+1] \neq 0$ и $V[i]/V[i+1] > T_1$ (у нас величина порога T_1 выбрана равной 2.2). Тогда данный элемент определяет начало р-удара и соответствующую метку в сигнале, при условии, что далее найдется конец этого р-удара, определяемый элементом V[j] массива V, для которого выполняются условия $i<j<i+m$, $V[j+1] \neq 0$, $V[j]/V[j+1] < T_2$ (величина порога T_2 выбрана 0.67). Тот же звук R содержит еще один р-удар, если его начало отстоит от конца предыдущего не более чем на *m* окон. В противном случае мы имеем дело уже с другим звуком R в пределах анализируемого участка, не помеченного как N. Пример — слово "Урарту" (см. рис. 8).

Рис. 8. Визуализация сигнала для слова "Урарту" с 2-мя отмеченными R.

Возвращаясь к примеру, представленному на рисунке 2 мы, таким образом, получаем результат, представленный на рисунке 9:

Рис. 9. Визуализация сигнала на рисунке 2 с выделением последовательных р-ударов.

(Отметим, что здесь программе не удалось выделить последний р-удар). Теперь мы можем сегментировать звук R (см. рис. 10). Считаем его началом начало первого р-удара. Опыт показывает, что метку конца звука разумно проставлять на расстоянии 3-х окон справа от конца последнего выделенного р-удара. Начальная метка звука маркируется символом R.

Рис. 10. Результат выделения R в слове "народ".

3. Обратимся к авторской априорной сегментации (см. [9]), в которой начала отрезков гласных маркируются символом W, а начала отрезков звонких согласных — символом С. После определения границ сегмента «R», необходимо совместить его метки и метки априорной сегментации. При этом рассматриваются следующие случаи.

а). Задаются два положительных числа T_3 и T_4 ($>T_3$). Если метка начала сегмента «R» находится достаточно близко к началу сигнала (количество отсчетов между ними $<T_3$), то эта метка и все предшествующие метки априорной сегментации, кроме метки начала сигнала, удаляются, и маркировка начальной метки сигнала заменяется на «R».

б). Если упомянутое количество отсчетов $>T_4$, то все упомянутые метки и их маркировка остаются на месте.

в). Если упомянутое количество отсчетов $\geqslant T_3$, но $\leqslant T_4$, то производится распознавание [а], [и], [о], [у], [э], [г] на отрезке в 3 окна от начала сигнала. При результате [г] действуем, как в случае а). При других результатах действуем, как в случае б).

Отметим, что использование 2-х порогов и распознавание начала актуальны в случае мягкого звука [г] (это иллюстрирует пример на рис.11). В случае твердого [р] без них можно было бы обойтись.

Рис. 11. Визуализация сигнала для слова "риска".

При построении априорной сегментации у нас используется минимальная допустимая длина T_5 для гласного «W» и минимальная допустимая длина T_6 для звонкого согласного «C» в середине слова, выраженные в количестве окон по 256 отсчетов.

г). Если выделенное «R» целиком попадает внутрь сегмента «W» в середине слова, то производится проверка на длину нового сегмента, образовавшегося между меткой начала сегмента «W» и первой меткой «R». Если эта длина меньше T_5 , то обозначение метки начала сегмента «W» заменяется на «R» и начальная метка «R» не добавляется в сегментацию. В противном случае обозначение сегмента остается прежним и добавляется новая метка для начала сегмента «R». Производится также проверка на длину нового сегмента между меткой конца сегмента «R» и меткой конца сегмента «W» с порогом T_5. Если его длина больше либо равна порогу, метка конца «R» добавляется в сегментацию и новый образовавшийся сегмент помечается как «W», в противном случае новая метка не добавляется.

д). Если выделенное «R» целиком попадает внутрь сегмента «C» в середине слова, то делается то же, что и в предыдущем случае с заменой «W» на «C» и T_5 на T_6.

е). Если метка начала и метка конца сегмента «R» попадают в разные сегменты, то для метки начала применяются правила, фигурирующие в случаях г) и д), а метка конца не добавляется в сегментацию.

Рисунок 12 представляет сегментацию сигнала на рисунке 2 с учетом выделения R:

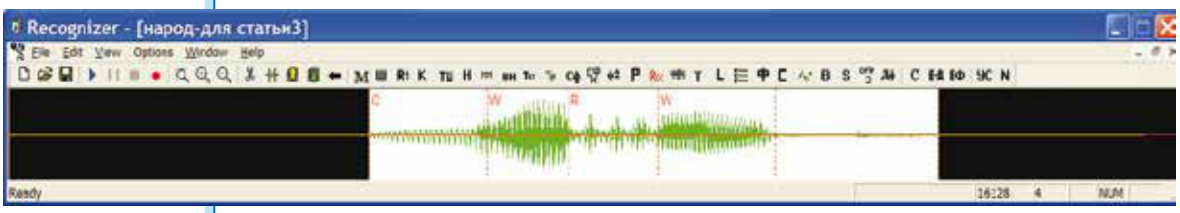

Рис. 12. Результат сегментации сигнала на рисунке 2 (маркировка заключительного глухого опущена).

Литература

1. *Зализняк, А.А.* Грамматический словарь русского языка. Словоизменение / А.А. Зализняк. — М.: Аст-пресс. — 2008. — 880 с.
2. *М.Х. Карабалаева.* Обнаружение и выделение звука [р] в речевом сигнале / Карабалаева М.Х., Ниценко А.В., Шелепов В.Ю. // Искусственный интеллект. — 2011. — № 1. — С. 168–174.

3. *А.А. Конев.* Выделение вокализованных звуков в слитной речи / Конев А.А., Тихонова В.И. // Сборник трудов XVI сессии Российского акустического общества. Том III — М.: ГЕОС, 2005. — С. 47–50.
4. *N. Dhananjaya.* Acoustic analysis of trill sounds / Dhananjaya N., Yegnanarayana B., Bhaskararao P. // The Journal of the Acoustical Society of America. — 2012. — No. 131 (4). — pp. 3141–3152.
5. *Maria-Josep Sole.* Aerodynamic characteristics of trills and phonological patterning // Journal of Phonetics. — 2002. — No.30. — pp. 655–688.
6. *N. Dhananjaya.* Features for automatic detection of voice bars in continuous speech / Dhananjaya, N., Rajendran, S., and Yegnanarayana, B. // Proceedings of Interspeech, Brisbane, Australia. — 2008. — pp. 1321–1324.
7. *N. Dhananjaya.* Voiced/nonvoiced detection based on robustness of voiced epochs / Dhananjaya, N., Yegnanarayana, B.// IEEE Signal Process. — 2010. — Lett. 17. — pp. 273–276.
8. *Шрюфер Е.* Обробка сигналів: цифрова обробка сигналів. — Київ: "Либідь". — 1992. — 295 с.
9. Сегментация и дифонное распознавание речевых сигналов / А. К. Бурибаева, Г. В. Дорохина, А. В. Ниценко, В. Ю. Шелепов // Тр. СПИИРАН. — Вып. 31 (2013). — С. 20–42.

A METHOD FOR DETECTION AND LOCALIZATION OF [R] SOUND IN SPEECH SIGNAL

Shelepov V. Ju.,
doctor of physical and mathematical Sciences, Professor, chief scientific officer of PI IPAI (Public Institution Institute of Problems of Artificial Intelligence)

Nitsenko A.V.,
researcher PI IPAI (Public Institution Institute of Problems of Artificial Intelligence)

Abstract

The article presents a method for detecting and localizing a hard and soft trill sound [r] in speech with the use of some features of this sound. An algorithm has been developed that allows localizing the sound [r] regardless of the phonetic environment, thereby increasing the accuracy of automatic speech segmentation, as well as reducing the number of candidate words for recognition.

Keywords: speech segmentation [r],-sound localization, variation, constancy point, bandpass filter