УДК 004.85

СЕГМЕНТАЦИЯ РЕЧЕВЫХ СИГНАЛОВ НА ВОКАЛИЗОВАННЫЕ И НЕВОКАЛИЗОВАННЫЕ УЧАСТКИ НА ОСНОВЕ ОДНОВРЕМЕННОЙ МАСКИРОВКИ*

А. А. Конев, Р. В. Мещеряков, Е. Ю. Костюченко

Томский государственный университет систем управления и радиоэлектроники, 634050, г. Томск, просп. Ленина, 40
E-mail: key@keva.tusur.ru

Рассмотрена модель одновременной тональной маскировки, выделяющая компоненты речевого сигнала, воспринимаемые слуховой системой человека. Предложен алгоритм одновременной маскировки на основе этой модели. Показано, что сигнал после одновременной маскировки представляется бинарной структурой, отражающей гармоническую структуру вокализованной последовательности. Экспериментально доказано, что данная структура может быть использована для выделения ключевых (с точки зрения восприятия слуховой системой) участков речи. На базе этой структуры создан алгоритм высококачественной сегментации речевого сигнала на вокализованные и невокализованные участки, не требующий обучения перед применением. По результатам тестирования совместного использования алгоритмов одновременной маскировки и сегментации речевого сигнала получены оценки качества их работы.

Ключевые слова: речевой сигнал, одновременная маскировка, сегментация речевого сигнала, вокализованные и невокализованные участки.

DOI: 10.15372/AUT20180407

Введение. При восприятии звука преобразование механических колебаний в нервные импульсы осуществляется основной мембраной внутреннего уха, её длина составляет около 35 мм. Каждой точке вдоль основной мембраны можно поставить в соответствие частоту звука, который вызывает максимальный отклик в данной точке. Чем больше расстояние от этой точки, тем ниже амплитуда отклика. Восприятие сигналов сложной формы (в том числе речевого сигнала) характеризуется тем, что отклик будет происходить на все частотные компоненты сигнала. Если амплитуда отклика на компоненту с собственной частотой окажется ниже, чем на другие, то эта компонента слуховой системой восприниматься не будет. Такой эффект называется одновременной маскировкой и позволяет создать дикторонезависимый и не требующий предварительного обучения алгоритм сегментации на вокализованные (с голосовым источником звука) и невокализованные (с шумовым источником звука) участки.

Цель предлагаемой работы — оценка качества нового высокоточного алгоритма сегментации участков на вокализованные и невокализованные.

Математическая модель. Построение модели одновременной маскировки базируется на рассмотрении основной мембраны как набора частотных резонансных фильтров. Модель восприятия речевых сигналов на периферии слуховой системы, включающая математическое представление набора частотных резонансных фильтров, описана в [1–4]. В рамках данной модели основная мембрана представлена в качестве дискретных отсчётов по длине мембраны, каждому из которых соответствует собственный резонансный фильтр.

*Работа выполнена при финансовой поддержке Министерства образования и науки РФ на 2017–2019 гг. (проект № 8.9628.2017/8.9).

Главная задача при построении математической модели одновременной маскировки — оценка влияния частотных компонент, являющихся резонансными для соседних частотных каналов фильтрации, на восприятие сигнала текущим каналом. Подобное влияние можно описать с помощью маскирующей функции, зависящей от интенсивности сигнала в каждом из соседних каналов и от их весовых коэффициентов. Весовые коэффициенты находятся по функции, за основу определения которой была взята реакция фильтров [1, 2] на синусоидальный сигнал:

$$W_0(k, k_1) = \frac{Ww(k, k_1)}{w(k)}, \tag{1}$$

где $Ww(k, k_1)$ — реакция фильтров на синусоидальный сигнал:

$$Ww(k, k_1) = \exp\Big(-2{,}88\Big(Q[k_1]\Big(1 - \frac{\omega[k_1]}{\omega[k]}\Big)\Big)^2\Big); \tag{2}$$

$w(k)$ — нормирующий множитель:

$$w(k) = \sum_{k_1=0}^{K-1} Ww(k, k_1); \tag{3}$$

k, k_1 — номера текущего (маскируемого) и маскирующего частотных каналов фильтрации; $Q(k_1)$ — добротность фильтра с номером k_1 [5]; $\omega(k)$ — резонансная частота фильтра с номером k [5]; K — количество каналов фильтрации.

Данные соотношения используются для получения реакции фильтров. Входящие в них коэффициенты и эксперимент по подбору значений представлены в работах [5, 6].

Маскирующая функция принимает вид

$$M_0(t, k) = \sum_{k_1=0}^{K-1} I(t, k_1) W_0(k_1, k), \tag{4}$$

где t — текущий дискретный отсчёт времени; $I(t, k_1)$ — интенсивность сигнала на канале фильтрации k_1.

Результатом маскировки будет разность значений интенсивности и маскирующей функции в текущем канале:

$$P_0(t, k) = I(t, k) - M_0(t, k). \tag{5}$$

При $P_0 < 0$ маскирующая составляющая превосходит по интенсивности полезную, полностью перекрывая её, в результате чего полезная информация, содержащаяся в данных фрагментах сигнала, не воспринимается слуховой системой человека. Тогда функция

$$P_0(t, k) = \begin{cases} 1, & \text{если } I(t, k) - M_0(t, k) \geq 0, \\ 0, & \text{если } I(t, k) - M_0(t, k) < 0, \end{cases} \tag{6}$$

считается признаком, который выделяет информативные участки сигнала по частоте.

Таким образом, после одновременной маскировки появляется ещё одно представление сигнала, где кроме координат время—частота существует координата, принимающая бинарные значения: $P_0 = 0$ не воспринимается слуховой системой, $P_0 = 1$ воспринимается.

Одновременная маскировка реальных речевых сигналов (фраза «Гаси огонь») приведена на рис. 1. По оси абсцисс даны временны́е отсчёты при частоте дискретизации, равной 12 кГц, по оси ординат — номера частотных каналов фильтрации. Чёрным цве-

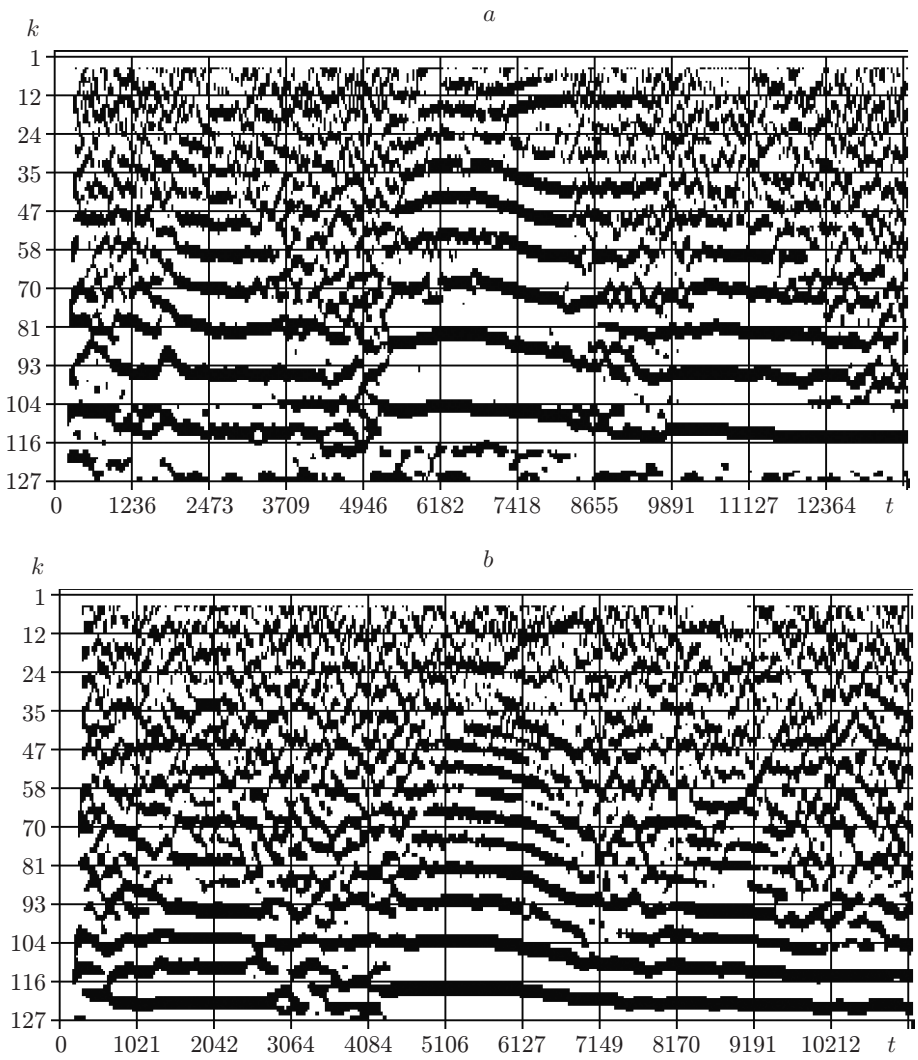

Рис. 1. Структура речевого сигнала после одновременной маскировки: a — диктора-женщины, b — диктора-мужчины

том представлены составляющие речевого сигнала, воспринимаемые слуховой системой ($P_0 = 1$), а белым — невоспринимаемые.

Результаты маскировки и рисунки, аналогичные рис. 1, a, b, были построены для всех реализаций фонем. Экспериментально установлено, что структура всех вокализованных звуков включает, как минимум, два интервала воспринимаемых слуховой системой частотных компонент сигнала ($P_0 = 1$). Подход к сегментации на вокализованные и невокализованные [7, 8] участки основан на определении наличия подобных интервалов на каждом дискретном отсчёте времени. Для реализации данного подхода создаются бинарные шаблоны, включающие частотный срез сигнала после одновременной маскировки. Каждый шаблон состоит из пяти интервалов: двух интервалов из единиц и трёх — из нулей. Пример шаблона приведён в правой части рис. 2. Шаблоны должны отражать структуру сигнала при разных значениях частоты основного тона. Поэтому для каждого номера частотного канала фильтрации, на котором возможно обнаружение основного тона, создаётся собственный шаблон.

Последовательность построения шаблонов для определения вокализации:

1) поиск интервала номеров частотных каналов, на которых возможно обнаружение основного тона;

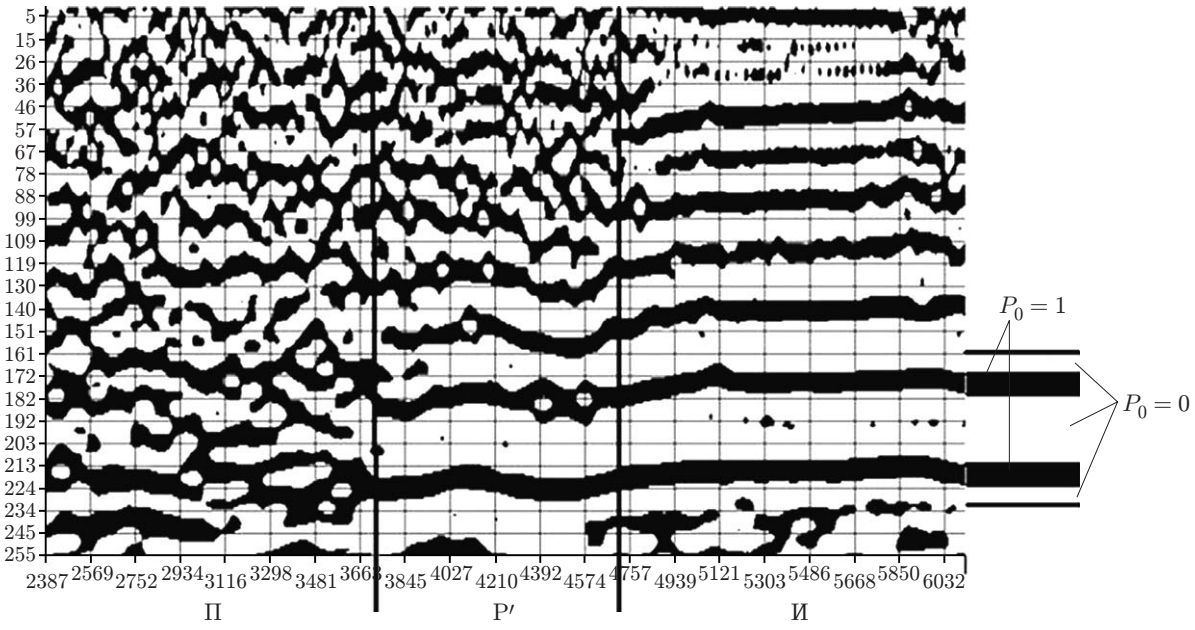

Рис. 2. Подход к построению шаблонов вокализованного сигнала

2) синтез тестовых сигналов для каждого частотного канала из полученного интервала (тестовый сигнал состоит из двух гармоник);

3) формирование набора шаблонов как реакции фильтров на тестовые сигналы с различной частотой основного тона.

Алгоритм сегментации речевого сигнала на вокализованные и невокализованные участки. На основе описанного подхода разработан алгоритм сегментации речевых сигналов на вокализованные и невокализованные участки, который условно можно разделить на два этапа:

1) принятие решения о вокализации (о наличии периодической структуры) сигнала на текущем дискретном отсчёте времени;

2) принятие решения о вокализации сегмента.

Первый этап вокализации сигнала на текущем дискретном отсчёте времени включает в себя определение меры различия d. Оценивается различие между массивом значений, полученных после одновременной маскировки, и каждым шаблоном с помощью

$$d(i,t) = \sum_k F(k,i) \oplus P_0(k,t), \tag{7}$$

где $d(i,t)$ — мера различия сигнала и шаблона; i — номер шаблона; $F(k,i)$ — шаблон для сигнала с частотой основного тона, соответствующей k; $P_0(k,t)$ — структура анализируемого речевого сигнала, получаемая после одновременной маскировки; $\oplus$ — исключающее «или».

Минимумы меры различия находятся в точках совпадения с шаблоном, синтезированным для частотного канала, определяемого частотой основного тона анализируемого сигнала. Таким образом, из набора шаблонов выбирается тот, у которого значение меры различия является минимальным. Принятие решения о вокализации сигнала на текущем дискретном отсчёте времени обусловлено сравнением минимального значения меры различия $d_{\min}$ с заданным порогом min. Если $d_{\min} < \min$, то сигнал на данном отсчёте времени признаётся вокализованным.

Второй этап — это принятие решения о вокализации сегмента речевого сигнала, состоящий из нескольких подэтапов:

— определение непрерывных участков сигнала с одинаковым признаком вокализации входящих в него временны́х отсчётов;

— изменение признака вокализации участков, неверно определённых как вокализованные;

— изменение признака вокализации участков, неверно определённых как невокализованные.

Сам сегмент представляет собой интервал времени, все отсчёты которого имеют один и тот же признак вокализованности. Граница сегмента в данном случае — это временной отсчёт, при переходе от которого к следующему происходит смена признака вокализованности.

Применение только первого подэтапа при расстановке границ сегментов приводит к появлению большого числа вокализованных сегментов малой длительности, а также невокализованных участков, располагающихся в основном на месте образования звонкой смычки. Чаще всего длительность этих участков настолько мала, что не воспринимается слуховой системой как звук речи. Таким образом, необходимо ввести сравнение полученных участков со значением минимальной длительности.

Выбор значения минимальной длительности вокализованного участка определяется результатами психоакустических исследований. Например, в [9] представлены данные о времени, необходимом для восприятия изменения высоты звука человеком. Для частоты основного тона, равной 100 Гц, время восприятия составляет примерно 50 мс, для 1000 Гц — 20 мс. В [2] приведены похожие данные: на низких частотах для распознавания высоты тона требуется примерно 60 мс, для частот от 1 до 2 кГц — 15 мс. В то же время для сложных звуков время увеличивается и для звуков речи может составлять 20–30 мс. Это значит, что для восприятия человеком звука как вокализованного необходима длительность участка не менее 20–30 мс. При реализации алгоритма в качестве порогового было выбрано значение 33 мс, т. е. вокализованные участки, длительность которых меньше пороговой, признавались как невокализованные.

Порог, с которым сравнивалось значение длительности невокализованных участков, был выбран равным 25 мс. Сравнение с этим порогом позволяет исключить короткие невокализованные участки, возникающие внутри вокализованных сегментов.

На рис. 3 в качестве примера представлен результат автоматической сегментации фразы «Гаси огонь» на вокализованные и невокализованные участки. На рисунке приведены значения меры различия $d_{\min}$ на каждом отсчёте времени, порога min, границы между сегментами, полученные при работе алгоритма.

Экспериментальные результаты. Для исследования описанного алгоритма сегментации были проведены следующие эксперименты:

1) выбор значения порога min для определения вокализации речевого сигнала;

2) оценка надёжности сегментации русской слитной речи;

3) оценка надёжности сегментации английской слитной речи.

Параметры, используемые для обработки сигналов: частота дискретизации F_s = = 8 кГц, разрядность сигнала 16 бит, количество каналов 1, верхняя частота анализа $F_{\text{в}} = 2500$ Гц, нижняя частота анализа $F_{\text{н}} = 50$ Гц, количество каналов фильтрации $K = 128$, верхняя граница частоты основного тона $F_{0\text{в}} = 400$ Гц, нижняя граница частоты основного тона $F_{0\text{н}} = 70$ Гц, количество шаблонов для определения вокализованной структуры сигнала 56.

Для оценки надёжности сегментации русской слитной речи использовался речевой материал, включающий отрывок текста, состоящий из 36 сегментов, который был произнесён десятью дикторами (пять мужчин и пять женщин), и отрывок другого текста из

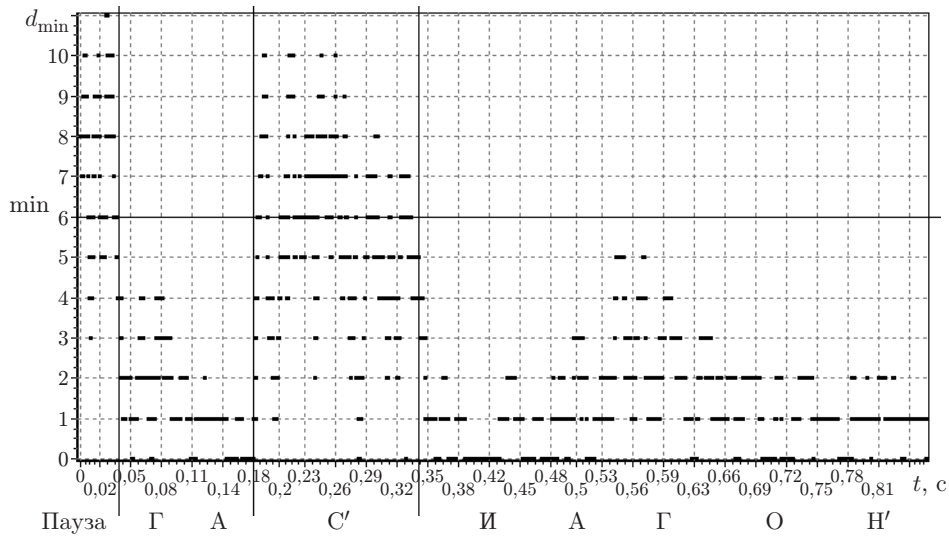

Рис. 3. Результат сегментации речевого сигнала на вокализованные и невокализованные участки

190 сегментов, произнесённый диктором-мужчиной. Общее количество сегментов составило 550. Весь речевой материал был отсегментирован вручную. Полученные временны́е границы между сегментами приняты за эталонные.

Оценка надёжности сегментации англоязычных речевых сигналов проводилась по 30 фразам (208 сегментов), произнесённым на английском языке диктором-мужчиной (носителем языка).

Значение порога min варьировалось от 4 до 8. Максимальные значения надёжности были получены при min = 6, поэтому дальнейшие результаты приведены для данного значения порога.

Надёжность автоматической сегментации оценивалась по следующим критериям: P_- — пропуск существующих границ (количество пропущенных границ по отношению к общему количеству границ в эталоне); P_+ — определение несуществующих границ (количество несуществующих границ по отношению к общему количеству границ в эталоне); $P_{0,01}$ — процент правильно определённых временны́х значений границ (количество границ, временны́е значения которых отличаются от эталонных не более чем на 0,01 с, по отношению к общему количеству границ в эталоне).

Полученные в результате эксперимента усреднённые значения предложенных критериев надёжности автоматической сегментации приведены в таблице.

Разработанный алгоритм не зависит от пола диктора, так как экспериментальные значения критериев надёжности практически не отличаются для дикторов-мужчин и дикторов-женщин.

Значения критериев надёжности автоматической сегментации практически не отличаются для русского и английского языков. Это подтверждает, что сегментация речевого

Критерий надёжности	Значение критерия для русского языка	Значение критерия для английского языка
P_-	0,02	0
P_+	0,09	0,06
$P_{0,01}$	0,91	0,9

сигнала по наличию голосового источника в системе речевосприятия человека является общей для различных языков и не зависит от их фонетического состава.

Заключение. В данной работе рассмотрен подход к математическому представлению одновременной маскировки, а также его применение для сегментации речевого сигнала. Математическое представление основано на особенностях восприятия речевых сигналов слуховой системой человека. Алгоритм сегментации на вокализованные и невокализованные участки, реализованный на основе структуры сигнала после одновременной маскировки, показал высокую надёжность — более 90 % границ между сегментами были определены с погрешностью менее 0,01 с. Полученные результаты подтверждают адекватность предложенного подхода.

СПИСОК ЛИТЕРАТУРЫ

1. **Бондаренко В. П., Мещеряков Р. В., Коцубинский В. П.** Выделение особенностей структуры речевого сигнала // Сб. тр. XIII сессии Росс. акустич. общ-ва. М.: ГЕОС, 2003. Т. 3. С. 63–66.
2. **Алдошина И. А.** Основы психоакустики. Ч. 1 // Звукорежиссер. 1999. № 6. URL: http://lib100.com/music/osnovi_psihoakustiki/pdf/ (дата обращения: 20.06.2018).
3. **Бекеши Г., Розенблат В. А.** Механические свойства уха // Экспериментальная психология. М.: Изд-во иностр. лит., 1963. Т. 2. С. 682–723.
4. **Винников Я. А., Титова Л. К.** Кортиев орган. Гистофизиология и гистохимия. Л.: Изд-во АН СССР, 1961. 260 с.
5. **Конев А. А., Мещеряков Р. В.** Разрешающая способность набора фильтров, моделирующего слуховую систему человека // Сб. тр. XX сессии Росс. акустич. общ-ва. М.: ГЕОС, 2008. С. 23–26.
6. **Бондаренко В. П.** Вопросы моделирования эффектов маскировки // Сб. тез. докл. VIII всесоюз. сем. «Автоматическое распознавание слуховых образов» (АРСО-VIII). Львов, 1974. Т. 1. С. 45–47.
7. **Сапожков М. А.** Речевой сигнал в кибернетике и связи. М.: Гос. изд-во лит-ры по вопросам связи и радио, 1963. 450 с.
8. **Буланин Л. Л.** Фонетика современного русского языка. М.: Высш. шк., 1970. 206 с.
9. **Златоустова Л. В., Потапова Р. К., Потапов В. В., Трунин-Донской В. Н.** Общая и прикладная фонетика: Учеб. пособие. М.: Изд-во МГУ, 1997. 416 с.

Поступила в редакцию 11 октября 2017 г.