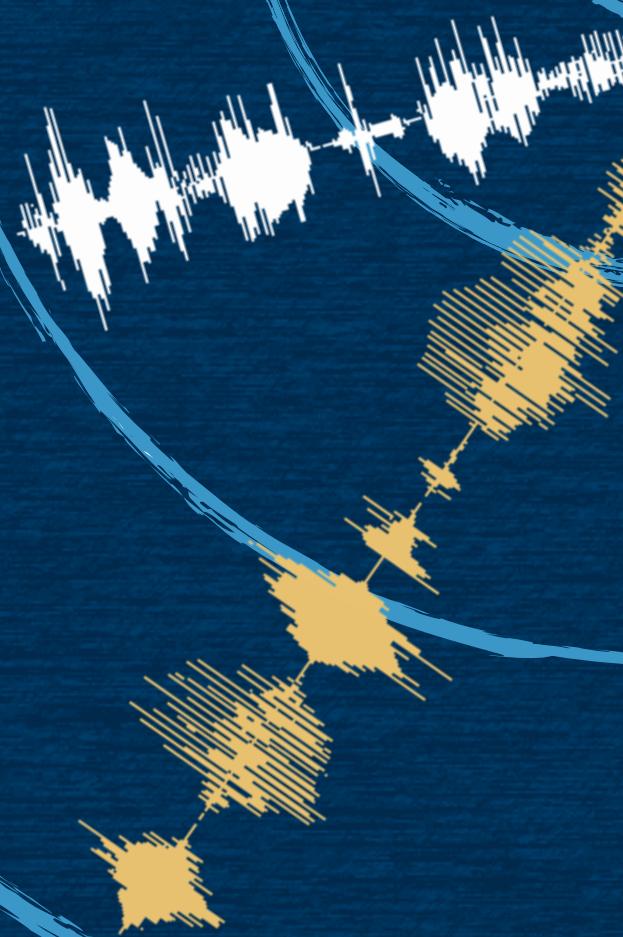


# РЕЧЕВЫЕ ТЕХНОЛОГИИ



ПОМОЖЬ НАУКЕ  
ПРАКТИЧЕСКИЙ ЖУРНАЛ

Speech technology

1-2'

2018



# Речевые Технологии

---

## 1-2/2018

**Главный редактор:**

**Харламов Александр Александрович,**  
доктор технических наук наук, *kharlamov@analyst.ru*

**Состав редколлегии:**

**Заместитель главного редактора: Потапова Родмонга Кондратьевна,**  
доктор филологических наук, профессор, *rkapotapova@yandex.ru*

**Голенков Владимир Васильевич,** доктор технических наук, профессор,  
*golen@bsuir.by*, Беларусь

**Женило Валерий Романович,** доктор технических наук, профессор,  
*zhenilo@yandex.ru*

**Жигулёвцев Юрий Николаевич,** кандидат технических наук, *ynzh@mail.ru*

**Карпов Алексей Анатольевич,** доктор технических наук, *kagrov@iias.spb.ru*

**Кривнова Ольга Федоровна,** доктор филологических наук, профессор,  
*okrivenova@mail.ru*

**Кудубаева Сауле Альжановна,** кандидат технических наук,  
*saule\_58@mail.ru*, Казахстан

**Кушнир Алексей Михайлович,** кандидат психологических наук,  
*kushnir-narobrig@yandex.ru*

**Кушнир Дмитрий Алексеевич,** кандидат технических наук,  
*kushdal@yandex.ru*

**Лобанов Борис Мефодьевич,** доктор технических наук,  
*lobanov@newman.bus-net.by*, Беларусь

**Ляксо Елена Евгеньевна,** доктор биологических наук, *lyakso@gmail.com*

**Максимов Евгений Михайлович,** доктор технических наук, *maximovem@inbox.ru*

**Матвеев Юрий Николаевич,** доктор технических наук, *matveev@mail.ifmo.ru*"ru

**Мещеряков Роман Валерьевич,** доктор технических наук, профессор,  
*mrv@ieeee.org*

**Петровский Александр Александрович,** доктор технических наук, профессор,  
*palex@bsuir.by*

**Ромашкин Юрий Николаевич,** кандидат технических наук, *gomaup@yandex.ru*

**Ронжин Андрей Леонидович,** доктор технических наук, профессор,  
*ronzhin@iias.spb.su*

**Сажок Николай Николаевич,** кандидат технических наук,  
*sazhok@gmail.com*, Украина

**Сулейманов Джавдет Шевкетович,** академик Академии наук Татарстана,  
профессор, *alsu\_73@list.ru*

**Чучупал Владимир Яковлевич,** кандидат технических наук,  
*v.chuchupal@gmail.com*

**Milos Zelezny,** *zelezny@kky.zcu.cz*, Чехия



## Содержание

Чучупал В.Я.

Неявная модель произношения для автоматического распознавания речи .....	3
--	---

Вашкевич М.И., Азаров И.С., Петровский А.А.

Оценка мгновенной частоты основного тона речевого сигнала на основе многоскоростной обработки .....	12
---	----

Крейчи С.А., Кривнова О.Ф., Тихонова Е.А.

Помехоустойчивость слоговых таблиц при восприятии речи в шуме .....	25
---	----

Солонина Е.Г.

Акустические и перцептивные признаки коартикуляционной назализации гласных в русском языке .....	37
--	----

Ляксо Е.Е., Фролова О.В., Григорьев А.С., Городный В.А.

Сравнительный анализ характеристик голоса и речи детей типично развивающихся, с расстройствами аутистического спектра, синдромом Дауна и умственной отсталостью .....	50
---	----

Вашкевич М.И., Азаров И.С., Петровский А.А.

Speech enhancement in a smartphone-based hearing aid ....	63
---	----

Фролова О.В., Бедалова Ш.Г., Ляксо Е.Е.

Особенности речевого развития детей дошкольного возраста с нарушениями развития, воспитывающихся в детском доме .....	82
---	----

Бирин Д. А., Булашевич А.Е., Грекис М.Ю.

Задача автоматической расстановки знаков пунктуации в распознанной спонтанной русской речи .....	94
--	----

Козлачков С.Б., Дворянкин С.В., Василевская Н.В.

Фонетическая функция А.А. Пирогова и помехоустойчивость канала речевой коммуникации .....	105
---	-----

### Редакция:

Редактор: Татьяна Иванова

Корректор: Людмила Асанова

Дизайн: Анна Ладанюк, Марина Столбова

Вёрстка: Ксения Мельникова

Адрес редакции: 109341, Москва, ул. Люблинская, д. 157, корп. 2.

Тел.: 8 495 345 59 00

Подписано в печать 27.02.2018. Формат 60×90%. Бумага офсетная. Печать офсетная.

Печ. л. 14,25. Заказ № 0628. Издательский дом «Народное образование».

Отпечатано в типографии НИИ школьных технологий.

143500, Москва, ул. Люблинская, д. 157, корп. 2. Тел.: (495) 345 52 00/59 00/59 01

© «Народное образование»

# Неявная модель произношения для автоматического распознавания речи

**Владимир Яковлевич Чучупал,**  
в.н.с., Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН

## Аннотация

Вариативность произнесения слов в естественной разговорной речи является одним из основных источников ошибок при ее автоматическом распознавании. Примером подобной вариативности является пропуск или подмена отдельных звуков, вызванная неполной или нечеткой артикуляцией в быстрой речи.

В статье описана неявная модель произношения, которая реализована посредством сглаживания параметров акустических моделей соседних звуков.

Предлагается использовать контексто-зависимые параметры сглаживания, которые обусловлены текущим фонетическим, просодическим и языковым контекстом звуков. Хотя подход к моделированию вариативности произношения уже обсуждался в литературе, метод контексто-зависимого сглаживания моделей смежных звуков, насколько известно автору, пока не был представлен.

Эксперименты на речевом корпусе данных, который содержал как читаемую, так и естественную речь, показали корректность предложения использовать переменные параметры сглаживания, значение которых обусловлено фонетическим и просодическим контекстом.

**Ключевые слова:** автоматическое распознавание речи, обработка естественного языка, акустическое моделирование речи, модели вариативности произношения

## ВВЕДЕНИЕ

Это было подтверждено симуляционными экспериментами ~\cite{Saraclar Nock, McAllaster} в которых использование корректных фактических транскрипций вместо канонических привело к снижению уровня ошибок распознавания почти в два раза.

Существует два основных подхода к моделированию вариативности произношения. Явный подход описывает вариабельность произнесения слова путем описания возможных изменений в базовой фонемной транскрипции ~\cite{Wester} слов.

Неявный подход ~\cite{Saraclar} описывает изменчивость произношения посредством изменений в структуре акустических моделей звуков, не изменяя фонемные транскрипции.



Системы распознавания речи обычно реализуют явный подход, поскольку он выглядят естественнее и может быть просто описан в терминах классической модели распознавания речи.

Пусть  $X = \{x^t\}$ ,  $t = 1, \dots, T$  последовательность наблюдений, векторов речевых параметров, а  $W = \{w_i\}$ ,  $i = 1, \dots, N$  последовательность слов. Наиболее вероятная последовательность слов  $W^*$  при известных  $X$  может быть получена из следующего выражения [5]:

$$\begin{aligned} W^* &= \arg \max_W P(W|X) \\ &= \arg \max_W \frac{P(X|W)P(W)}{P(X)} \\ &= \arg \max_W P(X|W)P(W). \end{aligned} \quad (1).$$

Первый сомножитель  $P(X|W)$  в числителе (1) соответствует правдоподобию заданных наблюдений и вычисляется с помощью акустической модели. Вероятность последовательности слов  $P(W)$  вычисляется с помощью модели языка. Знаменатель  $P(X)$ , по сути, является нормализующим членом.

Обозначим транскрипцию слова  $w$  как  $t^w$ , множество из всех транскрипций слова  $w$  обозначим как  $T^w$ . Множество всех возможных транскрипций для последовательности слов  $W$  обозначим через  $T^W$ . Обозначение  $t^w$  будет использовано для обозначения любого элемента  $T^W$ . Тогда условие (1) можно аппроксимировать (аппроксимация Витерби) следующим выражением:

$$W^* = \arg \max_{W, t^W \in T^W} P(X|t^W)P(t^W|W)P(W). \quad (2).$$

Оценка величины  $(t^W|W)$  выполняется на основе модели вариативности произношения. Параметрами этой модели являются фонемные транскрипции из  $T^W$ , а также их относительные частоты  $\{P(t^W|W), t^W | T^W\}$ . Таким образом, явная модель вариативности произношения может быть описана на основе методов определения фонемных транскрипций слов и их вероятностей.

Наиболее очевидный способ выбора возможных фонемных транскрипций основан на использовании фонемного транскриптора, т.е. системы распознавания фонем в потоке речи. На практике это пока недостаточно эффективно из-за низкой точности существующих транскрипторов. В качестве оценки вероятности фонемных транскрипций естественно использовать их экспериментальные относительные частоты. Однако это часто не представляется возможным, поскольку требует очень больших корпусов речевых данных.

В литературе описаны способы преодоления описанных трудностей, которые, в частности, основаны на дискриминантном и непараметрическом подходах [3, 6, 7, 8, 9], однако выигрыш в пословной точности распознавания речи, который получается при использовании явных моделей произношения, далеко не так существенен, как этого можно было бы ожидать, исходя из результатов симуляционных экспериментов.

Основная идея явного моделирования заключается в том, что все возможные изменения произношения могут быть достаточно точно представлены изменениями в базовой фонемной транскрипции слов, т.е. заменами, вставками и удалениями фонем. Экспериментальный анализ показывает, что альтернативное описание изменчивости произношения, особенно в спонтанной речи, может быть сделано на основе использования моделей, которые способны представлять частичные изменения фонемного качества, т.е. путем неявного моделирования произношения [4].

Модели неявного моделирования вариативности произношения в отличие от явных реализуются на основе более сложных акустических моделей фонем, которые используются в базовых фонемных транскрипциях. Например, марковских моделей в виде сети из состояний. Тем не менее, как следует из [4, 11, 12, 13, 14, 15] результат использования неявных моделей — повышение точности распознавания — существенно не отличается от такового для явных моделей.

Опубликованы исследования, в которых показано, что уровень вариативности произношения часто имеет условный характер [16], т.е. зависит от окружающего контекста. Это обстоятельство можно интерпретировать как возможность предсказать уровень вариативности произношения слов на основе текущих просодических, синтаксических и семантических характеристик речи.

## КОНТЕКСТО-ЗАВИСИМАЯ НЕЯВНАЯ МОДЕЛЬ ВАРИАТИВНОСТИ ПРОИЗНОШЕНИЯ

Пусть  $m$  и  $\$n\$$  обозначают модели звуков,  $P(x|m)$ ,  $P(x|n)$  обозначают условные вероятности для наблюдений речевых параметров  $\$x\$$  при заданных моделях. Тогда модель  $m$ , которая сглаживается моделью  $n$ ,  $P(x|m,n)$ , может быть определена аналогично [14]:

$$P_\lambda(x|m,n) = \lambda * P(x|m) + (1 - \lambda)P(x|n), \quad 0 \leq \lambda \leq 1. \quad (3)$$

Выражение (2) позволяет описать, пусть в упрощенной форме, некоторые часто наблюдаемые в спонтанной речи произносительные изменения.

Например, значение  $\lambda$ , равное 0, означает, что звук  $m$  полностью заменен звуком  $\$n\$$ . Аналогично  $\lambda = 0,5$  описывает ситуацию неполной замены фонетического качества, которая упрощенно соответствует эффектам назализации, озвончения или оглушения звуков в спонтанной речи.

Пусть звук  $m$  наблюдается на временном интервале  $s(m), \dots, e(m)$  с соответствующими параметрами  $x_{s(m)}, \dots, x_{e(m)}$ . Оптимальное значение коэффициента сглаживания  $\lambda$  в (2) может быть найдено аналогично оценке оптимальных весов смесей при обучении модели смеси нормальных распределений [17]:

$$\lambda_{m,n} = \frac{\sum_{t=s(m)}^{e(m)} P(x_t|m)}{\sum_{t=s(m)}^{e(m)} (P(x_t|m) + P(x_t|n))}. \quad (4)$$

Для корпуса данных  $U$ , который состоит из  $R$  высказываний  $U = \{u^r | r = 1, \dots, R\}$ , значение  $\lambda_{m,n}$  может быть вычислено усреднением локальных оценок (4) для всех наблюдений пар фонем  $(m,n)$  аналогично оценкам параметров марковских моделей [17]?



$$\hat{\lambda}_{(m,n)} = \frac{\sum_{r=1}^R \sum_{(m,n) \in u_r} \lambda_{(m,n)} P(m)}{\sum_{r=1}^R \sum_{(m,n) \in u_r} P(m)}, \quad (5)$$

где  $P(m)$  правдоподобие звука  $m$  при наблюдении  $(m,n)$ .

Пусть  $V$  — это вектор контекстных признаков, то есть признаков фонемного, позиционного, просодического, лингвистического и синтаксического контекста, наличие которых [16] коррелирует с наблюдаемым уровнем произносительной вариативности:

$V(c, l, r; nPh, pPOS, ROS, wPOS, POS, eWrd, LM)$ :

where

$c$ : центральная фонема,

$l$ : левый фонемный контекст,

$r$ : правый фонемный контекст,

$nPh$ : следующая фонема,

$pPOS$ : позиция в слове,

$ROS$ : темп речи,

$wPOS$ : позиция слова во фразе,

$POS$ : часть речи слова,

$LM$ : значение модели языка.

(6)

Экспериментально проверим предположение о том, что уровень произносительной вариативности, описываемый значением  $P(\lambda|V)$ , действительно зависит от признаков (6).

## ЧИСЛЕННОЕ ИССЛЕДОВАНИЕ

Доказательство того, что модели звуков [2] с использованием интерполированных параметров могут существенно снизить уровень ошибок, лучше всего сделать на основе эксперимента по распознаванию речи. Однако такой эксперимент требует внедрения интерполированных моделей в процедуры обучения и распознавания.

Предварительные эксперименты, которые доказывают эффективность предложенных моделей, пусть и косвенно, могут быть выполнены путем оценки параметра  $\lambda$  на корпусе данных, чтобы показать, что значение параметра существенно зависит от контекстных признаков (6).

Для экспериментальных исследований использовался речевой материал из корпусов данных с русской устной речью: TeCoRus [18], RuSpeech [19] и PronExRu [20].

Речевой материал был разделен на три выборки, предназначенные для обучения, настройки и тестирования. Обучающая выборка состояла из речевого материала корпусов RuSpeech и TeCoRus, которые содержали читаемую речь от 200 дикторов. Настроочные данные использовались для оценки значений параметров интерполяции. Они состояли из тестового материала корпуса RuSpeech, всего 1000 высказываний от 10 человек. Наконец, тестовый материал состоял из данных PronExRu (200 высказываний, 9 человек).

Обучающие данные использовались для создания гендерно-ориентированных акустических моделей. Оценка значений признаков в (6), за исключением признака темпа речи ROS (rate of speech), выполнялась двумя этапами с использованием результатов автоматического распознавания речи, полученных на первом этапе.

Оценка параметра ROS была выполнена с использованием алгоритма [13]. Чтобы понизить вычислительную сложность, вместо полного перебора возможных комбинаций длительностей звуков использовалась процедура сэмплинга длительностей.

В ходе предварительных экспериментов дискретные значения признаков были оценены для каждого центрального состояния марковских моделей. Затем значения параметров  $\lambda$  были вычислены с использованием (5).

Чтобы иметь возможность ранжировать признаки, строились два бинарных дерева решений. Одно дерево было построено с использованием читаемого речевого материала корпуса TeCoRus. Другое дерево было построено с использованием спонтанной речи корпуса PronEx.

Набор вопросов для выращивания деревьев содержал всевозможные вопросы о наличии признаков (6) для текущего состояния (например: «Применяется ли это состояние к существительному?», «Является ли это состояние частью слова?»), кроме того, вопросник включал набор стандартных вопросов, которые используются при построении фонемных бинарных деревьев решений для синтеза алфавита контексто-зависимых состояний марковских моделей аллофонов. Всего использовалось 82 вопроса: 30 вопросов, относящихся к признакам из (6), и 52 вопроса, относящиеся к фонемному контексту звука и его фонемному качеству.

В качестве критерия расщепления вершин при выращивании деревьев использовалась величина информационного выигрыша [21], изменения энтропии при разделении родительской вершины на две дочерние после использования вопроса:

$$\Delta H_\lambda(t, f) = H_\lambda(t) - \frac{|t_+|}{|t|} H_\lambda(t_+) - \frac{|t_-|}{|t|} H_\lambda(t_-). \quad (7)$$

Здесь  $H_\lambda(t)$ ,  $H_\lambda(t_-)$ ,  $H_\lambda(t_+)$  обозначают величину энтропии для данных в родительской вершине  $t$  и энтропию данных в дочерних вершинах  $t_-$  и  $t_+$ , которые образовались после бинарного разделения родительской вершины  $t$ .

Чтобы численно оценить важность каждого признака, рассчитывались значения весов признаков. Для этого для каждого данного признака по всем вершинам дерева, где вопрос об этом признаком был выбран как лучший разделитель вершины, значения величины информационного выигрыша суммировались. То есть вычислялось суммарное изменение выигрыша (7) для данного признака. Затем эти суммарные изменения нормировались величиной максимального значения к 100 (т.е. самый важный признак имел вес 100):

$$I(f) = \sum_{t \in T} \Delta H_\lambda(t, f), \quad (8)$$

где суммирование выполняется по всем вершинам дерева.

Результат построения деревьев содержится в таблице 1, которая содержит список наиболее важных признаков для настроенной и тестовой частей корпуса данных.



Таблица 1

## Наиболее важные признаки вариативности

Признак	Вес для настроочных данных	Признак	Вес для тестовых данных
ROS-noun	100	RATE-F	100
RATE-MDL	72	L-Labial	70
L-Soft	55	STRESS	63
RATE-F	42	POS-noun	54
R-VoiceLess	40	L-VOW	52
L-Sonant	37	PpMdl	51
R-Forv	35	R-Labial	42
L-Forv	31	NEXT-VOW	39
POS-adj	30	RATE-FST	39
R-Sil	29	R-Forv	35
STRESS	29	R-Forw	29

Сокращения в названиях признаков имеют следующее значение:

"POS-noun" обозначает «существительное»,  
"POS-adj" обозначает «имя прилагательное»,  
"RATE-F" обозначает «быстрый темп речи»,  
"RATE-mdl" обозначает «умеренный темп речи»,  
"STRESS" обозначает «ударный звук»,  
"PpMdl" обозначает «звук в середине слова»,  
"R-sil" обозначает «звук после паузы»,  
"L-soft" обозначает «предыдущий звук — мягкий согласный»,  
"R-VoiceLess" обозначает «звук перед паузой»,  
"L-Sonant" обозначает «предыдущий звук звонкий согласный»,  
"R-Forv" обозначает «следующий звук согласный передний»,  
"R-Forw" обозначает «следующий звук — передний гласный»,  
"NEXT-VOW" обозначает «следующий звук — гласный»,  
"R-Labial" обозначает «следующий звук — губной согласный».

Значения сглаживающего параметра  $\lambda$ , измеренные с помощью (4), большую часть времени находились в диапазоне от 0,2 до 0,8, и значение правдоподобия данных для сглаженных моделей было выше, чем при исходных акустических моделях.

Как следует из таблицы 1, наиболее важным признаком для читаемого речевого материала (речевого корпуса RuSpeech) является признак части речи. Для спонтанной речи (корпус Pronex) наиболее важным признаком является признак темпа речи. Оба эти признака относятся к предлагаемому набору признаков вариативности произношения (6). Из одиннадцати наиболее важных признаков для материала читаемой речи, как следует из таблицы, 6 признаков, которые входят в набор (6). Для материала спонтанной речи таких признаков 5.

Если рассмотреть наиболее важные признаки с весом более 50, то для читаемой речи таких признаков всего три, и два из них относятся к набору (6). Для спонтанной речи таких признаков шесть, причем четыре из них из набора (6).

Наиболее важными для обоих корпусов данных являются признаки, связанные с темпом речи, частью речи и положением звука относительно ударения. Исходя из этих предварительных результатов, можно утверждать, что значение  $\lambda$ , определенное в соответствии с [5], действительно зависит от значений признаков, перечисленных в [6]. Предлагаемые признаки вариативности произношения, основанные на просодических и позиционных контекстах слова и звука, коррелируют с наблюдаемыми изменениями акустического качества, т.е. с изменчивостью произношения.

Таким образом, модель (3), основанная на сглаживании параметров соседних акустических моделей, может использоваться для учета эффектов изменчивости произношения при распознавании естественной речи. Такая модель может быть реализована, например, путем организации пост-обработки как второй проход при распознавании.

## **ЗАКЛЮЧЕНИЕ**

В статье предложена неявная модель вариативности произношения как способ повышения точности автоматического распознавания речи. Произносительные изменения в разговорной речи предлагается учитывать за счет использования акустических моделей звуков со сглаженными параметрами, так, что величина параметров, которые регулируют уровень, зависит от контекстных и позиционных признаков звука и слова, в котором он находится.

Предложен набор из нескольких потенциальных контексто-зависимых признаков вариативности и численно исследована их ценность как предсказателей вариативности. Для этого на материале читаемой и спонтанной речи с использованием вопросника, который содержал как стандартные для построения фонетических решающих деревьев вопросы, так и вопросы, относящиеся к потенциальным признакам вариативности, были построены решающие деревья. Все признаки были ранжированы по важности в соответствии с метрикой информационного выигрыша.

Было установлено, что ряд предложенных контексто-зависимых признаков вариативности фактически коррелирует с наблюдаемой изменчивостью произношения, превосходя по важности большинство обычных признаков фонемного контекста звука. Таким образом, предложенная модель, основанная на сглаживании параметров, может быть использована для учета эффектов изменчивости произношения в естественном распознавании речи.

## **Список литературы**

1. Sarclar M., Nock H., Khudanpur S. Pronunciation modeling by sharing Gaussian densities across phonetic models // Computer Speech and Language. 2000. Vol. 14(4). P. 137–160.
2. McAllaster D., Gillick L., Scattone F., Newman M. Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch. // Int.Conf. Speech and Language Processing, 1998, Sydney, P. 1847–1850.



3. *Wester M.* Pronunciation modeling for ASR — knowledge-based and data-derived methods // Computer Speech and Language, 2003. Vol. 17, P. 69–85.
4. *Saraclar M., Khudanpur S.* Pronunciation change in conversational speech and its implications for automatic speech recognition // Computer Speech and Language 2004. Vol. 18(4). P. 375–395.
5. *Jelinek F.* Statistical Methods for Speech Recognition // The MIT Press, Cambridge, Massachusetts, 1997.
6. *Lehr M., Gorman K., Shafran I.* Discriminative pronunciation modeling for dialectal speech recognition. // Proc. Conf. of International Speech Communication Association, Interspeech, 2014, Singapoure, pp. 1458–1462, 2014.
7. *Byrne B., Finke M., Khudanpur S., McDonough J., Nock H., Riley M., Saraclar M., Wooters C., Zavaliagkos G.* Pronunciation modelling for conversational speech recognition: a status report from WS97. // IEEE Workshop on Automatic Speech Recognition and Understanding. 1997, USA, P. 26–33. 10.1109/ASRU.1997.659004
8. *Hutchinson B., Droppo J.* Learning Non-Parametric Models of Pronunciation in automatic speech recognition // Proc. International Conference on Acoustics, Speech, and Signal Processing, ICASSP, USA, 2011, P. 4904–4907.
9. *Schramm H.* Modeling Spontaneous Speech Variability for Large Vocabulary Continuous Speech Recognition. // Doktors der Naturwissenschaften Dissertation, Technical University of Aachen, Germany, 2006.
10. *Livescu L., Glass J.* Feature-based pronunciation modeling for speech recognition // Proc. Human Lang. Tech. Conf. of the North American Chapter of the Assoc. for Comp. Ling., USA, 2004.
11. *Hain T., Woodland P.C.* Dynamic HMM selection for continuous speech recognition // Proc. of EuroSpeech, 1999. P. 1327–1330.
12. *Hain T.* Implicit modelling of pronunciation variation in automatic speech recognition // Speech Communication. Vol 46 (2005). P. 171–188.
13. *Zheng J., Franco H., Stolcke A.* Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition // Speech Communication. 2003. Vol. 41. P. 273–285.
14. *Liu Y.* Modeling partial pronunciation variations for spontaneous Mandarin speech recognition // Computer Speech \& Language. Vol. 17. No 4, 2003. P. 357–379.
15. *Spiess T., Wrede B., Fink G.A., Kummert F.* Data-driven Pronunciation Modeling for ASR using Acoustic Subword Units // Int Conf. InterSpeech, 2003. P. 2549–2552.
16. *Ostendorf M., Shafran I., Bates R.* Prosody Models For Conversational Speech Recognition // 2nd Plenary Meeting and Symposium on Prosody and Speech Processing, USA, 2003. P. 147–154.
17. *Rabiner L., Biing-Hwang J.* Fundamentals of Speech Recognition.// PTR, Prentice Hall Signal Processing Series, New Jersey, USA, 1993.
18. Чучупал В.Я., Маковкин К.А., Чичагов А.В., Кузнецов В.Б., Огарышев В.Ф. Речевой корпус данных TeCoRus // Свидетельство об официальной регистрации базы данных №2005620205, 2005 г.
19. *Arlazarov, V.L., Bogdanov, D.S., Krivnova, O.F., Podrabinovitch, A.Ya.* Creation of Russian Speech Databases: Design, Processing, Development Tools, . Proceedings of the Intern. Conference SPECOM'2004, 4Pp. 650–656 , S-Pb., Russia, 2004.
20. Russian spoken speech corpus //Database registration certificate~2016620687, Rospatent, Moscow, 2016 (in Russian).
21. *Quinlan, J.R.* Induction of decision trees. // Machine Learning 1, 81–106, 1986.

## **IMPLICIT PRONUNCIATION VARIATION MODEL FOR AUTOMATIC SPEECH RECOGNITION**

**Vladimir J. Chuchupal,**

*leading scientific researcher, Computing center. A. A. Dorodnicyn FITS Yiwu wounds*

### **Abstract**

The variations in pronunciation of words in natural speech are one of the main sources of automatic speech recognition errors. The examples of such variations include the pronunciation variations that are caused by a fuzzy or an incomplete articulation that is frequently observed in spontaneous speech.

The implicit pronunciation model is proposed that is implemented by means of smoothing of parameters of the adjacent acoustical phone models in phonemic transcription. It is proposed to use the context-dependent smoothing, so that the values of the smoothing parameters are conditioned by the current position and prosodic contexts of a phone.

While the pronunciation variation modeling approach on the base of combination of acoustical models has already been discussed in literature, the method based on the context-dependend smoothing of the adjacent models as far as we know has not been published yet.

The experiments on the speech corpuses that contained both the read and spontaneous speech showed the correctness of the proposal for the use of the context-dependent smoothing parameters which are conditioned by the features of phonemic context and prosody.

**Keywords:** automatic speech recognition, natural speech processing, acoustic modeling, pronunciation variation modeling



# Оценка мгновенной частоты основного тона речевого сигнала на основе многоскоростной обработки

**Максим Иосифович Вашкевич,**  
кандидат технических наук, доцент Белорусского государственного университета информатики и радиоэлектроники (БГУИР)

**Илья Сергеевич Азаров,**  
доктор технических наук, доцент БГУИР

**Александр Александрович Петровский,**  
доктор технических наук, профессор кафедры электронных вычислительных средств БГУИР

## Аннотация

В работе предлагается алгоритм оценки частоты основного тона, основанный на представлении речевого сигнала синусоидальной моделью с мгновенными параметрами. Алгоритмом предусмотрена следующая последовательность шагов: 1) декомпозиция сигнала на субполосные составляющие; 2) определение мгновенных параметров синусоидальной модели субполярных сигналов; 3) вычисление функции формирования кандидатов периода основного тона; 4) поиск локального контура частоты основного тона. Особенностью алгоритма является то, что ширина полос пропускания фильтров, используемых для декомпозиции, а также длительность кадра анализа масштабируются для каждого кандидата периода основного тона путем передискретизации сигнала. В работе делается сравнение предлагаемого алгоритма с широко используемыми оценщиками частоты основного тона RAPT, YIN, SWIPE', IRAPT и PEFAC. Предлагаемый алгоритм демонстрирует хорошее частотное и временное разрешение для сигналов, имеющих значительную частотную модуляцию, и показывает хорошую производительность как для чистых, так и для зашумленных сигналов.

**Ключевые слова:** частота основного тона, многоскоростная обработка

## ВВЕДЕНИЕ

Надежное определение частоты основного тона требуется во многих приложениях обработки речи. В большинстве параметрических моделей, применяемых при кодировании, преобразовании и синтезе речевых сигналов, требуется оценка вокализованности/невокализованности и значение частоты основного тона. Использование алгоритма определения частоты основного тона с хорошим временным разрешени-

ем особенно необходимо для анализа/синтеза нестационарных звуков, которые обычно происходят на границах вокализованных сегментов и в моменты переходов. В то же время точная оценка контура частоты основного тона имеет большое значение при частотном анализе речевых сигналов, синхронизированном с частотой основного тона [1]. Понятие мгновенной частоты может быть естественным образом применено к частоте основного тона, если предположить, что речевой сигнал описывается гармонической моделью [2, 3]. Контур мгновенной частоты основного тона может быть извлечен из вокализованных участков речи, если рассматривать их как непрерывный и нестационарный процесс [4, 5].

Идея точной оценки частоты основного тона, разработанная в [6, 7], основана на декомпозиции сигнала на узкополосные компоненты и использовании их мгновенных частот в качестве исходных данных. Данный подход нашел применение в анализе речи и певческого голоса [8, 9], а также при создании устойчивого к ошибкам оценщика основного тона в [10]. Тем не менее у данного подхода есть фундаментальные ограничения, связанные с принципом неопределенности: нельзя достичь одинаково высокого разрешения для всего диапазона частот основного тона, используя банк фильтров с фиксированными параметрами. Для частотно-временного анализа в случае низкого голоса лучше использовать кадры большой длины, а для высоких голосов предпочтительно иметь короткую длину кадра и более широкие полосы у фильтров анализа. Компромиссное решение было найдено в [7], где использовались кадры длительностью 50 мс и фильтры с шириной полосы в 70 Гц, которое позволяло довольно точно оценить основной тон на женских голосах, но, как оказалось, подвержено грубым ошибкам на низких мужских голосах [11].

В данной статье описан алгоритм тонкой оценки частоты основного тона, основанный на многоскоростной схеме анализа сигнала. Главной идеей является достижение улучшения в точности оценки за счет подстройки параметров банка анализирующих фильтров для каждого кандидата периода основного тона. Предлагаемый алгоритм является эффективным для анализа как низких, так и высоких голосов. В алгоритме также предполагается, что вариация частоты основного тона пропорциональна ее текущему значению. Увеличение ширины полосы анализирующих фильтров, необходимое для коротких кандидатов периода, приводит к смешиванию гармоник в субполосных сигналах при анализе низких голосов. Для компенсации этого эффекта предлагается использовать специального вида функцию формирования кандидатов периода (ФФКП) основного тона, которая менее чувствительна к смешению гармоник. В работе приводится как теоретическое обоснование предлагаемого алгоритма, так и практические аспекты его реализации. В заключительной части статьи оценивается производительность алгоритма на чистых и зашумленных сигналах.

## **1. МНОГОСКОРОСТНАЯ СХЕМА ВЫЧИСЛЕНИЯ ФУНКЦИИ ФОРМИРОВАНИЯ КАНДИДАТОВ ПЕРИОДА ОСНОВНОГО ТОНА**

Предлагаемый оценщик частоты основного тона основан на синусоидальной модели, которая представляет детерминированную часть сигнала в виде суммы периодических компонент с нестационарными параметрами:

$$s(n) = \sum_{k=1}^K A_k(n) \cos(\phi_k(n)) + r(n), \quad (1)$$



где  $\Phi_k(n) = \sum_{i=1}^n \omega_k(n) + \phi_k(0)$ ,  $K$  — количество периодических компонент и  $r(n)$  — шумовая компонента. Параметры данной модели (мгновенная амплитуда  $A_k(n)$  и частота  $\omega_k(n)$  в рад/отсчет) используются в качестве начальных данных для оценки частоты основного тона. Для получения параметров модели сигнал  $s(n)$  раскладывается на комплексные субполосные составляющие ДПФ-модулированным банком фильтров, краткое описание которого приводится далее.

Равномерная сетка частот, соответствующая центральным частотам банка фильтров анализа, определяется как  $k\omega_{step}$ ,  $k = 1, 2, \dots, K$ ,  $K\pi/\omega_{step}$ , где  $\omega_{step}$  — шаг по частоте в рад/отсчет. Импульсная характеристика  $k$ -го анализирующего фильтра определяется выражением:

$$h_k(n) = 2 \frac{\sin(\omega_{bw}n)}{\pi n} w(n) e^{jkn\omega_{step}}, \quad (2)$$

где  $w_{bw}$  — половина ширины полосы пропускания фильтра и  $w(n)$  — четная оконная функция.

Выход каждого канала банка фильтров является аналитическим сигналом  $S_k(n)$  с ограниченной полосой, который можно представить как свертку входного сигнала  $s(n)$  с импульсной характеристикой:

$$S_k(n) = \sum_{i=-\infty}^{\infty} h_k(i)s(n-i) = \operatorname{Re}(S_k(n)) + j\operatorname{Im}(S_k(n)). \quad (3)$$

Мгновенные параметры субполосных компонент могут быть получены следующим образом:

$$A_k(n) = \sqrt{\operatorname{Re}(S_k(n))^2 + \operatorname{Im}(S_k(n))^2}, \quad (4)$$

$$\phi_k(n) = \arctan\left(\frac{-\operatorname{Im}(S_k(n))}{\operatorname{Re}(S_k(n))}\right), \omega_k(n) = \dot{\phi}_k(n). \quad (5)$$

Чтобы избежать точек разрывов, в функции (5) применялась процедура развертывания фазы (*phase unwrapping*).

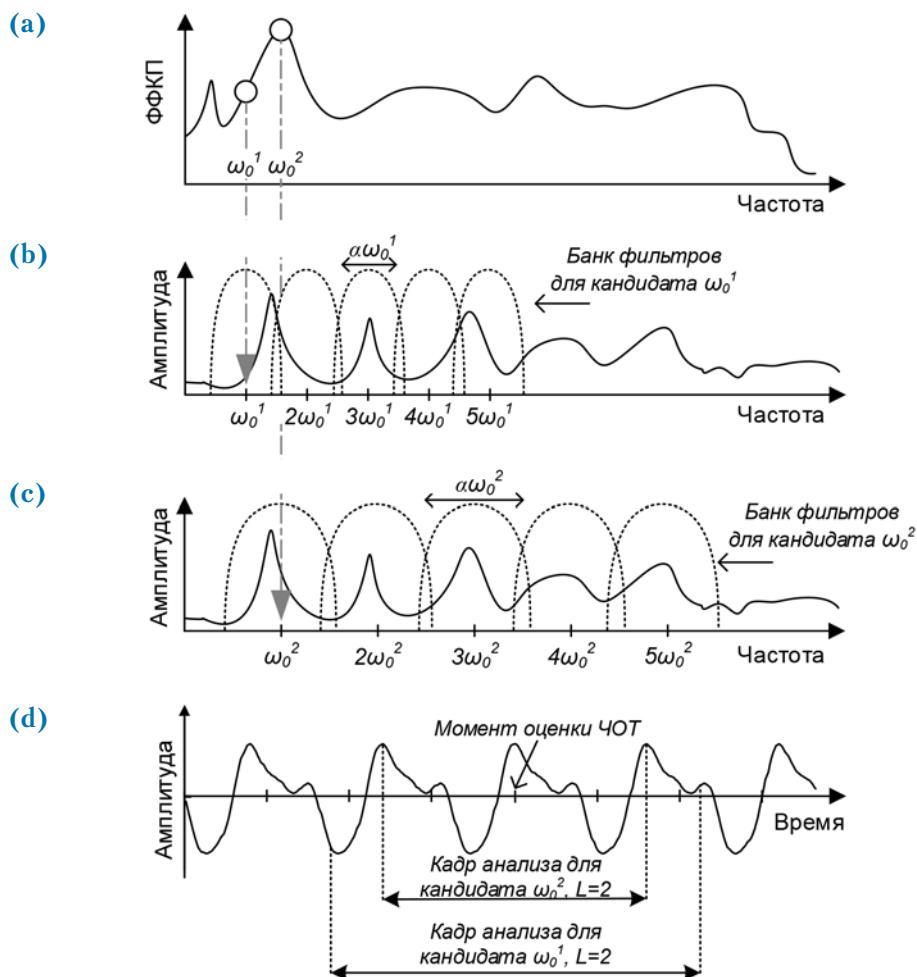
Мгновенные параметры  $A_k(n)$ ,  $\phi_k(n)$  используются в качестве начальных данных для вычисления функции формирования кандидатов периода основного тона. Учитывая предположение, что вариация частоты основного тона пропорциональна её текущему значению, параметры анализирующего банка фильтров должны быть масштабированы для каждого кандидата периода следующим образом:

$$w_{step}(w_0) = w_0, w_{bw}(w_0) = \alpha w_0, \quad (6)$$

где  $w_0$  — частота кандидата в рад/отсчет и  $\alpha$  — допустимая относительная вариация тона. Длительность кадра анализа ( $N$ ) должна быть подобрана, чтобы включать целое число периодов основного тона:

$$N = 2\pi L / w_0, \quad (7)$$

где  $L$  — число периодов в кадре анализа. Данная идея иллюстрируется на рисунке 1.



*Рис. 1. Масштабирование анализирующего банка фильтров для каждого: (а) – функция формирования кандидата периода (ФФКП), (б) – амплитудный спектр и банк фильтров для кандидата  $\omega_0^1$ , (с) – амплитудный спектр и банк фильтров для кандидата  $\omega_0^2$ , (д) – исходный сигнал*

Изменение параметров банка фильтров для каждого кандидата периода в общем случае является вычислительно затратным процессом. Альтернативой данному подходу может служить применение банка фильтров с фиксированными параметрами, но к сигналу с изменяющейся частотой дискретизации. В этом случае можно определить частоту  $F_s$  дискретизации, кратной частоте кандидата периода ( $f_0$ ):

$$F_s = Rf_0, \quad (8)$$

где  $f_0$  – частота в Гц и  $R$  – целое число. Используя (8) и учитывая, что  $f_0 = w_0 F_s / (2\pi)$ , выражение (7) принимает вид фиксированного по длительности кадра анализа для всех кандидатов на период:

$$N = RL. \quad (9)$$



Параметр  $R$  определяет число гармоник, которые оставляются в передискретизированной версии сигнала:

$$K = \begin{cases} \frac{R-1}{2}, & \text{для нечетных } R \\ \frac{R}{2}-1, & \text{для четных } R. \end{cases} \quad (10)$$

Поскольку практическую значимость для определения частоты основного тона имеют лишь несколько первых гармоник, то для анализа можно использовать очень короткие по длительности кадры. Учитывая [8], выражение (6), описывающее масштабирование параметров банка фильтров, принимает вид:

$$\omega_{step} = 2\pi/R, \omega_{bw} = \alpha\omega_{step}. \quad (11)$$

Параметры синусоидальной модели, необходимые для вычисления ФФКП, извлекаются из сигнала при помощи многоскоростной схемы, показанной на рисунке 2.

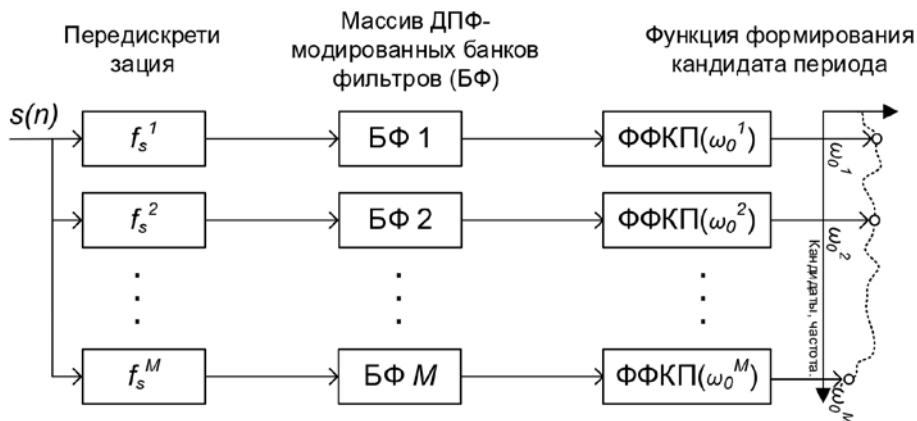


Рис. 2. Многоскоростная схема вычисления функции формирования кандидата периода основного тона ( $M$  – количество кандидатов периода)

## 2. ФУНКЦИЯ ФОРМИРОВАНИЯ КАНДИДАТА ПЕРИОДА ОСНОВНОГО ТОНА

В качестве функции формирования кандидата периода, как правило, используются различные метрики, основанные на автокорреляционной функции. Например, в [12] используется нормированная кросскорреляционная функция

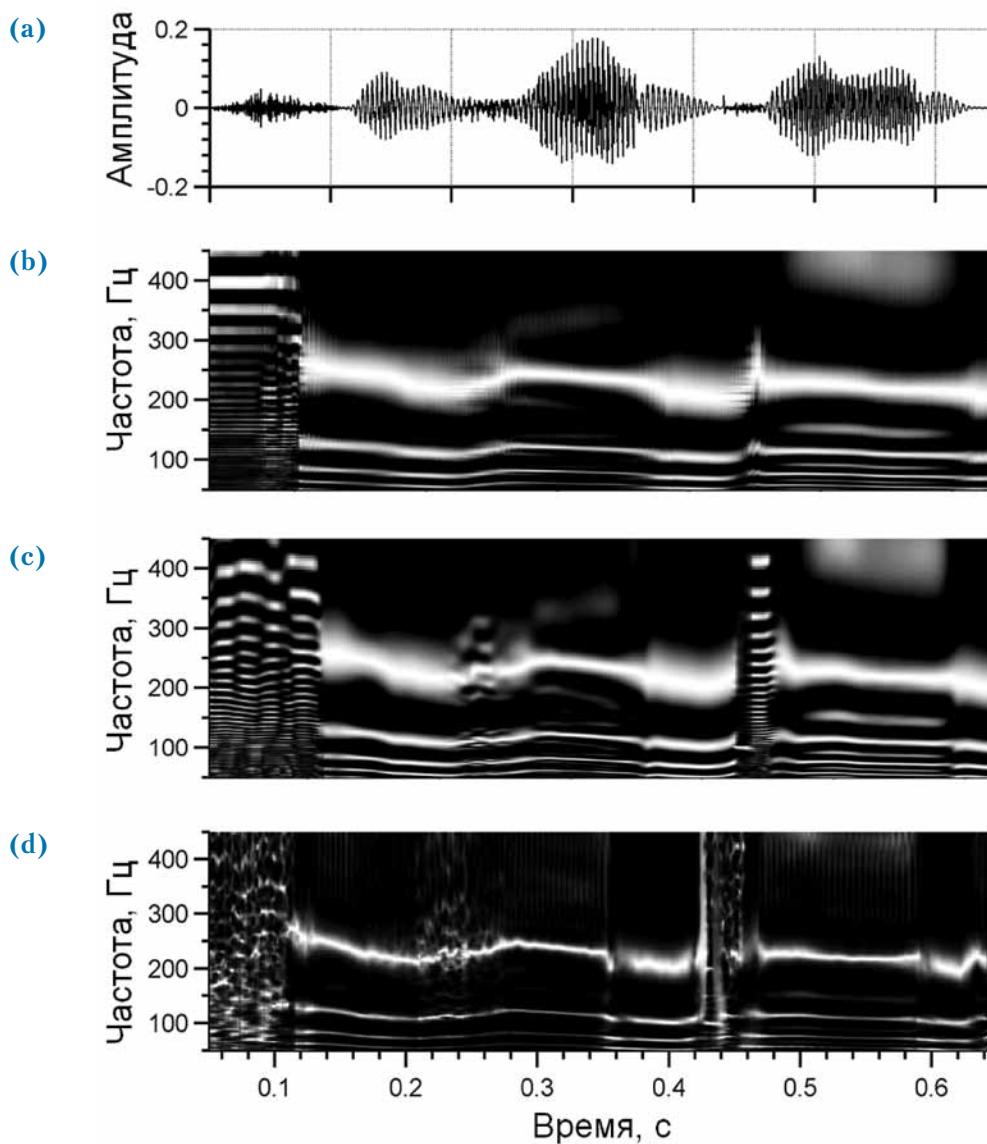
$$\phi(n, l) = \frac{\sum_{i=0}^{N+l-1} s(n)s(n+l)}{\sqrt{e(0)e(l)}}, \quad (12)$$

где  $l$  – задержка в отсчетах,  $e(l) = \sum_{i=0}^{N+l-1} s(n)^2$ .

Функция (12) усредняет данные внутри кадра анализа и поэтому дает сглаженные значения. Чтобы улучшить временное разрешение, в [7] предложено использовать нормированную кросскорреляционную функцию на основе синусоидальной модели сигнала:

$$\phi_{inst}(n, l) = \frac{\sum_{k=1}^K [A_k(n)]^2 \cos(\omega_k(n)l)}{\sum_{k=1}^K [A_k(n)]^2}. \quad (13)$$

В данной функции предполагается, что ширина полос фильтров анализа уже, чем минимально допустимое значение частоты основного тона, вследствие чего каждая гармоника сигнала всегда попадает в отдельный канал.



*Рис. 3.* Формирование кандидатов периода: (а) – исходный сигнал, (б) – ФФКП  $\phi()$ , (в) – ФФКП, полученная на основе синусоидальной модели [7]  $\phi_{inst}()$ , (г) – предлагаемая ФФКП  $\phi_{ms}()$  ( $V=1$ )



Многоскоростная схема анализа, описанная выше, подвержена явлению смешивания гармоник в каналах, которое возникает для высокочастотных кандидатов периода основного тона, когда обрабатываемый голос является низким. Вследствие этого появляются редкие единичные выбросы в высокочастотной области  $\phi_{inst}(n, l)$ . Для того, чтобы уменьшить влияние эффекта смешивания гармоник, предлагается использовать следующую функцию формирования кандидата периода, которая использует мгновенные параметры, полученные для  $2V + 1$  смежных отсчетов:

$$\phi_{ms}(n, l) = \prod_{v=-V}^V \sum_{k=1}^K A_k(n+v) \cos(\omega_k(n+v)l). \quad (14)$$

В каждом отдельном канале схемы на рисунке 2 выполняется вычисление (14), отвечающее за конкретное значение задержки  $l = 1/w_0^m$ , где индекс  $m = 1, \dots M$ , для чего в канал подается кадр сигнала, передискретизированный с коэффициентом  $R/l$  (в соответствии с [8]). Уравновешивание значений  $\phi_{ms}(n, l)$  для различных кандидатов выполняется путем нормализации к единичной энергии каждого передискретизированного кадра сигнала. Использование в (14) амплитуд, не возвещенных в квадрат, в отличие от (13) позволяет сделать вклад амплитуд различных гармоник более сбалансированным. Как правило, эффект смешивания гармоник возникает на коротких временных периодах и может быть существенно уменьшен умножением нескольких термов вида  $\sum_{k=1}^K A_k(n) \cos(w_k l)$ . На рисунке 3 показаны кандидаты периодов, сформированные функциями  $\phi(n, l)$ ,  $\phi_{inst}(n, l)$  и предлагаемой функцией для короткого речевого фрагмента. Очевидно, что функция  $\phi_{ms}(n, l)$  имеет более высокое частотное и временное разрешение по сравнению с  $\phi(n, l)$  и  $\phi_{inst}(n, l)$ .

### 3. АЛГОРИТМ ОЦЕНКИ ЧАСТОТЫ ОСНОВНОГО ТОНА

Предлагаемый алгоритм оценки частоты основного тона состоит из следующих шагов<sup>1</sup>:

- 1) выполнить передискретизацию фрейма входного сигнала  $s(n)$  для каждого кандидата периода основного тона с частотой дискретизации [8];
- 2) нормировать энергию каждого передискретизированного фрейма к 1;
- 3) оценить мгновенные параметры синусоидальной модели согласно выражениям (2)–(5). Данный шаг повторяется для  $2V + 1$  перекрывающихся кадров каждого фрейма. В качестве оконной функции в (2) использовать окно Хемминга;
- 4) вычислить функцию формирования кандидатов периода частоты основного тона (14), используя соответствующий набор параметров;
- 5) умножить полученное значение функции формирования кандидатов периода основного тона на взвешивающую функцию для ограничения низкочастотных кандидатов периода:  $w_{weight}(\omega_0) = 0,2 \frac{\omega_0}{\pi} + 0,8$ ;

<sup>1</sup> Условное название алгоритма «halcyon». Matlab-реализация алгоритма доступна по ссылке <http://dsp.tut.su/halcyon.html>

- 6) поиск наилучшего непрерывного контура частоты основного тона методом динамического программирования, максимизирующего сумму ФФКП на локальной последовательности кадров; в результате данного шага выбирается лучший кандидат  $\omega_{0,best}(n)$ , который является грубой (начальной) оценкой частоты основного тона;
- 7) вычислить уточненную оценку основного тона  $\omega_{0,fine}(n)$ , используя мгновенные параметры синусоидальной модели полученные для лучшего кандидата:

$$\omega_{0,fine}(n) = \frac{1}{\sum_{k=1}^K A_k(n)} \sum_{k=1}^K \frac{1}{k} \omega_k(n) A_k(n). \quad (15)$$

Вычислительная сложность алгоритма (число умножений, требуемое для оценки одного значения частоты основного тона) является невысокой, поскольку в основе реализации банка фильтров лежит быстрое преобразование Фурье. Ориентировочно вычислительная сложность оценивается как

$$O(IKN + 2(V+1)KN\log(N)),$$

где  $I$  — порядок НЧ фильтра, используемого при децимации/интерполяции в процессе передискретизации.

Для практической реализации алгоритма были использованы следующие значения параметров:  $K = 8$ ,  $L = 4$ ,  $R = 2K + 1 = 17$ ,  $N = 68$ ,  $M = 100$ ,  $I = 121$ ,  $V = 1$ . Диапазон поиска частоты основного тона — 50–450 Гц. Данный диапазон разбит линейно в логарифмическом масштабе на 100 интервалов, каждый из которых соответствует одному кандидату периода частоты основного тона. Длительность перидескритеризованных кадров варьируется от 80 (самый низкочастотный кандидат) до 9 мс (самый высокочастотный кандидат).

#### **4. ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА ТОЧНОСТИ АЛГОРИТМА**

Предлагаемый оценщик частоты основного тона (Halcyon) сравнивался с пятью известными и широко применяемыми алгоритмами RAPT [12], YIN [13], SWIPE' [14], IRAPT [7], PEFACT [15]. Сравнение выполнялось в терминах: 1) процента грубых ошибок (gross pitch error — GPE) и 2) среднего значения мелких ошибок (mean fine pitch error — MFPE). GPE вычисляется как процент вокализованных фреймов с ошибкой в оценке основного тона, превышающей от настоящего значения основного тона, при вычислении MFPE фреймы, содержащие грубые ошибки, не учитывались.

Для оценки временного разрешения алгоритма и его устойчивости к быстрому изменению основного тона были синтезированы модельные сигналы с изменяющейся частотой основного тона в диапазоне от 100 до 350 Гц. Полученные экспериментальные результаты были разделены на 6 групп в зависимости от скорости изменения частоты основного тона, измеряемой в процентах изменения тона на миллисекунду (0–0,3, 0,3–0,6, 0,6–0,9, 0,9–1,2, 1,2–1,5, >1,5). Усреднённые значения ошибок показаны на рисунке 4.

Алгоритмы IRAPT и Halcyon показывают более высокую устойчивость к изменению частоты основного тона — процент грубых ошибок для них остается незначительным до значения 1,5 %/ мс. График MFPE показывает, что предлагаемый алгоритм превосходит все остальные по частотно/временному разрешению. На рисунке 5 приведен пример анализа модельного сигнала с быстрым изменением тона. Алгорит-

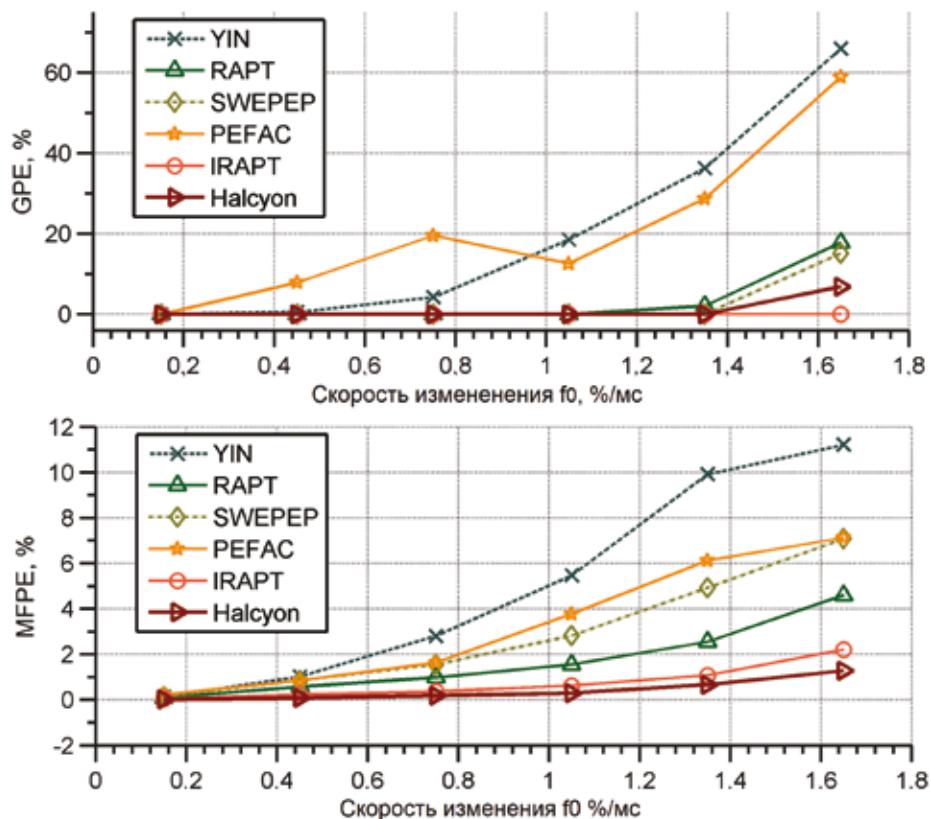


Рис. 4. Оценка временного разрешения алгоритмов выделения основного тона

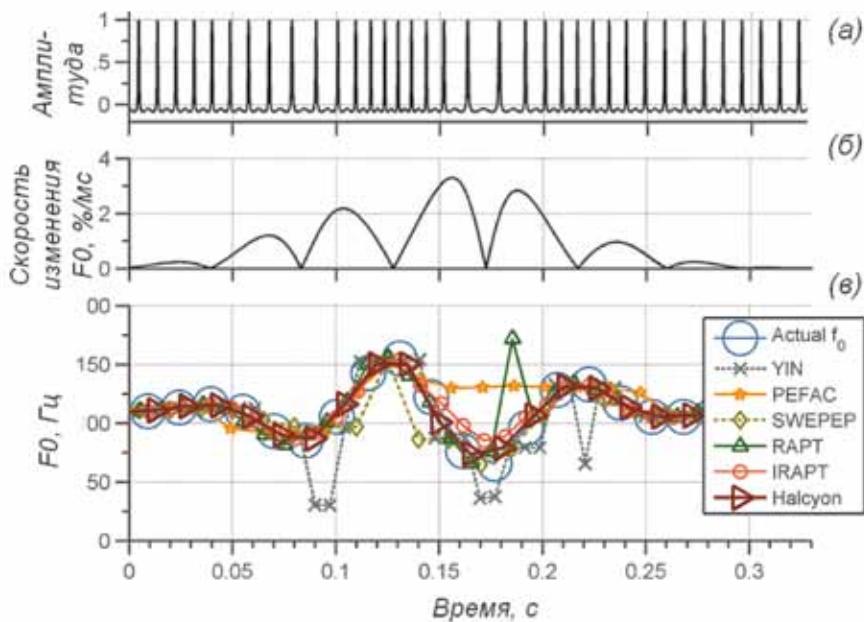


Рис. 5. Анализ сигнала с изменяемым тоном. (а) – исходный сигнал, (б) – скорость изменения основного тона, (с) – настоящий и рассчитанный контур частоты основного тона

мы IRAPT, SWIPE' и Halcyon дают оценку, весьма близкую к истинным значениям, остальные алгоритмы не демонстрируют такой точности.

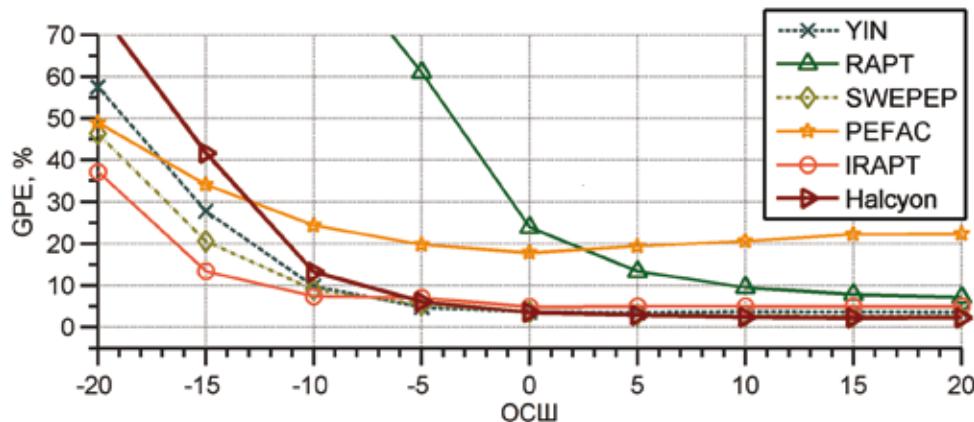
Для экспериментов с использованием натуральной речи использовалась база PTDB-TUG [16]. Усредненные результаты экспериментов для чистой речи приведены в таблице 1. По сравнению с IRAPT 1 предлагаемый алгоритм демонстрирует в два раза меньший процент грубых ошибок (GPE) благодаря многоскоростной схеме анализа.

**Таблица 1**  
**Сравнение алгоритмов оценки частоты основного тона  
с использованием речевых сигналов**

	Мужской голос		Женский голос	
	GPE%	MFPE%	GPE%	MFPE%
RAPT	3.687	1.737	6.068	1.184
YIN	3.184	1.389	3.960	0.835
SWIPE'	0.756	1.505	4.273	0.800
PEFAC	20.521	1.383	31.192	0.972
IRAPT 1	1.625	1.608	3.777	0.977
Halcyon	0.743	1.268	3.600	1.039

Для проверки на устойчивость к шумам к чистым речевым сигналам добавлялся шум двух видов (белый и речеподобный (англ. *babble*)) с различным значением ОСШ от -20 до 20 дБ. Усредненные результаты оценки тона в зашумленной речи показаны на рисунках 6,7.

Для белого шума все алгоритмы за исключением RAPT показывают хороший результат для ОСШ-10 дБ и выше. Для речеподобного шума результаты работы алгоритмов быстро ухудшаются, начиная со значения ОСШ 0 дБ. В целом предлагаемый алгоритм показывает приемлемую устойчивость к аддитивным шумам, учитывая, что в нем используются значения параметров синусоидальной модели, оцененные на очень коротких фреймах анализа (до 9 мс) для высокочастотных кандидатов частоты основного тона.



**Рис. 6.** Точность измерения основного тона (аддитивный белый шум)

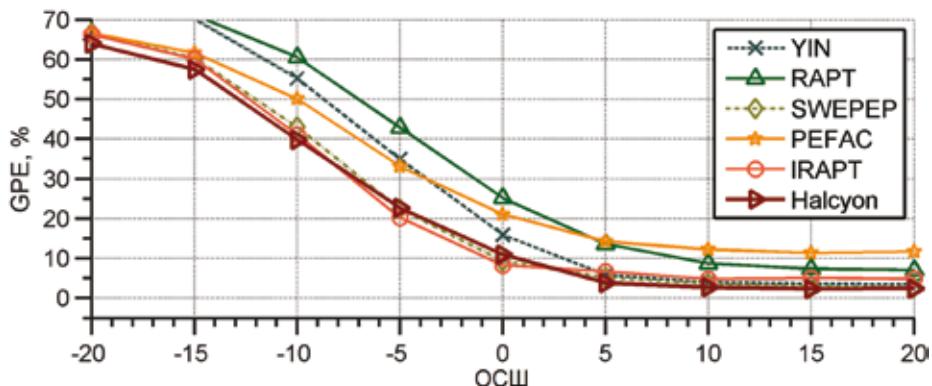


Рис. 7. Точность измерения основного тона (аддитивный речеподобный шум)

## ЗАКЛЮЧЕНИЕ

В работе представлен алгоритм выделения частоты основного тона, который может быть применен в задачах анализа искусственно сгенерированных и речевых сигналов. В алгоритме выполняется декомпозиция сигнала на узкополосные составляющие, каждая из которых описывается синусоидальной моделью с мгновенными параметрами амплитуды, частоты и фазы. Для каждого кандидата периода частоты основного тона выполняется масштабирование анализирующего банка фильтров, что способствует точной оценке тона как для низких, так и для высоких голосов. Экспериментальные результаты показывают значительную устойчивость предлагаемого алгоритма к модуляциям основного тона, а также его высокое частотное и временное разрешение.

## Благодарность

Работа выполнялась при поддержке компании ITForYou (Москва), а также Белорусского республиканского фонда фундаментальных исследований (грант № Ф17У-003).

## ЛИТЕРАТУРА

1. F. Zhang, G. Bi, Y. Q. Chen, "Harmonic transform," in Vision, Image and Signal Processing, IEE Proceedings, vol.151, no.4, pp.257–263, 2004.
2. R. J. McAulay, T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 34, no. 4, pp. 744–754, 1986.
3. J. Laroche, Y. Stylianou, E. Moulines, "HNS: Speech modification based on a harmonic+noise model," in ICASSP-93 — IEEE International Conference on Acoustic, Speech, and Signal Processing, April 27-30, Minneapolis, USA, Proceedings, 1993. — pp. 550–553.
4. J. O. Hong, P. J. Wolfe, "Model-based estimation of instantaneous pitch in noisy speech," in INTERSPEECH 2009 — 10th Annual Conference of the International

- Speech Communication Association, September 6–10, Brighton, UK, Proceedings, 2009. — Pp. 112–115.
- 5. *B. Resch, M. Nilsson, A. Ekman and W. B. Kleijn* “Estimation of the Instantaneous Pitch of Speech,” IEEE Transactions on Audio, Speech and Language Processing, vol. 15, No. 15, Pp. 819–822, 2007.
  - 6. *T. Abe, T. Kobayashi, S. Imai*, “Harmonics tracking and pitch extraction based on instantaneous frequency,” in ICASSP-95 — IEEE International Conference on Acoustic, Speech, and Signal Processing, May 9–12, Detroit, USA, Proceedings, 1995. — Pp. 756–759.
  - 7. *E. Azarov, M. Vashkevich, A. Petrovsky*, “Instantaneous pitch estimation based on RAPT framework,” in EUSIPCO'12 — European Signal Processing Conference, August 27–31, Bucharest, Romania, Proceedings, 2012. — Pp. 2787–2791.
  - 8. *E. Azarov, M. Vashkevich, A. Petrovsky*, “Instantaneous harmonic representation of speech using multicomponent sinusoidal excitation,” in INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, August 25–29, Lyon, France, Proceedings, 2013. — Pp. 1697–1701.
  - 9. *E. Azarov, M. Vashkevich, A. Petrovsky*, “Guslar: A framework for automated singing voice correction,” in ICASSP-2014 — IEEE International Conference on Acoustic, Speech, and Signal Processing, May 4–9, Florence, Italy, Proceedings, 2014. — Pp. 7919–7923.
  - 10. K. Hotta, K. Funaki, “On a Robust F0 Estimation of Speech based on IRAPT using Robust TV-CAR Analysis,” in APSIPA 2014 — Annual Summit and Conference Asia-Pacific Signal and Information Processing Association, 2014, December 9–12, Siem Reap, Cambodia, Proceedings, 2014. — Pp. 1–4.
  - 11. *E. van den Berg, B. Ramabhadran*, “Dictionary-based pitch tracking with dynamic programming,” in INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association, September 14–18, Singapore, Proceedings, 2014. — Pp. 1347–1351.
  - 12. *D. Talkin*, “A Robust Algorithm for Pitch Tracking (RAPT)” in “Speech Coding & Synthesis,” W B Kleijn, K K Paliwal eds, Elsevier ISBN 0444821694, 1995.
  - 13. *A. Cheveigné, H. Kawahara* “YIN, a fundamental frequency estimator for speech and music,” Journal of the Acoustical Society of America, vol. 111, No. 4, Pp. 1917–1930, 2002.
  - 14. *A. Camacho, J. G. Harris*, “A sawtooth waveform inspired pitch estimator for speech and music,” Journal of the Acoustical Society of America, Vol. 123, No. 4, Pp. 1638–1652, 2008.
  - 15. *S. Gonzalez, M. Brookes*, “PEFAC — A Pitch Estimation Algorithm Robust to High Levels of Noise,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 22, No. 2, Pp. 518–530, 2014.
  - 16. *G. Pirker, M. Wohlmayr, S. Petrik, F. Pernkopf* “A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario,” in INTERSPEECH 2011 — 12th Annual Conference of the International Speech Communication Association, August 28–31, Lyon, France, Proceedings, 2011. — Pp. 1509–1512.

## **ESTIMATION OF INSTANTANEOUS FUNDAMENTAL FREQUENCY OF SPEECH BASED ON MULTIRATE SIGNAL PROCESSING**

***Maksim I. Vashkevich,***

*Candidate of technical Sciences, associate Professor of the Belarusian state University of Informatics and Radioelectronics (BSUIR)*

***Iliy S. Azarov,***

*Doctor of technical Sciences, associate Professor, BSUIR*

***Aleksandr A. Petrovsky,***

*Doctor of technical Sciences, Professor of the chair of electronic computing BSUIR*



### **Abstract**

The paper presents an algorithm for accurate pitch estimation that takes advantage of the sinusoidal model with instantaneous parameters. The algorithm decomposes the signal into subband components, extracts their instantaneous parameters and evaluates period candidate generating function (PCGF). In order to achieve high accuracy for low and high-pitched sounds it is assumed that possible pitch variation range is proportional to current pitch value. The bandwidths of the decomposition filters and length of the analysis frame are scaled for each period candidate by multirate sampling. The algorithm is compared to other widely used pitch extractors on artificial quasiperiodic signals and natural speech. The proposed algorithm shows a remarkable frequency and time resolution for pitch-modulated sounds and performs well both in clean and noisy conditions.

**Keywords:** fundamental frequency, pitch, multirate signal processing

# Помехоустойчивость слоговых таблиц при восприятии речи в шуме

**Станислав Антонович Крейчи,**  
научный сотрудник филологического факультета МГУ  
им. М.В. Ломоносова

**Ольга Фёдоровна Кривнова,**  
доктор филологических наук, ведущий научный сотрудник филологического  
факультета МГУ им. М.В. Ломоносова

**Екатерина Александровна Тихонова,**  
бакалавр филологического факультета МГУ им. М.В. Ломоносова

## Аннотация

В работе описана методика составления тестовых таблиц с разной степенью помехоустойчивости для исследования восприятия речи в шуме. Основная цель данного исследования заключалась в том, чтобы построить и протестировать слоговые артикуляционные таблицы, в которых учитывалась бы разная степень помехоустойчивости элементов к шуму, а также выявить наиболее устойчивые к шуму акустико-фонологические признаки. В результате аудитивного тестирования было установлено, что слоговая разборчивость составленных таблиц примерно одинакова в отсутствии шумовой помехи и значительно отличается при белом и розовом шуме.

**Ключевые слова:** тестовые таблицы, аудитивное тестирование, помехоустойчивость, белый, розовый шум, слоговая разборчивость, акустико-фонологические признаки.

## ВВЕДЕНИЕ

Исследования слухового восприятия речевых сообщений в шуме не перестают быть актуальными при решении многих задач в области разработки новых средств связи, а также с общефонетической точки зрения, так как помогают расширить наши представления о восприятии речи и детализировать существующие модели слуховой обработки речевого сигнала человеком [1].

Общие положения этих моделей сводятся к тому, что минимальные звуковые единицы языка (фонемы) хранятся в памяти носителя языка в виде целевых артикуляций, которые необходимы для их произнесения при порождении речи. Механизм восприятия речи обрабатывает только те звуки/фрагменты речевого сигнала, которые могут быть соотнесены с целевыми артикуляциями и их различительными признаками. Предполагается, что неречевые сигналы обрабатываются отдельно, другими механизмами. Независимость работы этих механизмов позволяет отделять речь от помех в зашумленных условиях восприятия.



Осмысление речевого сигнала строится на идентификации входящих в него слов и установлении синтаксических и семантических связей между ними [2]. Распознаванию слова в сообщении предшествует распознавание его фонемного состава, полного или частичного. Однако распознавание отдельных фонем не осознаётся слушающим, в связи с чем возникает вопрос о том, на каком этапе восприятия, каким образом и в какой мере оно влияет на восприятие и распознавание слова. Одна из популярных гипотез состоит в том, что в памяти носителя языка каждой фонеме, наряду с целевой артикуляцией, соответствует свой слуховой образ, представляющий собой определенную конфигурацию полезных акустических признаков-параметров, на формирование которой в речевом сигнале и направлена целевая артикуляция фонемы. В рамках данного подхода слуховой образ фонемы (или, иначе, ее акустическая модель) рассматривается либо как результат усреднения по всем акустическим признакам всех когда-либо слышанных вариантов данной фонемы, либо как её вариант, наиболее часто встречающийся в речи. При этом предполагается, что при восприятии речи фонема опознается не как целостная единица, а через распознавание её фонологически значимых признаков. Если для определения точного значения признака недостаточно акустической информации, то определяется не отдельная фонема, а класс фонем, различающихся по этому неопознанному признаку. Данный принцип является основополагающим для большинства моделей восприятия речи. Они имеют между собой некоторые различия, касающиеся, например, набора полезных акустических признаков, имеющих фонологическую нагрузку, тем не менее их можно объединить под общим названием «модели, использующие дифференциальные признаки» или сокращенно «признаковые модели».

В признаковых моделях активно используется также понятие акустических ключей. Акустические ключи, или дескрипторы — это «сведения о наличии/отсутствии в речевом сигнале релевантных для каждого фонологического признака акустических событий, об отношениях следования между ними, а также сведения об их временных, частотных и энергетических модификациях» [3]. Различные акустические ключи в разной степени подвергаются искажению под воздействием помех [4], что и определяет помехоустойчивость соответствующих фонологических признаков и ассоциированных с ними речевых единиц. Кроме акустических факторов на восприятие речи могут также влиять и другие факторы (статистические, общелингвистические, контекстные, стилистические и пр.).

## **ПОМЕХОУСТОЙЧИВОСТЬ СЛОГОВЫХ ТАБЛИЦ ПРИ ВОСПРИЯТИИ РЕЧИ В ШУМЕ**

В настоящей работе экспериментально исследовалось восприятие речи в условиях внешнего шума. Целью исследования было выявление наиболее устойчивых к шуму акусто-фонологических признаков, а также анализ влияния шумовой помехи на стратегию восприятия речевого сигнала [5].

Основная задача исследования состояла в том, чтобы сравнить влияние акустических и частотных факторов на качество распознавания речевого сигнала в нормальных условиях и в шуме в условиях минимального использования неакустических признаков при идентификации речевых единиц. Для этого в качестве материала для исследования были выбраны бессмысленные единицы речи (отдельные слоги). В составленные нами артикуляционные таблицы были включены слоги структуры согласный-гласный (СГ). Фонемы в слоге структуры СГ характеризуются наибольшей произносительной слитностью по сравнению с другими звукосочетаниями. Это означает, что в сочетаниях СГ сильнее всего проявляется контекстное влияние целевых артикуляций фонем друг на друга, что приводит к изменению звучания как гласного, так и согласного по сравнению с их возможным изолированным произнесением и другими контекстами. С учётом этого целесообразно в качестве акустической характеристики слога СГ рассматривать не только собственные признаки входящих в него фонем, согласного и гласного, а признак слога как такового, который заключается в степени фонетического сходства/различия входящих в него фонем, т.е. в степени слогового контраста [6].

Согласно Л.В. Бондарко [6], в интегральном признаком «слоговой контраст» можно выделить следующие составляющие:

1. **Контраст по частоте основного тона (ЧОТ):** реализуется в наибольшей степени в сочетаниях глухих согласных с гласным (т.к. на участке согласного тон отсутствует, а на участке гласного присутствует), а в сочетаниях со звонкими согласными заключается в изменении ЧОТ при переходе от согласного к гласному.
2. **Контраст по длительности:** реализуется максимально в слогах с твёрдыми глухими смычными (взрывными) согласными, длительность которых намного меньше длительности гласного.
3. **Контраст по формантной структуре:** реализуется во всех слогах, кроме слогов с начальными сонорными согласными, имеющими, как и гласные, чётко выраженную формантную структуру.
4. **Контраст по интенсивности:** в ударных слогах гласный характеризуется большей интенсивностью, чем согласный, хотя в слогах с сонорными согласными этот контраст ослабляется.
5. **Контраст по частоте F2:** количественно характеризуется величиной перепада частоты второй форманты (F2) на начальном переходном участке гласного (таблица 1).

При восприятии речи в шуме наблюдается так называемый «эффект маскировки» — изменение слуховой чувствительности к маскируемому сигналу под воздействием маскирующего [7]. В нашем эксперименте мы использовали следующие виды шума:

- белый шум — стационарный шум, спектральные составляющие которого равномерно распределены по всему диапазону частот;
- розовый шум — стационарный шум, равномерно убывающий по логарифмической шкале частот.



Таблица 1

Значения контраста по F2 в слогах СГ (С — любой согласный)

Величина перепада между значением F2 в начале и в конце переходного участка, Гц с учетом знака	Слог
- 400	па, ба, ма, фа, ва, ла
0 ... – 100	ка, га, ха
100	та, да, на, са, за, ца, ра, ша
500	все слоги С'а
-100	по, бо, мо, фо, во, ло
400	то, до, но, со, зо, до, ро, то
1000	п'о, б'о, м'о, ф'о, в'о
1100	т'о, д'о, н'о, с'о, з'о, л'о, р'о, к'о, г'о, х'о
0... – 100	ко, го, хо
0	бу, му, фу, ву, ку, гу, ху, лу
500	ту, ду, су, зу, ну, шу, жу
1100	п'у, б'у, м'у, ф'у, в'у
1200	т'у, д'у, с'у, з'у, л'у, р'у, г'у, н'у
-500	Ле
-400	пе, бе, ме, фе, ве
0	те, де, се, зе, не, ше, же, ре
100	С'е
-900	пы, бы, мы, фы, вы, лы, ты, ды, сы, зы, ны, ры
0	С'и

Эффект маскировки слоговых контрастов в этих шумах должен проявляться, по-видимому, следующим образом:

- Маскировка наиболее выражена, когда частоты маскируемого и маскирующего звука близки. Таким образом, можно ожидать, что эффект маскировки белым шумом не будет зависеть от частотных характеристик маскируемого звука, в то время как эффект маскировки розовым шумом будет сильнее для звуков с низкими частотами;
- степень маскировки увеличивается с увеличением интенсивности маскирующего звука. В наших экспериментальных условиях, когда интенсивность шума будет постоянной, а интенсивность участков речевого сигнала разной (в зависимости от собственных акустических свойств произносимой единицы), этот принцип можно переформулировать так: «степень маскировки увеличивается с уменьшением интенсивности маскируемого звука относительно шума»;
- при условии большой интенсивности маскера, маскировка заметно более выражена по отношению к звукам высокой частоты. Речевой сигнал в принципе имеет относительно небольшой диапазон частот,

однако можно предположить, что те акустические ключи, которые расположены в более низкой частотной области, будут подвергаться маскировке в меньшей степени, чем те, которые располагаются в более высокой частотной области.

При составлении тестовых слоговых таблиц мы руководствовались следующими соображениями:

- Речевые единицы, характеризующиеся большей интенсивностью, можно считать более помехоустойчивыми.
- Такие акустические ключи, как наличие/отсутствие основного тона и наличие/отсутствие формантной структуры, можно считать более помехоустойчивыми, так как они находятся в области более низких частот, и, следовательно, в меньшей степени подвергаются маскировке.
- F-картину гласного можно считать помехоустойчивым («надёжным») акустическим ключом для распознавания слога.
- Конфигурационные характеристики формант (формантные треки) также можно считать надёжным акустическим ключом для распознавания слога.
- Шумные согласные, у которых наибольшая и достаточно выраженная интенсивность шума приходится на область низких частот, можно считать более помехоустойчивыми, чем согласные, у которых при прочих равных условиях наибольшая интенсивность шума приходится на область высоких частот.

С учетом изложенных соображений для проведения перцептивного эксперимента были составлены три тестовые таблицы, одна из которых содержала наиболее помехоустойчивые слоги (т. 1), вторая — средние по помехоустойчивости (т. 2) и третья — наименее помехоустойчивые (т. 3). При распределении слогов по таблицам мы руководствовались следующими критериями:

### **1) Шумность-сонорность согласного**

Сонорные согласные обладают ярко выраженной формантной структурой и в целом являются более надёжными акустическими ключами для восприятия слога, чем шумные согласные. К сонорным по данному признаку примыкают согласные [в] и [й], которые также обладают выраженной формантной структурой.

### **2) Интенсивность шумового компонента шумного согласного**

Глухие и звонкие шумные согласные отличаются друг от друга наличием основного тона, который маскируется шумом в меньшей степени. Однако для того, чтобы различать глухие согласные между собой, так же как и звонкие, необходимо располагать и акустической информацией о шумовом компоненте согласного. Степень маскировки шумового компонента согласного зависит как от области его частотной локализации в спектре, так и от его интенсивности и длительности. Слоги с более интенсивными шумными согласными можно считать наиболее помехоустойчивыми по этому признаку. К ним относятся прежде всего слоги, содержащие сибилянты: [с], [с'], [з], [з'], [ш], [ш'], [ц], [ч']. Наименьшей интенсивностью шумных участков и, следовательно, наименьшей помехоустойчивостью должны обладать в русском языке согласные [х], [х'], [ф], [ф'], [т], [п], [п'], [к], [к'].



### **3) Интенсивность гласного**

Участки гласных содержат большое количество акустических ключей, важных для распознавания слогов. По критерию интенсивности к наиболее помехоустойчивым будут скорее всего относиться слоги с гласным [а], который является наиболее интенсивным из гласных. К наименее помехоустойчивым будут относится слоги с гласными компонентами [и] и [у], наименьшими по интенсивности.

### **4) Слоговой контраст по F2**

F-картина гласного является надёжным акустическим ключом для распознавания речи в шуме. Слоги, в которых перепад F2 максимально информативен, должны быть наиболее помехоустойчивыми по данному признаку. Однако надо учитывать не только степень выраженности данного контраста, то есть величину перепада между значением F2 на начальном и стационарном участке, но и информативность F-картины. Под информативностью мы понимаем количество контекстов, для которых характерен тот или иной переходный участок: чем меньше это количество, тем более информативной считается F-картина.

Например, F-картина гласного одинакова для всех слогов типа С'и, а также для всех слогов типа С'е и всех слогов типа С'a (см. табл. 1). В подобных случаях мы будем считать её малоинформационной.

Таким образом, слоги, в которых перепад F2 максимально информативен, мы будем считать наиболее помехоустойчивыми по данному признаку.

### **5) Частота встречаемости слога в речи**

В условиях недостаточности акустической информации для однозначной идентификации сообщения, при отсутствии контекстной информации, слушающий принимает решение в пользу наиболее частотного варианта из возможных. В связи с этим более помехоустойчивыми являются более частотные слоги русского языка, а менее помехоустойчивыми — менее частотные.

Ниже приведена сводная таблица помехоустойчивых и помехонеустойчивых слогов по каждому из описанных критериев в отдельности.

### ***Методика подготовки артикуляционных (слоговых) таблиц с тремя типами помехоустойчивости***

На основе таблицы 2.1 была проведена дальнейшая работа по подготовке слоговых таблиц трех типов помехоустойчивости, которые служили материалом для перцептивного эксперимента.

Интегральная помехоустойчивость таблицы определяется помехоустойчивостью входящих в неё элементов. Показателем помехоустойчивости

**Таблица 2.1**

**Помехоустойчивость разных типов слогов**

Критерии	Устойчивые	Неустойчивые
Шумность-сонорность согласного	Слоги с согласными [м], [н], [р], [л], [й], [в]	Слоги с шумными согласными (кроме [в])
Интенсивность шумового компонента шумного согласного	Слоги с согласными [ш], [с], [ч], [ц] и их мягкими аллофонами	Слоги с согласными [п], [б], [ф], [х] и их мягкими аллофонами
Интенсивность гласного	Слоги с гласными [а], [е], [о]	Слоги с гласными [ы], [и], [ы]
Слоговой контраст по частоте F2	па, ба, ма, фа, ва, ла ка, га, ха по, бо, мо, фо, во, ло п'о, б'о, м'о, ф'о, в'о ко, го, хо п'у, б'у, м'у, ф'у, в'у ле пе, бе, ме, фе, ве	С'и С'е С'a
Частотность	Наиболее частотные	Наименее частотные

элемента таблицы является относительно высокий уровень его разборчивости в шуме, и наоборот, элементы, имеющие низкий уровень разборчивости в шуме, мы называем помехонеустойчивыми.

Элементами таблиц являются слоги русского языка структуры СГ. Всего в данной работе было составлено три тестовые таблицы, одна из которых содержала наиболее помехоустойчивые слоги (1), вторая — средние по помехоустойчивости (2) и третья — наименее помехоустойчивые (3).

При подготовке таблиц все слоги были условно разделены по каждому критерию, указанному в табл. 2.1 на помехоустойчивые (+) и помехонеустойчивые (-). Для критериев «интенсивность шумового компонента согласного», «контраст по F2» и «частотность» была также введена оценка «±», так как по этим критериям выделяются единицы, которые трудно однозначно отнести к той или иной группе. По совокупности оценок, определенных по каждому отдельному критерию, для каждого слога была определена та артикуляционная таблица (из трех), к которой он был отнесен в соответствии с суммарной оценкой прогнозируемой помехоустойчивости, см. таблицу 2.2.

При отнесении слога в ту или иную группу мы руководствовались также некоторыми дополнительными соображениями относительно потенциальной значимости отдельного признака для восприятия речи в шуме:

- признак шумности-сонорности согласного имеет наибольший вес, так как его важность для распознавания в шуме была доказана в экспериментальных исследованиях [Phatak et al., 2008; Wang & Bilger, 1973].
- признак слогового контраста по F2 имеет большой вес, так как мы исходим из предположения, что акустические ключи, находящиеся на участках гласных, наиболее устойчивы к шуму.



Таблица 2.2

*Влияние критериев помехоустойчивости на включение слога в артикуляционную таблицу определенной степени помехоустойчивости*

Шумность– сонорность	+	+	+	--	-	-	-	-	-	-	-	-
Интенсивность шумового компонента шумного согласного	-	-	-	+	+	±	±	±	-	-	-	-
Слоговой контраст по F2	+	-	±	±	±	+	±	±	+	-	-	-
Частотность	+	+	+	-	+	±	+	+	-	+	+	±
Интенсивность гласного	+	+	+	+	+	+	+	+	+	+	+	-
Тип (номер) таблицы	1	2	2	1	2	1	2	3	1	2	3	3

- критерий частотности оказывается решающим в тех случаях, когда по остальным (акустическим) признакам слоги относятся к помехонеустойчивым.

В результате все возможные слоги СГ с полноартикуируемым гласным (180 единиц) были размещены в следующие тестовые артикуляционные таблицы (в т. 3.1–3.3 слоги записаны в широкой фонетической транскрипции):

Таблица 3.1

*Слоги с высокой степенью помехоустойчивости  
(предположительно «легкая»)*

ба	Му	на	ч'е	Л'а	па	ш'а	та	ре	й'е
ле	Ва	ло	га	ше	ве	ха	ч'о	лы	ры
ра	л'и	ш'е	ну	ла	ше	са	й'о	це	ч'у
й'у	Мы	ру	ш'о	Ч'и	ву	цы	ро	су	ны
мо	Вы	го	ша	но	ме	лу	хо	ш'и	ма
не	с'и	со	ш'у	шы	й'а	ка	л'о	л'у	ко

Таблица 3.2

*Слоги со средней степенью помехоустойчивости  
(предположительно «средняя»)*

в'е	Жу	м'а	по	де	н'о	р'и	в'а	се	за
д'и	р'о	ке	бу	Р'у	жа	да	н'а	п'и	гы
фу	с'у	то	пу	цо	г'а	шу	ты	цу	р'а
зы	Шо	в'о	сы	Д'е	с'о	жы	ца	б'е	с'е
до	н'у	т'е	ду	фа	м'о	в'и	д'у	н'и	г'е
же	р'е	гу	жо	Й'и	бы	ку	м'е	м'и	ч'а
с'а	б'и	кы	н'е	ту	з'е	т'и	ге	д'о	зу
з'и	к'и	п'е	м'у						

*Таблица 3.3  
Слоги с низкой степенью помехоустойчивости (предположительно «трудная»)*

де	ф'о	пе	х'е	бо	зе	ху	г'и	б'у	з'а
х'о	ф'а	г'о	фо	з'у	г'у	ф'е	б'о	т'у	хы
к'а	Фе	т'а	фы	Д'а	пы	к'у	те	х'а	ды
ф'и	п'о	ф'у	зо	К'е	бе	хе	п'а	з'о	х'и
к'о	п'у	х'у	б'а	т'о	г'е				

### **ИССЛЕДОВАНИЕ ПОМЕХОУСТОЙЧИВОСТИ СОСТАВЛЕННЫХ СЛОГОВЫХ ТАБЛИЦ**

Помехоустойчивость составленных таблиц исследовалась посредством аудитивного тестиирования на разборчивость в нормальных условиях и в шуме. Слоговые элементы каждой таблицы были начитаны диктором (женский голос) — носителем орфоэпической нормы русского языка. Каждый слог в таблице повторялся два раза. Пауза между повторами составляла 2 секунды, а пауза между вторым повтором и следующим слогом — 3 секунды. На оцифрованные записи в звуковом редакторе Adobe Audition CS6 накладывался белый или розовый шум в соотношении +0 дБ. Каждая таблица предъявлялась слушателям в трёх вариантах записи: без наложения шума, с наложением белого шума и с наложением розового шума.

Экспериментальные комплекты аудиозаписей были составлены таким образом, чтобы внутри одного комплекта присутствовали все три таблицы и все три режима. Таким образом, получилось 6 экспериментальных комплектов: (1-nc, 2-wn, 3-pn), (1-nc, 2-pn, 3-wn), (1-pn, 2-nc, 3-wn), (1-pn, 2-wn, 3-nc), (1-wn, 2-pn, 3-nc), (1-wn, 2-nc, 3-pn), где **1, 2, 3** — соответственно лёгкая, средняя и трудная таблицы, а *nc*, *wn*, *pn* — соответственно нормальные условия, белый шум и розовый шум.

Испытуемыми были 15 студентов от 19 до 23 лет с нормальным слухом: 9 женского пола и 6 мужского пола. Родной язык для всех испытуемых — русский, все они являлись носителями стандартного московского произношения.

Каждому испытуемому предлагался для прослушивания один из экспериментальных комплектов. Задача испытуемого состояла в том, чтобы записать в протокол в русской орфографии слоги, которые он слышит на аудиозаписи. В случае затруднения ставился знак «0». Перед прослушиванием экспериментального материала предлагался тренировочный материал из 5 слогов — элементов «средней» таблицы в нормальных условиях. Тренировочный материал можно было прослушивать неограниченное количество раз. Громкость устанавливалась таким образом, чтобы испытуемый чувствовал себя комфортно. Каждая тестовая запись предъявлялась один раз в присутствии экспериментатора.

Всего в результате эксперимента было получено 2700 реакций. Разборчивость каждого слога рассчитывалась как процент правильного его распознавания от общего количества предъявлений слога в данном режиме. Ниже приводится средний интегральный показатель разборчивости по разработанным нами слоговым таблицам.



## Результаты перцептивного тестирования слоговых таблиц

Таблица 4  
Средний интегральный показатель разборчивости по таблицам

	Таблица 1	Таблица 2	Таблица 3
Без шума	0,98	0,97	0,95
Розовый шум	0,90	0,72	0,47
Белый шум	0,89	0,70	0,48

В результате аудитивного тестирования было установлено, что слоговая разборчивость составленных таблиц примерно одинакова в отсутствии шумовой помехи и значительно отличается при белом и розовом шуме. Это говорит о том, что учитываемые в таблицах признаки слога, такие как шумность-сонорность согласного, интенсивность шумового компонента шумного согласного, слоговой контраст по частоте второй форманты, интенсивность гласного и частотность слога, являются значимыми для оценки помехоустойчивости слога. Обнаружено также, что слоги, содержащие сочетания мягких заднеязычных согласных с гласными среднего и заднего ряда, обладают в условиях шумовой помехи разборчивостью, близкой к нулю, а слоги с сонорными согласными в целом оказались наиболее помехоустойчивыми. Среди сонорных следует выделить носовые согласные, которые испытуемые часто путали между собой. Это происходило в тех слогах, где данные согласные были мягкими. Объясняется это тем, что некоторая часть информации о формантной картине на переходном участке гласного всё же теряется в условиях шумовой помехи, и при наличии и-образного перехода на гласном не так заметно на слух понижение формант, вызванное соседством губного согласного, которое и является основным отличием в акустической картине слогов с носовыми [м'] и [н'].

В ходе эксперимента также выяснилось, что в шуме слоги с фрикативными согласными часто опознаются как слоги со смычными или аффрикатами того же места образования. Так, в ответах испытуемых [с] исходного слога часто заменяется на [ц], [ш'] на [ч'], [з'] на [д'] и т.п. Вероятно, относительно долгий шумный участок фрикативного согласного неравномерен по своей интенсивности. Из-за эффекта маскировки менее интенсивные участки фрикативного согласного не прослушиваются, а краткий пик интенсивности воспринимается как короткий по длительности шумовой участок, характерный для взрывных или аффрикат. Таким образом, к критериям помехоустойчивости можно добавить критерий способа образования, где «фрикативный» был бы помехонеустойчивым значением. Исключение составляют шипящие согласные [ж] и [ш], для которых в русском языке нет смычных того же места образования.

## ЗАКЛЮЧЕНИЕ

Основная цель данного исследования заключалась в том, чтобы построить и протестировать слоговые артикуляционные таблицы, в которых учитывалась бы разная степень помехоустойчивости элементов к шуму.

Несмотря на то что выработанная нами система критериев оказалась достаточной для составления слоговых таблиц с разной степенью устойчивости к шуму, полноту решения этой задачи можно увеличить. С учётом использованных нами критериев, а также в дополнение к ним сведений, которые были получены по результатам контрольного тестирования, можно разработать более детальную систему критериев оценки помехоустойчивости. С её использованием удастся, возможно, создать таблицы, которые будут ещё более существенно отличаться по помехоустойчивости друг от друга, и достичь большей однородности элементов внутри каждой таблицы с точки зрения их помехоустойчивости.

Кроме того, можно отметить еще несколько направлений, в которых можно было бы продолжать данное исследование.

Во-первых, слоговой материал таблиц можно расширить слогами другой структуры и слогами, в которых используются не только ударные аллофоны гласных фонем. Материалом для исследований подобного рода могут служить не только слоги. Интересно также восприятие речи в шуме на материале более крупных единиц речи.

Во-вторых, в нашем исследовании была сделана попытка выявить зависимость эффекта, который оказывает шум на восприятие речи, от спектральных характеристик шума — белого и розового. Значимых отличий для распознавания речи с наложением этих видов шума выявлено не было, но, возможно, если экспериментировать с разным соотношением сигнал/шум и различными спектральными характеристиками шума, могут быть получены иные значимые результаты.

Наконец, восприятие в шуме можно сравнивать с другими видами помех и выяснить, какие свойства речевых единиц обеспечивают их общую перцептивную устойчивость, а какие обеспечивают устойчивость только к какому-то конкретному виду помех.

## **ЛИТЕРАТУРА**

1. Венцов А.В., Касевич В.Б. Современные модели восприятия речи: критический обзор. Проблемы восприятия речи. СПб, 1994.
2. Кодзасов С.В., Кривнова О.Ф. Общая фонетика. М., 2001.
3. Зиновьева Н.В. Система акустических ключей к распознаванию фонетических единиц русского языка // Экспериментальная фонетика. М. 1989. С.11–35.
4. Елкина В.М., Юдина Л.С. Статистика слогов русской речи // Вычислительные системы. Новосибирск, 1964. Вып. 10. С. 58–78.
5. Ягунова Е. В. Восприятие согласных фонем и их дифференциальных признаков. Автореферат диссертации на соискание ученой степени кандидата филологических наук. СПб, 1994.
6. Штерн А.С. Перцептивный аспект речевой деятельности. СПб, 1992.
7. Бондарко Л.В. Звуковой строй современного русского языка. М, 1977.
8. Алдошина И. Основы психоакустики. Москва, 2000.
9. Phatak S., Lovitt A., Allen J. Consonant confusions in white noise // J. Acoust. Soc. Am. 124, 1220 (2008).
10. Wang M.D.; Bilger R.C. Consonant confusions in noise: a study of perceptual features. // J. acoust. Soc. Am. 54: 1248–1266 (1973).



## **IMMUNITY OF SYLLABIC TABLES IN PERCEPTION OF SPEECH IN NOISE**

***Stanislav A. Krejci,***

*researcher of the faculty of Philology of Moscow state University  
they. M. V. Lomonosova*

***Olga F. Krivnova,***

*doctor of Philology, leading researcher  
philological faculty of Moscow state University. M. V. Lomonosov*

***Ekaterina A. Tikhonov,***

*bachelor of philological faculty of Moscow state University. M. V.  
Lomonosov*

### **Abstract**

This paper describes the methodology of constructing the test tables for study of speech perception in noise. The problem of this study was to build and test syllable articulation tables, which would take into account different degrees of elements' immunity to noise, and to identify those phonological features that are the most resistant to noise. It was found that syllable intelligibility of composed tables is about the same in the absence of noise and significantly different in white and pink noise.

**Keywords:** test tables pattern , speech perception, noise immunity, white noise, pink noise, syllable intelligibility, acoustic and phonological characteristics.

# Акустические и перцептивные признаки коартикуляционной назализации гласных в русском языке

**Екатерина Геннадьевна Солонина,**  
аспирантка кафедры ТиПЛ филологического факультета МГУ  
им. Ломоносова

## Аннотация

В статье излагаются результаты исследования, посвященного изучению коартикуляционной назализации гласных русского языка в рамках фонетического слова. Анализируются случаи непосредственного соседства с носовыми согласными гласных различной степени редукции, в позиции двусторонней и односторонней назализации. Основными акустическими признаками назализации по результатам исследования являются две форманты назализации на частотах ~250–300 Гц и ~900–1100 Гц, при этом в односторонних контекстах инерционная назализация проявляет себя в большей степени, чем предвосхищающая. В позиции конечного слога, по сравнению с другими слоговыми позициями, гласный, следующий за носовым согласным подвергается большей назализации, тогда как гласный, предшествующий конечному носовому согласному, напротив, отражает наименьшую степень назализации. Перцептивный эксперимент показал, что при восприятии назализованных гласных главными факторами, влияющими на успешность распознавания признака назализации, являются позиция носового согласного и степень редукции гласного: в случаях ударного гласного признак назализации распознается лучше, чем для заударных гласных, при этом инерционная назализация распознается в большем количестве случаев, чем предвосхищающая. Результаты исследования могут быть использованы в практических задачах, таких как автоматическое распознавание и синтез речи.

**Ключевые слова:** русский язык, фонетика, гласные, коартикуляция, назализация.

## ВВЕДЕНИЕ

Речевая коммуникация человека является сложной системой, в которой происходят различные процессы порождения и обработки речевого сигнала. Он характеризуется большой вариативностью звуковых элементов, на которые оказывают действие как внешние, так и внутренние факторы. Звуковые цепочки в речи не имеют дискретного характера, каждый звук (реализация фонемы) характеризуется переходными участками, на которых происходит перестроение артикуляционных органов от одной целевой артикуляции к другой, и качество звука, в особенности гласного, во многом зависит от его фонетического окружения. При этом контекстное влияние может также распространяться и на стационарные фазы их произнесе-



ния. Артикуляционно-акустическая зависимость реализации фонем от контекста выражается прежде всего в явлении коартикуляции. Ярким примером одного из таких коартикуляционных эффектов может служить контекстная назализация гласных. Поскольку назализация гласных в русском языке не имеет смыслоразличительной функции, этот признак может свободно возникать и варьироваться в зависимости от индивидуальных произносительных особенностей говорящего, однако существуют и универсальные тенденции развития коартикуляционной назализации, степень проявления которой меняется в зависимости от контекста гласного. Назализация в соседстве с носовыми согласными является частным и в то же время типичным случаем реализации коартикуляционных эффектов, относящихся к самому процессу речепорождения и обусловленных механизмами, устройством и работой речевого аппарата человека.

В общем случае коартикуляция определяется как взаимовлияние целевых артикуляций соседних фонем в звуковой последовательности [5], в результате чего происходит перекрытие фаз реализации артикуляционных жестов на сегментах соседних звуков. Это позволяет обеспечить слитность и высокую скорость передачи информации за счет облегчения перехода от одной фонемы к другой, сохраняя при этом высокий уровень разборчивости речи. В основе коартикуляции лежит в первую очередь невозможность мгновенно перестроить артикуляторы в положение, необходимое для произнесения последующей фонемы, в результате чего каждая звуковая реализация состоит из фазы выдержки, на которой в идеале и должна осуществляться целевая артикуляция, и переходных участков, экскурсии (до выдержки) и рекурсии (после выдержки). В потоке речи рекурсия, т.е. конечная фаза артикуляционного жеста одной фонемы, обычно совпадает или пересекается во времени с экскурсией, т.е. начальной фазой артикуляционного жеста последующей, но артикуляционное взаимовлияние может наблюдаться и на фазе выдержки. По расположению влияющего звука относительно целевого (исследуемого) различают коартикуляцию инерционную / прогрессивную (влияние предшествующего звука) и предвосхищающую / регрессивную (влияние последующего). Из числа других причин возникновения коартикуляционных явлений можно отметить то, что говорящий не нацелен на само речепроизводство, его цель — донести до слушающего информацию, затратив при этом минимум усилий и времени. Предполагается, что звуковая модель слова в сознании говорящего задается и активизируется целиком, а не по фонемам, поэтому при отсутствии фонематических артикуляционных ограничений положение речевых органов для произнесения последующей фонемы может подготавливаться заранее, а определенные элементы артикуляции могут сохраняться и после произнесения предыдущей фонемы [3]. Кроме того, различные артикуляторы имеют разную массу и степень подвижности и могут быть связаны между собой различными мышцами, что создает для каждого артикулятора свои особенности перемещения и может обеспечивать большую или меньшую длительность коартикуляционных эффектов для разных фонетических признаков в зависимости от свойств конкретных артикуляционных органов, участвующих в их реализации.

Коартикуляционная назализация гласных возникает за счет запаздывающего или предвосхищающего движения небной занавески, необходимого для произнесения соседнего носового согласного. При произнесении носового согласного небная занавеска опускается, открывая тем самым носовой проход для воздушного потока. Носовая полость в этом случае принимает на себя функции основного резонатора [7]. При прогрессивной коартикуляции небная занавеска возвращается в закрытое положение не сразу в силу инертности артикуляторов, и на экскурсии последующего гласного звука носовой проход все еще остается открытим. В случае следования носового согласного за гласным небная занавеска начинает подниматься на гласном еще до начала ротовой артикуляции носового согласного, предвосхищая ее. Необходимо отметить, что небная занавеска является одним из наименее подвижных артикуляционных органов и наиболее обособленных от других, что делает назализацию одним из наиболее свободно возникающих коартикуляционных типов, особенно в том случае, когда в языке отсутствует фонологический признак назализации гласных.

Анализ существующих фонетических работ показывает недостаточность исследования признаков коартикуляционной назализации гласных в русском языке. В качестве основного акустического признака назализации отмечается появление в спектральной картине гласного дополнительной форманты назализации ( $F_n$ ) на частоте 200–300 Гц [5], являющейся наиболее выраженной областью в акустической картине носовых согласных и распространяющейся на соседний гласный. Помимо указанной форманты назализации различными исследователями также отмечаются и другие акустические признаки. Л.В. Бондарко указывает на появление дополнительных формант на участках 500–1000 Гц или 1500–2000 Гц и ослабление второй ротовой форманты [1]. Данные о дополнительной форманте между F1 и F2 приводятся в работах Л.В. Златоустовой [4]. Однако под вопросом остается регулярность и закономерность появления данных спектральных признаков. Помимо формант назализации важной особенностью спектра носовых согласных является наличие антиформант — участков ослабления энергии в определенной частотной области спектра в связи с альтернативным носовым выходом воздушного потока [6]. Наиболее подробное изучение акустических признаков контекстной назализации гласных было проведено на материале английского языка. В работе [8] отмечено проявление в формантной структуре назализованного гласного двух формант назализации на частотах 300–400 Гц и 800 Гц для мужского голоса. Кроме того, указывается не только появление дополнительных формант, но и влияние антиформант носового согласного на соседний гласный, проявляющееся как ослабление ротовых формант на участках, оказывающихся под влиянием антиформант носовых согласных [8]. В исследовании коартикуляционной назализации гласных в американском варианте английского языка [9] в качестве основного признака назализации также отмечается дополнительная форманта  $F_n$ , располагающаяся ниже частоты первой ротовой форманты гласного, которая для гласных с низкой F1 может совпадать с первой собственной формантой гласного, что, как отмечают авторы, затрудняет интерпретацию акустической картины назализации.

Звуковые единицы могут быть описаны не только с позиции физиологии речепорождения или акустических свойств, но и с точки зрения их восприятия и анализа перцептивных характеристик [1]. Все эти задачи входят в сферу исследования при экспериментально фонетическом анализе речи. Исследование перцептивной системы позволяет выявить, какие акустические признаки являются наиболее важ-



ными для носителей конкретного языка при идентификации звуковой единицы. Для изучения собственной идентификации акустических признаков сигнала в качестве стимульного материала обычно используются изолированные звуки или слоги [2], поскольку в процессе такого эксперимента аудитор перекодирует слуховой сигнал в структуру фонемной записи без отождествления стимула с имеющимися в сознании испытуемого словарными единицами.

В нашей работе была поставлена задача рассмотреть гласные, подвергающиеся коартикуляционной назализации, и выявить акустические и перцептивные признаки их назализации. В рамках акустического исследования была проанализирована также степень проявления признаков назализации в зависимости от контекстных характеристик целевого гласного, характера коартикуляции и степени редукции гласного. Перцептивный эксперимент был направлен на определение контекстов распознавания носителями русского языка признака назализации изолированных ударных и редуцированных гласных, находящихся исходно в позиции непосредственной близости с носовым согласным. Проявление коартикуляционной назализации гласных рассматривалось только в рамках фонетического слова, назализация на границе слов не исследовалась.

## **МАТЕРИАЛ И МЕТОДИКА ИССЛЕДОВАНИЯ**

Материалом исследования послужила дикторская запись изолированных слов и словосочетаний в мужском произнесении, изначально предназначавшаяся для задач синтеза русской речи. Запись была сделана в лабораторных условиях и включала в себя представительную часть возможных контекстных реализаций гласных.

Для исследования были выбраны примеры как двустороннего, так и одностороннего соседства гласных с носовыми согласными. В случае, когда в соседстве с гласным находился единственный носовой звук, в качестве второго согласного выступал взрывной того же места образования как артикуляционно наиболее близкий носовому согласному, поскольку ротовая артикуляция такого согласного отличается от носовых только положением небной занавески. В материал исследования были включены все гласные фонемы русского языка (/а/, /о/, /е/, /ү/, /и/) и носовые согласные /н/ и /м/, их твердые и мягкие варианты.

В акустическом исследовании итоговое количество анализируемых гласных в соседстве с носовым согласным составило 426 единиц. Основное разбиение гласных по типам производилось на основании положения целевого гласного относительно носового согласного, а также в зависимости от степени редукции гласного. По положению целевого гласного относительно носового согласного контексты были разбиты на следующие типы: положение между двумя носовыми (НГН) / положение после носового согласного перед взрывным (НГС) / положение после носового согласного в позиции абсолютного конца слова (НГ#) / положение перед носовым после взрывного (СГН) / положение пе-

ред носовым в позиции абсолютного начала слова (#ГН). Количество примеров для каждого из выделенных позиционных контекстов гласного представлено в таблице 1.

*Таблица 1  
Количественное распределение материала по положению целевого гласного относительно носового согласного в непосредственном окружении*

Контекст	НГН	НГС	НГ#	СГН	#ГН
Количество исследуемых примеров	43	141	97	136	9

Разбиение контекстов по принципу положения целевого гласного относительно ударения в слове осуществлялось по пяти типам: ударный, первый предударный, первый заударный в неконечной позиции, первый заударный в позиции абсолютного конца слова и остальные случаи.

*Таблица 2  
Количественное распределение материала по положению целевого гласного относительно ударения*

Позиция гласного относительно ударения	Количество примеров в материале
Ударный	155
Первый предударный слог	98
Первый заударный слог в неконечной позиции	57
Первый заударный слог в позиции абсолютного конца	56
Остальные позиции	60

В качестве инструментария для вычисления частотных параметров и анализа формантных картин целевых звуков использовалась компьютерная программа для обработки речи Speech Analyzer, версия 3.1, при помощи которой исследовались осциллограммы, динамические спектрограммы и мгновенные спектры звуков.

На предварительном этапе исследования была произведена сегментация целевого гласного и его непосредственного окружения, подсчитана длительность целевого гласного, которая использовалась далее при измерении длительности участка его назализации. В качестве базы для сравнения и выявления эффектов назализации были взяты определенные для данного голоса частоты формант гласных из контекста между взрывными зубными и переднеязычными согласными как артикуляционно наиболее близкими носовым согласным, а также собственные формантные характеристики носовых согласных. Анализ акустических признаков назализации осуществлялся на гласных, находящихся в позиции между двумя носовыми согласными, их спектральная картина сопоставлялась с гласными в позиции между двумя взрывными согласными с целью выявления формантных характеристик, показывающих назализованные эффекты. На следующем этапе анализа исследовались контексты односторонней коартикуляционной назализации, для которых определялась длительность участка назализации гласного.



В перцептивном эксперименте в качестве материала использовалась аудиозапись мужского голоса, на основе которой проводился акустический анализ. Важной задачей в рамках данного эксперимента было выявление возможных различий в восприятии назализованных гласных в зависимости от степени их редукции, в связи с чем экспериментальный аудиоматериал включал два типа контекстов: с положением гласного в соседстве с носовым согласным в ударном слоге и с положением гласного в соседстве с носовым согласным в первом заударном слоге, исключая позицию абсолютного конца слова как снижающую степень редукции гласного. Такой выбор материала позволяет проверить гипотезу влияния степени редукции на восприятие назализованности гласных, поскольку указанные группы гласных представляют собой типы, наиболее контрастирующие по степени редукции. В состав экспериментальных ударных и заударных гласных вошло 56 единиц каждого типа. В материал эксперимента вошли также филлеры, в качестве которых выступали гласные, вырезанные из контекста между двумя взрывными согласными. Степень редукции гласных филлеров совпадала со степенью редукции экспериментальных гласных стимулов. Заполнение материала филлерами происходило по схеме  $\Phi *-*-\Phi-*-$ , где знаком \* указан целевой стимул. Исходя из этой схемы, количество гласных-филлеров составило 23, а общее гласных в каждой части эксперимента — 79.

Проведение эксперимента осуществлялось в программе iSpring, позволяющей удаленно проводить анкетирование с прослушиванием аудиоматериала. Перед началом тестирования аудиторам предлагалось выставить наиболее удобную для них громкость звучания, для чего был предоставлен тестовый аудиофрагмент, представляющий собой изолированный ударный гласный, находившийся изначально в контексте между взрывным и плавным согласными; перед прохождением тестирования аудиторам предлагалась инструкция для ознакомления с условиями эксперимента. Участникам предъявлялись поочередно аудиофайлы, содержащие тройное повторение целевого гласного без контекста. Задача аудиторов состояла в том, чтобы записать буквами русского алфавита то, что они слышат (метод «диктанта»). Переход к следующему аудиофайлу осуществлялся при нажатии специальной иконки на экране после заполнения строки для ответов, испытуемые, таким образом, сами регулировали скорость прохождения эксперимента. При нажатии на кнопку перехода к следующему звуку аудиофайл проигрывался автоматически, повторное прослушивание звуков было невозможно, невозможен был также переход к следующему звуку без заполнения актуальной строки ответов. Каждый изолированный гласный звучал трижды с интервалом  $0,5\pm0,05$  с.

В эксперименте приняли участие 12 человек в возрасте от 18 до 56 лет, носители русского языка без дефектов слуха. Оценка распознавания назализации гласного проводилась на основе наличия / отсутствия в ответах аудиторов носовых согласных, поскольку, в силу коартикуляции, гласный, вычлененный из речевого сигнала, несет информацию о звуках, находящихся в его непосредственной близости, в особенностях о предшествующем согласном. В результате изолированный звук, вырезанный из контекста, в большинстве случаев воспринимается

**Таблица 3**  
**Количественное распределение стимульного материала эксперимента по положению целевого гласного в слоге**

Позиция целевого гласного в слоге	Кол-во представленных в эксперименте ударных гласных	Кол-во представленных в эксперименте заударных гласных
НГН	5	4
НГС	20	20
СГН	12	32
НГ#	19	-

на слух как последовательность гласного с окружающими его согласными. Это свойство позволяет расценивать восприятие признака назализации гласного через указание в ответах аудиторов носового согласного в качестве соседнего элемента. Необходимо отметить, что при такой оценке результатов указание носового согласного в качестве соседствующего элемента однозначно свидетельствует о распознавании слушающим назализации гласного, тогда как в случаях отсутствия в ответах носового согласного утверждать, что аудиторы не восприняли гласный как назализованный, следует с определенной осторожностью. Возможно, признак назализации был идентифицирован, но участники эксперимента не проинтерпретировали назализацию как соседство носового согласного. Аудиторы могут распознавать признак назализации, но не относить его к влиянию носового согласного, а интерпретировать его, например, как индивидуальную особенность гласного в специфическом произнесении диктора.

## РЕЗУЛЬТАТЫ АКУСТИЧЕСКОГО ИССЛЕДОВАНИЯ

По результатам проведенного анализа при характеристике русских носовых согласных нельзя не отметить наличие двух особых формант на частотах ~250–300 и ~900–1100 Гц, что подтверждает полученные ранее данные других исследователей, в том числе на материале русского языка. При этом вторая форманта имеет заметно меньшую интенсивность и окружена частотными участками со сниженной энергией (антиформантами). Частотные значения усиленных участков для переднеязычного носового /н/ в среднем несколько выше, чем для губного /м/. Назализация на соседних гласных проявляется как отражение в формантной картине гласных признаков, присущих носовым согласным. При этом в зависимости от расположения собственных ротовых формант гласного проявление признаков назализации может иметь различный характер. Форманта назализации не всегда проявляется на графике как самостоятельный частотный пик. Так, для гласных верхнего подъема, имеющих низкую первую ротовую форманту, первая форманта назализации на частоте ~250–300 Гц может быть не видна на графиках, поскольку в таких случаях происходит наложение этой форманты носового согласного на собственную форманту гласного. В тех случаях, когда ротовая форманта гласного близка по частоте форманте назализации, она может приближаться по значению к Fn либо иметь увеличенную ширину.

Ниже приводится пример слова «банану» с графиком спектрального среза на середине ударного гласного, где мы можем наблюдать наличие дополнительной форман-



ты на частоте ~300 Гц, а также уменьшение частоты второй ротовой форманты до значения ~1180 Гц вместо ~1250–1350 Гц, характерной для ударного гласного между двумя зубными взрывными (по данным нашего диктора). В области выше частоты F2 (~1400 Гц) наблюдается значительный спад интенсивности спектра, что не отмечалось в случае неназализованных гласных.

Вторая форманта назализации ярко проявляется в случаях с гласными переднего ряда, для которых собственная F2 располагается выше области второй Fn (рис. 2).

Анализ длительностей участков назализации в односторонних контекстах показал, что инерционная назализация (в положении целевого гласного после носового согласного) имеет большую длительность, присутствует регулярно и в среднем затрагивает около половины длительности гласного. Минимальное распространение назализации в контекстах НГС составляет около трети от длительности гласного. В некоторых случаях инерционная назализация покрывает всю длительность гласного, см. ниже пример на рис. 3.

Анализ длительности назализации гласных проводился в односторонних консонантных контекстах. Измерение длительности назализованного участка связано с определенной степенью условности в связи с недискретной природой звучащей речи и плавными переходами формантных значений. Длительность назализованного участка измерялась с использованием динамических спектрограмм и с опорой на слуховой контроль, позволяющий прояснить наблюданную формантную картину гласного, что особенно важно для гласных, формантная картина назализации которых близка к значениям их ротовых формант.

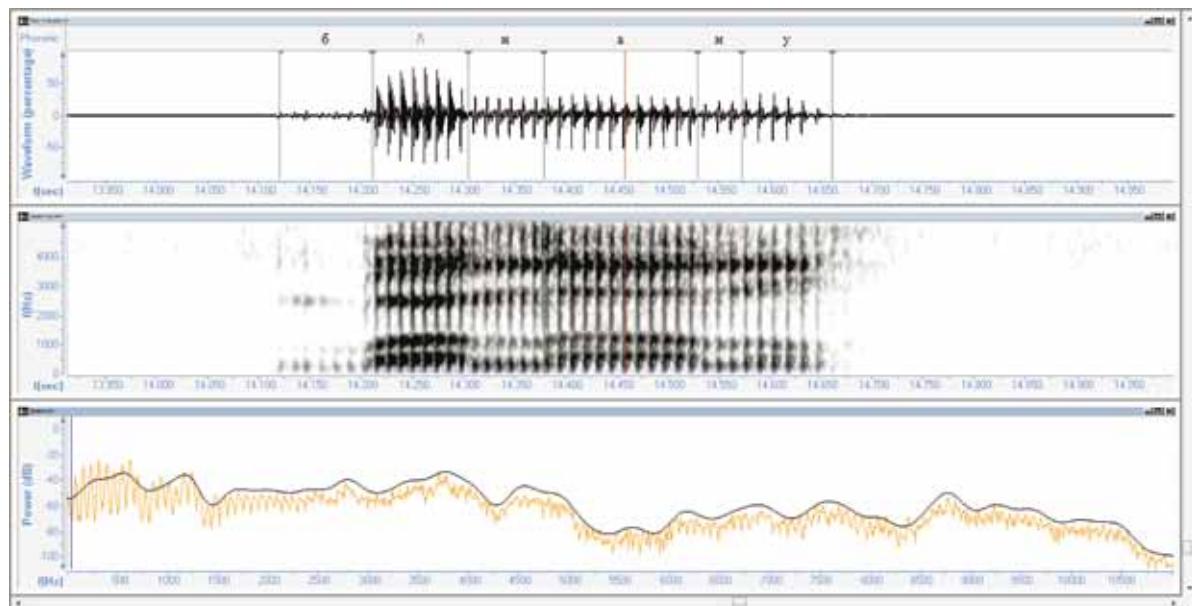


Рис. 1. Осциллограмма, спектrogramma и график спектрального среза, взятого на середине ударного гласного для слова «*банану*»

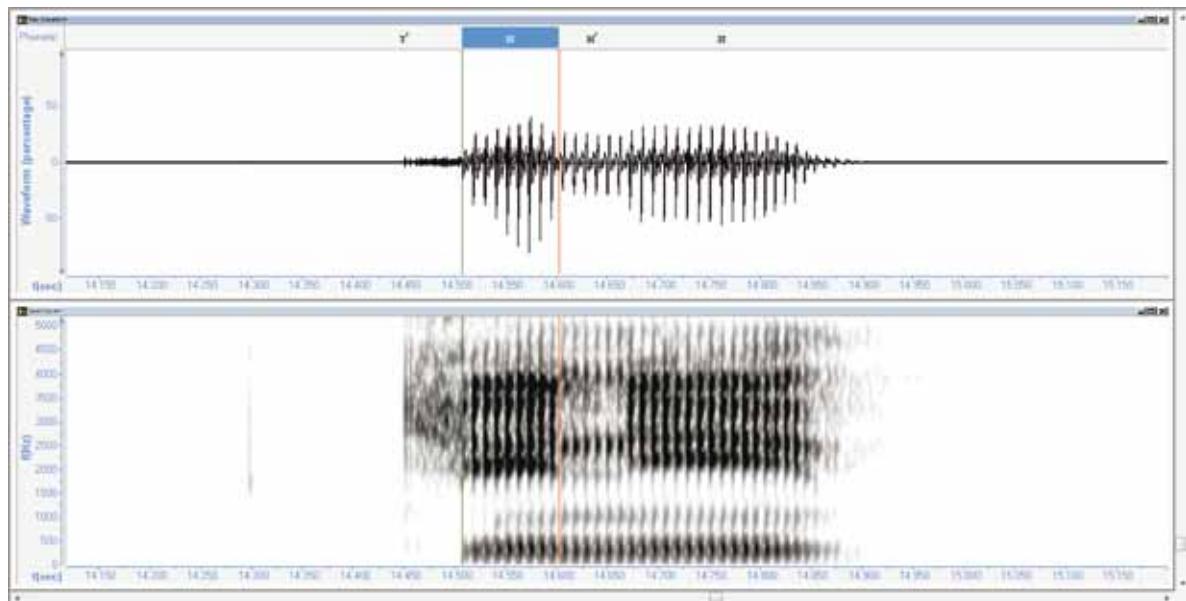


Рис. 2. Осциллограмма и спектрограмма слова «тяни», выделены границы первого предударного гласного

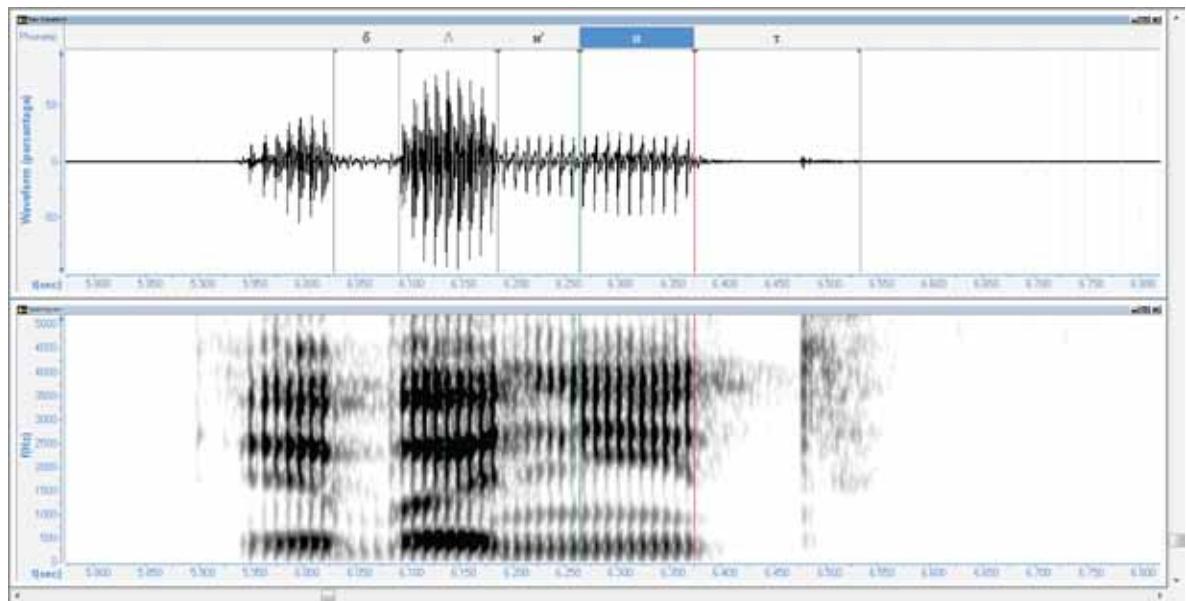


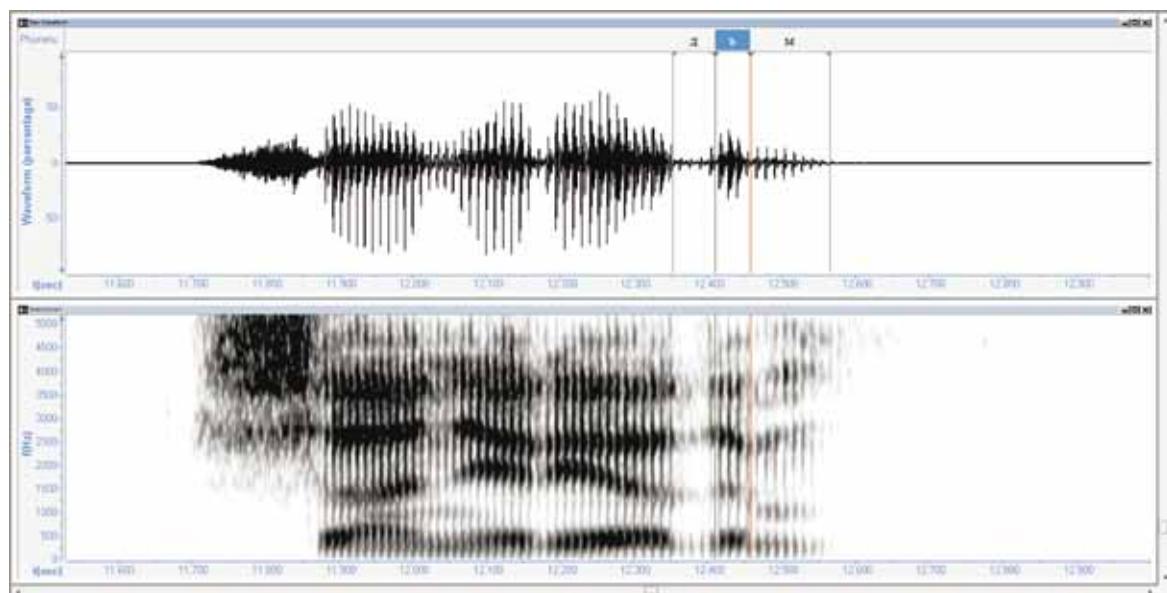
Рис. 3. Осциллограмма и спектрограмма слова «эбонит», ударный гласный полностью назализован



При предвосхищающей коартикуляции (в положении целевого гласного перед носовым согласным) проявление назализации менее продолжительно. В большинстве случаев назализация не распространяется на стационарный участок гласного и занимает около 30% от длительности гласного (см. выше рис. 2). Встречаются также примеры, в которых назализация гласного отсутствует, как в случае с кратким заударным гласным в конечном слоге «сани рядом» (рис. 4).

Особый случай представляют гласные конечного слога слова. В этой позиции инерционная назализация затрагивает значительно больший участок гласного (все случаи полной назализации, покрывающей всю длительность гласного в односторонних контекстах, обнаружились именно в позиции конечного слога), тогда как предвосхищающая назализация, напротив, имеет значительно меньшую длительность (большинство примеров полного отсутствия назализации принадлежат к именно конечным слогам). В контекстах НГС ударных гласных, где распределение конечных и неконечных слогов с целевыми гласными составило 10:10, частичная назализация гласных непоследнего слога охватывает около 50% их длительности, тогда как средний размер назализации гласных последнего слога в слове составляет 70–75% их длительности.

В заключение отметим, что при анализе спектральных данных надо иметь в виду, что присутствие формант назализации на акустических графиках гласных отображается либо как дополнительное частотное усиление, либо как увеличение ширины форманты, в том случае, когда форманта гласного и форманта назализации близки по частоте. Влияние антиформант проявляется в ослаблении интенсивности ротовых формант гласного на частотах, близких к антилокусам носового согласного.



**Рис. 4.** Осциллограмма и спектрограмма словосочетания «саны рядом», выделены границы заударного гласного в конечном слоге

Таким образом, проявление назализации различается в зависимости от типовой F-картины гласного и расположения локусных частот и антилокусов соседнего носового согласного.

## РЕЗУЛЬТАТЫ ПЕРЦЕПТИВНОГО ЭКСПЕРИМЕНТА

Общей особенностью восприятия назализации гласных является зависимость успешности ее распознавания от степени редукции гласного. По результатам анализа длительности назализованного участка ударные и редуцированные гласные не обнаружили существенного различия, тогда как при восприятии признак редукции гласного оказался значимым по влиянию на распознавание назализации. Так, признак назализации ударных гласных был распознан в 63% случаев, а для редуцированных гласных доля распознавания оказалась равной 32%.

*Таблица 4  
Количество распознаваний признака назализации  
в зависимости от степени редукции гласного*

	Ударный гласный	Заударный гласный
Кол-во распознаваний признака назализации	421 из 672 (63%)	216 из 672 (32%)

Различие в степени назализации в зависимости от положения носового согласного также находит отражение в количестве случаев успешного распознавания назализации: наибольшее количество распознавания отмечается в контекстах двусторонней назализации (82%). В позиции после носового перед взрывным согласным количество распознаваний составляет 65%, а при предвосхищающей назализации — 12%.

*Таблица 5  
Количество распознаваний признака назализации гласного  
в зависимости от позиции носового согласного в исходном слове*

	НГН	НГ#	НГС	СГН
Кол-во распознаваний признака назализации	89 из 108 (82%)	171 из 228 (75%)	313 из 480 (65%)	64 из 528 (12%)

Различие между контекстом НГН и контекстом НГ# оказалось статистически не значимым (критерий Х<sup>2</sup>: 2,2974 < 3,84146, при  $\alpha=0,05$ ), а различие в контекстах НГ# НГС и НГС СГН прошло статистическую проверку (критерий Х<sup>2</sup>: 6,85209 > 6,6349, при  $\alpha=0,01$  и 302,652 > 7,87944, при  $\alpha=0,005$ , соответственно).

Таким образом, основными факторами, влияющими на количество успешных распознаваний признака назализации гласного, оказались степень редукции и исходное положение целевого гласного относительно носового согласного.



## **ЗАКЛЮЧЕНИЕ**

При акустическом анализе было установлено различие в степени назализации гласного в зависимости от его положения относительно носового согласного, а также особое проявление назализации гласного в позиции конечного слога слова. Перцептивный эксперимент подтверждает различие в восприятии признака назализации в зависимости от положения гласного относительно носового согласного.

1. Важным фактором, влияющим на успешное распознавание назализованных гласных, является также степень редукции гласного: позиция первого заударного слога оказалась более трудной для восприятия признака назализации гласного по сравнению с ударным гласным.
2. Особое проявление назализации гласных в конечном слоге слова может быть объяснено общим спадом артикуляционных усилий к концу высказывания, в результате чего конечный согласный оказывает минимальное влияние на предшествующий звук, что уменьшает или, напротив, увеличивает назализацию гласного в случаях предвосхищающей и инерционной назализации соответственно.
3. Статистические выводы о проявлении назализации в зависимости от качества и редукции гласного были осложнены тем, что материал данного исследования не был сбалансирован по представленности и частотности фонем и их контекстов. В нашем материале по техническим причинам гласные различного качества имеют разную встречаемость и не все контексты оказались учтены.

В перспективе необходимо провести более глубокое и широкое исследование на сбалансированном фонетическом материале в условиях разного устного дискурса, прежде всего в слитной русской речи. Для прикладных и практических задач очень полезно было бы проанализировать контекстную коартикуляционную назализации гласных в русском языке в сравнении с другими языками, в том числе и такими, где имеется фонологическая назализация гласных, например французским или португальским языком. Выявленные закономерности можно было бы использовать в задачах автоматической идентификации языка и говорящего, а также для устранения произносительного акцента в иноязычной речи.

## **ЛИТЕРАТУРА**

1. Бондарко Л.В. Звуковой строй современного русского языка. – М., 1977.
2. Венцов А.В., Касевич В.Б. Проблемы восприятия речи. – М., 2003.
3. Зиндер Л.Р. Общая фонетика. – М.: Высшая Школа, 1979.
4. Златоустова Л.В., Потапова Р.К., Трунин-Донской В.Н. Общая и прикладная фонетика. – М., 1997.
5. Кодзасов С.В., Кривнова О.Ф. Общая фонетика. – М., 2001.
6. Потапова Р.К., Потапов В.В. Речевая коммуникация. От звука к высказыванию. – М., 2012.

7. Чистович Л.А., Венцов А.В. и др. Физиология речи. Восприятие речи человеком. – Л., 1976.
8. Harrington, Cassidy Techniques in Acoustic Phonetics, Kluwer: Dordrecht, 1999, Chapter 4 Acoustic Phonetics.
9. Olive J., Greenwood A., Coleman J. Acoustics of American English Speech. Springer-Verlag, 1992.

## **ACOUSTICAL AND PERCEPTUAL FEATURES OF COARTICULATORY NASALIZATION OF RUSSIAN VOWELS**

***Ekaterina G. Solonina,***  
*postgraduate student, Department of Tipl philological faculty of  
Moscow state University Lomonosov's*

### **Abstract**

This paper concerns the problem of coarticulatory nasalization of Russian vowels adjacent to nasal consonants. The objective of the current study was to determine acoustic and perceptual features of nasalized vowels in different phonetic contexts and measure the extent of nasalization. In comparison with the corresponding vowels within stop consonants (CVC) the spectrum of nasalized vowels is characterized by two main nasal formants: Fn1 in the 250–300 Hz region and Fn2 in the 900–1100 Hz region. Concerning the temporal extent of vowel nasalization the data show that vowels in the nasal context (NVN) are completely nasalized. But in the position of one adjacent nasal consonant nasalization of vowels following a nasal consonant (NVC) appears to be more extensive than in the case of vowels preceding a nasal consonant (CVN). The study reveals that position of the syllable on word boundaries may have influence on the extension of nasalization. Thus, vowels in syllable-final position from the context NVC is more nasalized in comparison with other syllable positions, whereas vowels from the context CVN appear to be the least nasalized in syllable-final position. Perceptual experiment shows better recognition of nasalization in cases of stressed syllable and carryover nasalization as compared with anticipatory one.

**Keywords:** Russian language, phonetics, vowels, coarticulation, nasalization.



# Сравнительный анализ характеристик голоса и речи детей типично развивающихся, с расстройствами аутистического спектра, синдромом Дауна и умственной отсталостью

**Елена Евгеньевна Ляксо,**  
доктор биологических наук, профессор, профессор кафедры ВНД  
и психофизиологии биологического факультета СПбГУ, руководитель  
группы по изучению детской речи

**Ольга Владимировна Фролова,**  
кандидат биологических наук, кандидат биологических наук,  
научный сотрудник кафедры ВНД и психофизиологии биологического  
факультета СПбГУ, группа по изучению детской речи

**Алексей Сергеевич Григорьев,**  
аспирант кафедры ВНД и психофизиологии биологического факультета  
СПбГУ, группа по изучению детской речи

**Виктор Александрович Городный,**  
магистрант кафедры ВНД и психофизиологии биологического  
факультета СПбГУ, Группа по изучению детской речи

## Аннотация

Целью исследования явилось выявление специфических особенностей голоса и речи детей с расстройствами аутистического спектра (РАС), синдромом Дауна (СД), умственной отсталостью разной степени выраженности (УО) по сравнению с типично развивающимися детьми (ТР). Анализ речи детей включал перцептивный эксперимент, направленный на выявление взрослыми значения слова, сказанного ребенком, и спектрографический анализ временных и частотных характеристик вокализаций и слов детей. Определяли длительность высказываний, слов, ударных и безударных гласных и их стационарных участков; значения частоты основного тона (ЧОТ) и диапазона ЧОТ высказываний и гласных, значения ЧОТ, формантных частот и их интенсивности на стационарных участках гласных. Выявлены акустические характеристики речи детей с синдромом Дауна, РАС, УО, отличные от соответствующих характеристик речи ТР детей.

**Ключевые слова:** детская речь, спектрографический анализ, индекс артикуляции, расстройства аутистического спектра, синдром Дауна, умственная отсталость

## **ВВЕДЕНИЕ**

Проблема оценки состояния говорящего (в том числе и патологического) по характеристикам голоса и речи широко изучается с использованием различных методических подходов. Нарушение развития или атипичное развитие детей часто сопровождается нарушением речи и/или использованием речи в процессе коммуникации. Так дети с расстройствами аутистического спектра (PAC) характеризуются нарушением эмоционального, интеллектуального и социального развития, специфическими особенностями речи и повторяющимся репертуаром поведения [1]. Обобщенный диагноз — расстройства аутистического спектра включает ряд заболеваний, которые характеризуются различной степенью тяжести нарушений речи [2]. Детям с синдромом Дауна (СД), несмотря на сложности с артикуляцией, присуща развитая коммуникация, однако темпы формирования коммуникативных навыков ниже, чем у типично развивающихся сверстников.

В отношении акустических характеристик речи детей с PAC данные противоречивы. В одних работах отмечают монотонность речи [3], так показано, что у детей школьного возраста изменения частоты основного тона (ЧОТ) на основании оценки коэффициента вариативности значений ЧОТ в каждом слове высказывания меньше, чем у здоровых детей [3]. В других исследованиях указывают на увеличенный диапазон ЧОТ в речи детей с PAC 4–6,5 [4] и 4–10 лет, как и в обращенной ребенку материнской речи, что может, по мнению авторов, свидетельствовать о задержке речевого развития таких детей [5]. Установлена специфика спектра (на основе усреднения спектра записи речи на протяжении одной минуты) речи детей с PAC по сравнению с ТР детьми, что позволяет говорить о возможности использования особенностей спектральных характеристик в качестве количественных объективных биомаркеров речи детей с PAC [4]. Анализируют преимущественно речь детей с синдромом Аспергера [6] и высокофункциональным аутизмом [7, 8]. Характеристики речи детей с PAC, имеющих среднюю и тяжелую степень проявления аутистических расстройств, практически не изучены. Исследования акустических и перцептивных характеристик речи детей с PAC, воспитывающихся в русскоязычной среде, единичны [9, 10, 11, 12].

В зарубежной литературе широко представлены исследования характеристик речи детей с СД, в которых показано снижение разборчивости речи и чёткости артикуляции [13]. Обсуждается вопрос о том, можно ли считать специфику формирования речи детей с СД обусловленной общей задержкой развития или правильно рассматривать ее в качестве особого пути речевого развития вследствие особенностей строения речевого тракта и генетических нарушений [например, 13, 14]. Дети с СД при произнесении слов совершают ошибки, отличные от ошибок, характерных для детей с фонологическими нарушениями [15]. Подростки и взрослые с СД неправильно произносят согласные, с которыми у детей при нормативном развитии не возникает проблем: /d/, /t/, /n/ и /v/ [16]. Более сложные нарушения показаны у детей с СД по сравнению с детьми с аналогичной задержкой развития уже на стадии лепета [17]. В исследованиях с применением спектрографического метода анализа установлено уменьшение разницы между значениями формантных частот кардинальных гласных [18, 19]. В целом чёткость артикуляции и разборчивость речи у людей с СД ниже, чем в норме.



Акустические характеристики речи детей с умственной отсталостью (УО) практически не изучены. Одним из подходов к изучению речи детей с УО является сопоставление с речью детей с РАС [20]. Установлено, что произношение и фонематический слух у детей с РАС нарушены в меньшей степени, чем у детей с УО. Повторы, штампы, характерные для детей с УО, у детей с РАС сохраняются дольше. Дети с УО используют диалоговую речь, но практически не пользуются монологической речью.

Целью настоящего исследования явилось определение акустических характеристик речи детей, специфичных для РАС, СД и УО.

## **МЕТОДИКА**

В исследовании приняли участие дети 5–12 лет — ТР по заключению педиатров ( $n=60$ ), с РАС (F84 — по МКБ 10 пересмотра, 1998;  $n = 30$ ); дети 5–7 лет — с СД (Q90,  $n=4$ ), умственной отсталостью (F70,  $n = 9$ ) и смешанными специфическими расстройствами психологического развития (CCP, F83,  $n = 14$ ). Дети с РАС разделены на две группы: дети с регрессом в развитии в возрасте 1,5–3 года (группа РАС-1) и дети, у которых риск развития диагностирован при рождении, РАС являются сопутствующими основному заболеванию. В анамнезах детей группы-2 (РАС-2) имеются органические нарушения мозга, задержка психического и речевого развития, гиперактивность, умственная отсталость и т.д. Эти дети после роддома какое-то время (от 2 недель до месяца) находились в больнице. Для оценки тяжести аутистических расстройств использовали шкалу CARS [21]. В обе группы вошли дети с разной степенью тяжести по аутизму, баллы по шкале CARS для детей двух групп на момент начала исследования значимо не различались.

Запись речи проводили на цифровой рекордер “Marantz PMD222” с выносным микрофоном “SENNHEIZER e835S”, с параллельной видеозаписью поведения ребенка на камеру Sony Handycam HDR-CX330. Запись речи и поведения детей проводили в домашних условиях, в помещении лаборатории, в детском саду и школе. Использовали стандартизованные ситуации записи: игру со стандартным набором игрушек, беседу с экспериментатором, просмотр мультфильма по iPad и пересказ его сюжета, просмотр картинок и ответы на вопросы по ним, повторение за экспериментатором слов [10, 11].

Ситуации записи были по возможности максимально сходными для всех детей, однако дети с РАС до экспериментальной сессии плавали в бассейне для снятия напряжения, в дальнейшем запись их вокализаций/речи осуществляли в присутствии родителей.

Инструментальную обработку речевого материала осуществляли на основе спектрографического анализа в звуковом редакторе “Cool Edit Pro”. Из речи детей выделяли слова, в которых определяли значения ЧОТ ( $F_0$ ), ее средние, максимальные и минимальные значения, вычисляли диапазон ЧОТ [ $F_{0max}$ - $F_{0min}$ ]; ударный и безударный гласный. Для ударного гласного определяли значение ЧОТ и выделяли стационарный участок. На стационарном участке гласного считали длительность,

значения ЧОТ и ее интенсивность ( $E_0$ ), значения трех формантных частот ( $F_1$ ,  $F_2$ ,  $F_3$ ) и их интенсивности ( $E_1$ ,  $E_2$ ,  $E_3$ ); производили нормирование значений интенсивностей формант по отношению к значениям интенсивности ЧОТ ( $E_n/E_0$ ). Строили формантные треугольники с вершинами значений кардинальных гласных /a/, /u/, /i/ и определяли их площади [22].

$$\text{Area} = 0,5 \times \{(F_2[i] \times F_1[a]) + (F_2[a] \times F_1[u]) + (F_2[u] \times F_1[i])\} - \\ - \{(F_1[i] \times F_2[a]) + (F_1[a] \times F_2[u]) + (F_1[u] \times F_2[i])\}, \quad (1)$$

где  $F_1[x]$  и  $F_2[x]$  — значения первой и второй формант соответствующих гласных.

Индекс артикуляции гласных звуков (VAI) вычисляли по формуле:

$$VAI = (F_1[a] + F_2[i]) / (F_1[i] + F_1[u] + F_2[a] + F_2[u]), \quad (2)$$

где  $F_1[x]$  и  $F_2[x]$  — значения первой и второй формант соответствующих гласных [23].

Строили гистограммы зависимости частоты форманты и ее нормированной интенсивности по отношению к интенсивности ЧОТ ( $E_n/E_0$ ).

Выбор для анализа определенного набора акустических признаков речевого сигнала обусловлен их информативностью при изучении акустических характеристик речи ТР детей и детей с РАС [9, 22].

Перцептивный эксперимент ( $n = 500$  аудиторов в возрасте  $22,2 \pm 4,8$  года) проводили с целью выявления возможности определения взрослыми значения слова ребенка, его эмоционального состояния, возраста, пола. В тестовых последовательностях каждое слово повторялось три раза, интервал между одинаковыми словами составлял 5 с, интервал между разными словами — 10 с. Эксперимент проводили в условиях открытого поля.

Статистическая обработка данных проведена в пакете «Statistica 10.0».

Исследование одобрено Этическим комитетом Санкт-Петербургского государственного университета.

## **РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ**

Анализ уровня речевого развития детей показал, что речь ТР детей включала слова, фразы и высказывания. Дети активно использовали речь в процессе коммуникации. С увеличением возраста детей в их лексиконе увеличивается количество слов, отражающих разное эмоциональное состояние. В речи ТР детей встречаются антонимы, представленные наречиями и прилагательными, глаголами.

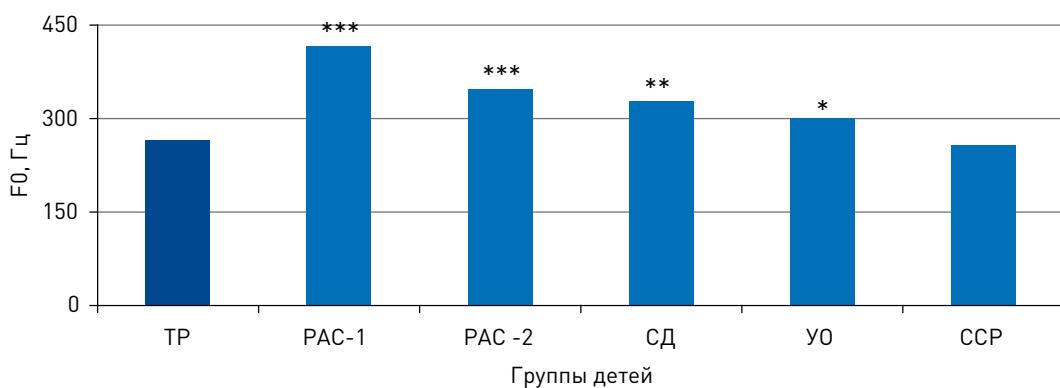
Дети с СД употребляли простые слова и речевые конструкции с нечеткой артикуляцией, они активно вступали во взаимодействие с взрослым с использованием всех «речевых» возможностей.

Репертуар детей с УО и ССП включал простые слова, фразы и высказывания, состоящие из нескольких фраз, которые дети использовали в ответных репликах в диалогах с взрослым.

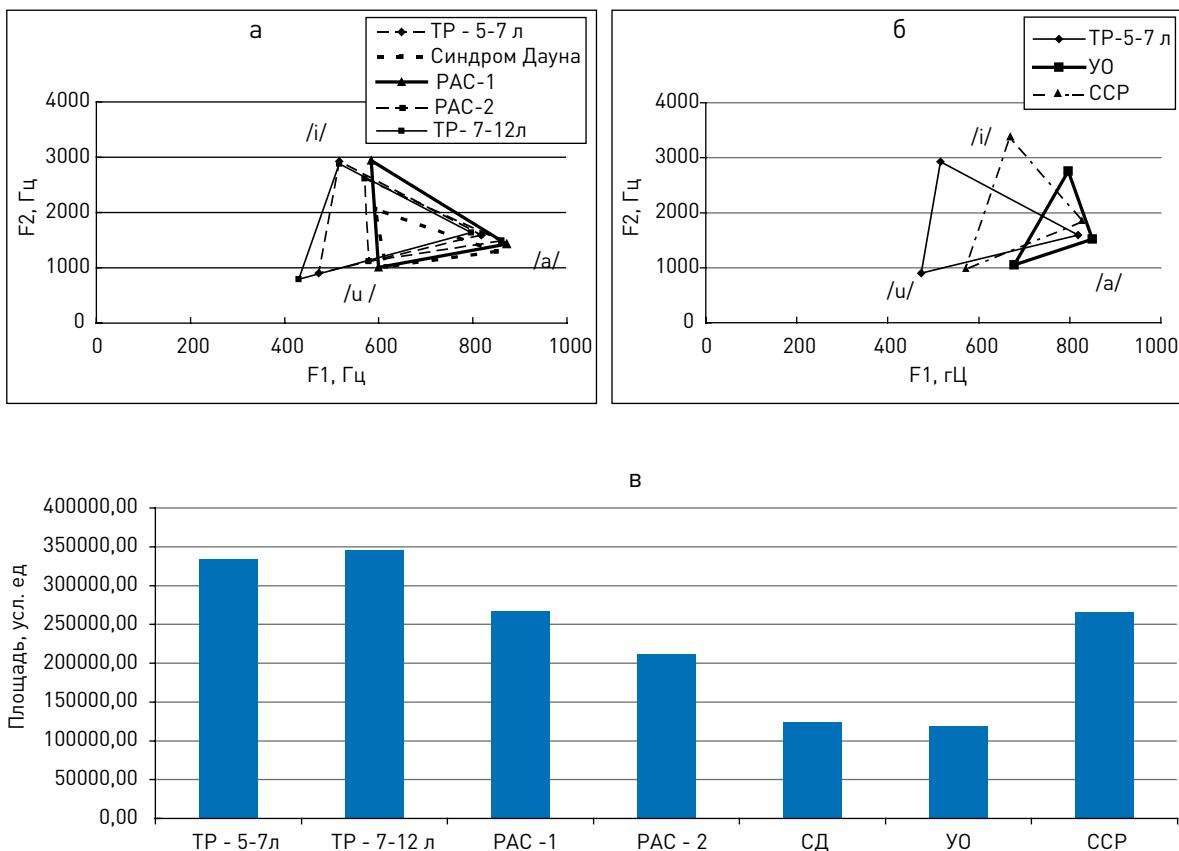


У детей с РАС наблюдали «нормальную» и «специфическую» речь. «Нормальная» речь детей с РАС представлена вокализациями, напоминающими лепетные конструкции, плачем, простыми словами и простыми фразами. Особенностью «нормальной» речи является несформированность разных уровней ее организации. Артикуляция, и/или просодика, и/или грамматика, и/или прагматика не соответствуют речи нормально развивающихся детей. «Специфическая речь» включает эхолалию — повторение слов, слов и фраз, и «свой язык» (введен условный термин) — звукосочетания с нечеткой артикуляцией, значение которых не понятно даже в конкретной ситуации. Для детей группы-1 характерно наличие вокализаций (13% детей), речи (47%), специфической речи (20%). У 20% детей группы-1 наряду с нормальной речью имеется специфическая речь. В группе-2 при наличии вокализаций (10%), речи (40%), специфической и нормальной речи (50% детей) дети, использующие только специфическую речь, отсутствуют. Различие между группами по сформированности речи у детей заключается в большем количестве детей группы-1, использующих речь ( $p<0,05$  — критерий Манна-Уитни), и значимо меньшим числом детей, использующих одновременно специфическую и нормальную речь ( $p<0,01$ ). У детей с РАС коммуникация с родителями, экспериментатором и другими детьми могла отсутствовать (47% детей), либо в процессе коммуникации с взрослым дети (53%) использовали простые реплики, состоящие из слова, одного слова, простой фразы, повторения части реплики взрослого. Важной характеристикой речевой функции у детей с РАС является нежелание использовать речь. Значимо большее количество детей группы-1 ( $p<0,01$ ) характеризуется отсутствием желания говорить.

Анализ акустических характеристик речи детей выявил более высокие значения ЧОТ у детей с РАС-1, РАС-2 ( $p<0,001$ ), синдромом Дауна ( $p<0,005$ ) и УО ( $p<0,01$ ) по сравнению с значениями ЧОТ у ТР детей и детей с ССР. По значениям ЧОТ ТР дети и дети с ССР не различаются. Речь детей группы РАС-1 характеризуется максимальными значениями ЧОТ (рис. 1).



**Rис. 1.** Значения ЧОТ на стационарном участке ударных гласных из слов и речевых конструкций ТР детей, детей с РАС, СД, УО и ССР. По вертикальной оси — значения ЧОТ, Гц, по горизонтальной — группы детей; \*\*\*  $p<0,001$ , \*\*  $p<0,005$ , \*  $p<0,01$  — критерий Манна — Уитни

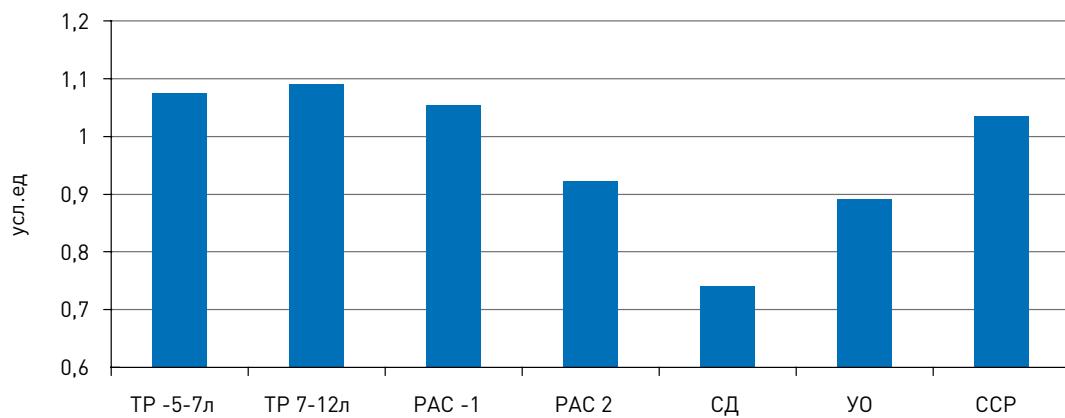


**Рис. 2.** Формантные треугольники ударных гласных /а/, /у/, /и/ из слов детей: ТР, с синдромом Дауна, PAC-1, PAC-2 (а), ТР, УО и CCP (б) на двухформантной плоскости с координатами F1, F2 и значения площадей формантных треугольников (в).  
На а, б – по горизонтальной оси – значения первой форманты – F1, Гц; по вертикальной – вторая форманты – F2, Гц;  
на в – по вертикальной оси – значения площадей формантных треугольников гласных, в усл. единицах

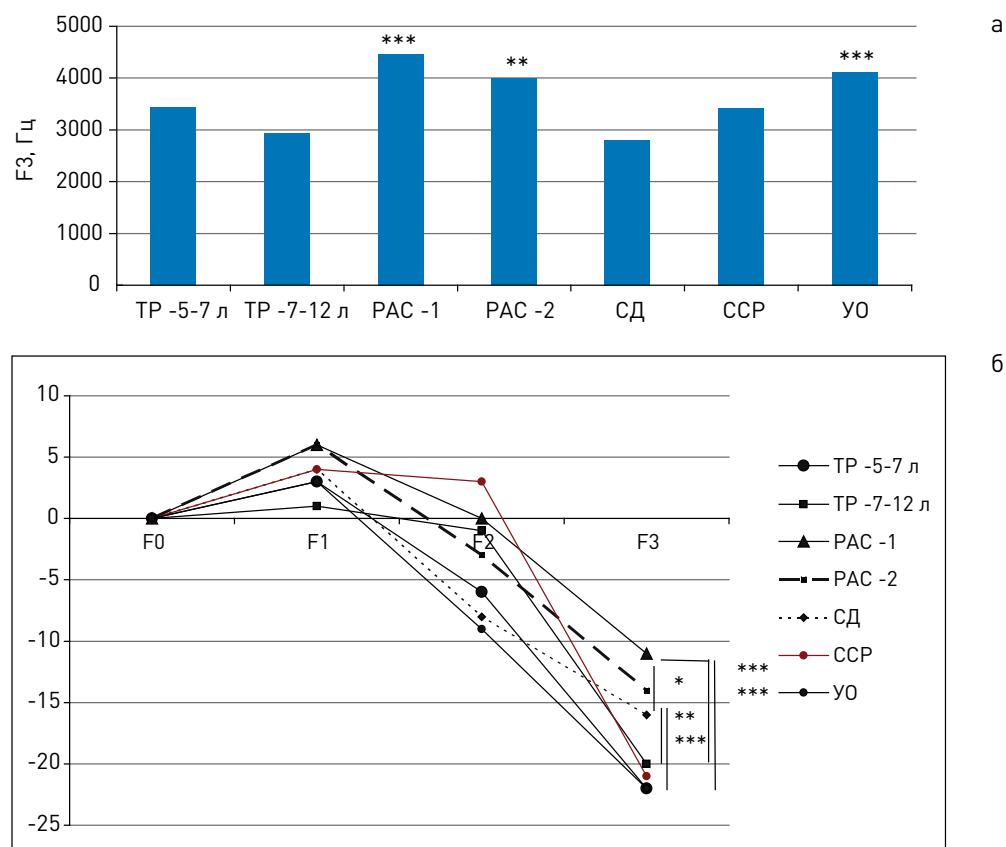
Формантные треугольники ударных гласных /а/, /у/, /и/ из слов (ситуация повторение слов) детей ТР, СД и PAC (рис. 2, а), ТР детей и детей с УО и CCP (б) различаются по форме и ориентации (рис. 2, а, б). Максимальные различия обусловлены смещением в высокочастотную область F1 гласных /у/ и /и/ у детей с атипичным развитием по сравнению с данными для ТР детей. Значения двух первых формант гласного /а/ значимо не различаются у детей всех групп.

Значение индекса артикуляции гласных минимально для детей с СД, что отражает специфику их нечеткой артикуляции, значимо не различаются у ТР детей в возрасте 5–7 и 7–12 лет, детей с PAC-1 и CCP (рис. 3).

Спектрограммы гласных в словах и вокализациях детей с PAC и УО отличается от спектрограмм ТР детей и детей с СД и CCP выраженным высокочастотными составляющими. Значения частоты F3 гласных значимо выше у детей с PAC-1 ( $p<0,001$ ), PAC-2 ( $p<0,01$ ) и детей с УО ( $p<0,001$ ) по сравнению с соответствующей характе-



**Рис. 3.** Значение индекса артикуляции гласных в словах детей ТР, с РАС, СД, УО и ССР. По вертикальной оси — значения площадей формантных треугольников гласных, в усл. единицах

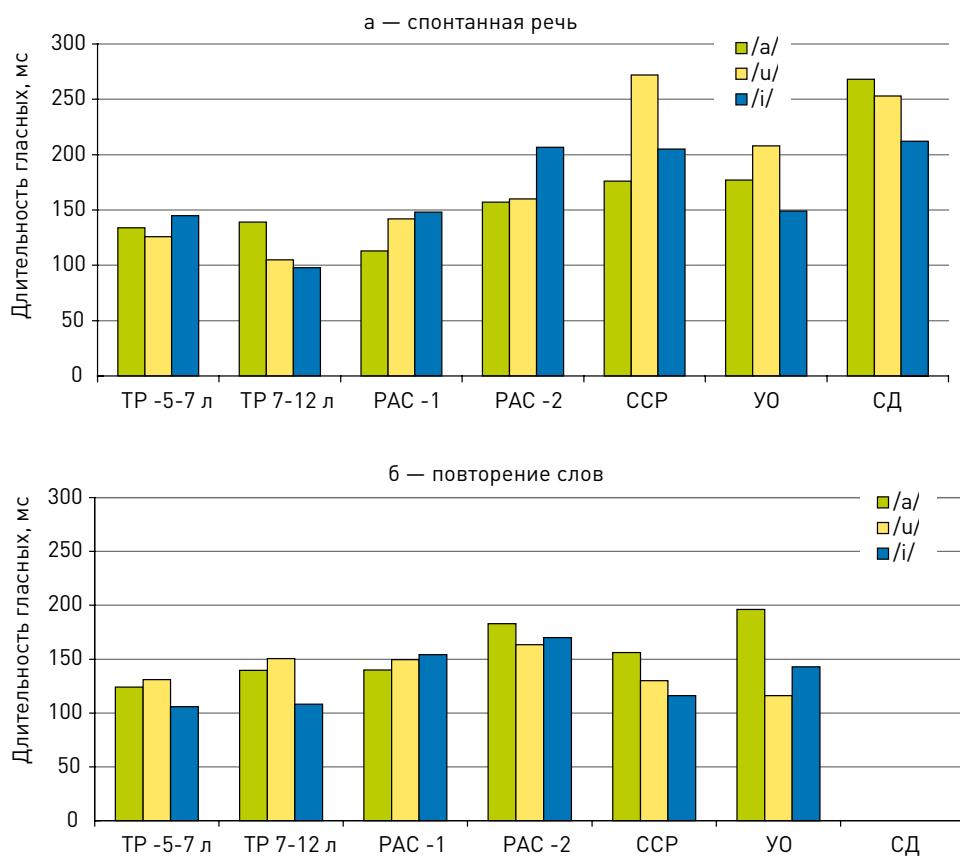


**Рис. 4.** Значения третьей форманты гласных (а) и интенсивности формантных частот гласного /а/, нормированные по отношению к интенсивности ЧОТ (б). На а — по вертикальной оси — значения  $F_3$ , Гц; на б — по горизонтальной оси — значения частот (F1, F2, F3), по вертикальной — значения интенсивностей формант по отношению к значениям интенсивности ЧОТ ( $E_n/E_0$ )

ристикой гласных ТР детей и детей с ССР и СД (рис. 4, а). Интенсивность F3 (E3) гласных, нормированная по отношению к интенсивности ЧОТ, (E3/E0) выше у детей с PAC-1 ( $p<0,001$ ) и PAC-2 ( $p<0,001$ ) при более высоких значениях у детей PAC-1 ( $p<0,005$ ). Значения E3/E0 гласных в словах детей PAC-1 значимо выше ( $p<0,001$ ), чем соответствующие значения у ТР детей, детей с УО и ССР (рис. 4, б). Значения E3/E0 гласных в словах детей PAC-2 значимо выше, чем у ТР детей 7-12 лет и детей с ССР ( $p<0,005$ ), ТР детей 5-7-летнего возраста и УО детей ( $p<0,001$ ). Интенсивность (нормированная) третьей форманты в гласных из слов и речеподобных конструкций детей с СД значимо не отличается от соответствующих значений гласных из слов детей с PAC-2 и меньше ( $p<0,05$ ), чем у детей с PAC-1. Таким образом, интенсивность третьей форманты является одним из специфических признаков, характеризующих речь детей с PAC.

Для детей с PAC баллы по шкале CARS связаны со средними значениями ЧОТ — F(1, 110) = 263,16  $p<0,0000$  ( $\text{Beta} = -0,83977$ ,  $R^2 = 0,70253$ ); значениями ЧОТ на стационарном участке — F(1, 110) = 250,99  $p<0,0000$  ( $\text{Beta} = -0,8338$ ,  $R^2 = 0,69528$ ); значениями F1 — F(3, 107) = 30,882  $p<0,0000$  ( $\text{Beta} = -0,48161$ ,  $R^2 = 0,46404$ ), и F3 ( $\text{Beta} = -0,40721$ ).

Длительность гласных из слов спонтанной речи максимальна у детей с ССР и СД по сравнению с длительностью гласных в словах ТР детей 5-12 лет и детей PAC-1. Длительность ударных гласных /a/, /u/, /i/ в словах детей с PAC-2 выше ( $p<0,001$ ), чем у детей с PAC-1. При повторении слов длительность гласных уменьшается



**Рис. 5. Длительность ударных гласных из слов детей в спонтанной речи (а) и при повторении слов (б)**

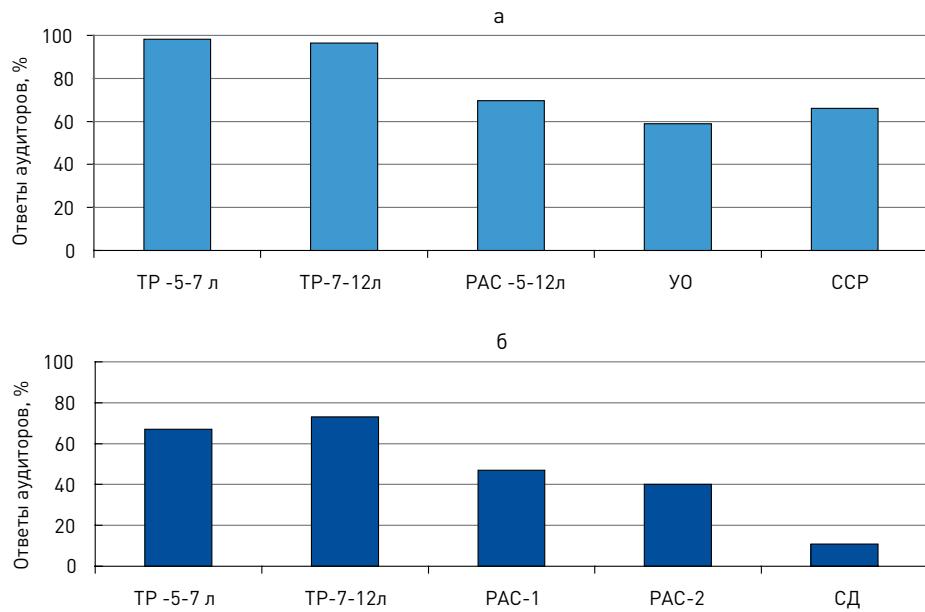


у всех детей по сравнению с спонтанной речью, при больших различиях в значениях у детей ССР и УО (рис. 5, а, б). Длительности ударных гласных в словах ТР и детей РАС-1 значимо не различаются.

Таким образом, сравнительный анализ акустических характеристик речи детей показал, что для детей с РАС характерны высокие значения ЧОТ, высокие значения F3 и E3/E0, характеризующие «атипичную» спектrogramму сигнала. Для детей с РАС, СД, УО и ССР выявлены различия по значениям ЧОТ, индексу артикуляции гласных и по площадям формантных треугольников гласных. Длительность гласных является значимым признаком только для детей с СД, УО и ССР.

Взрослые, носители русского языка, распознают значение повторяемых детьми слов лучше, чем слов из спонтанной речи (рис. 6, а, б). Значение большего количества слов ТР детей взрослые распознают при повторении (98,2 и 96,5% — слов детей 5–7 и 7–12 лет соответственно), чем в спонтанной речи (67 и 73% слов), и лучше, чем слов детей с РАС, УО, ССР и СД. В спонтанной речи детей с СД взрослые распознали 10,8% слов, в 55,5% речевых конструкций правильно определяли гласный, в 33,7% — несколько гласных. Фонетический анализ посредством транскрибирования в системе SAMPA слов и речевых конструкций детей с СД показал наличие всех гласных с частотой встречаемости /a/ — 0,44, /e/ — 0,12, /i/ — 0,155, /o/ — 0,048, /u/ — 0,092, /ɨ/ — 0,116 и редуцированных фонем /U/ — 0,008, /l/ — 0,02, шва — 0,028. Однако у каждого из детей количество гласных различно и составляет от двух до пяти и редуцированные фонемы.

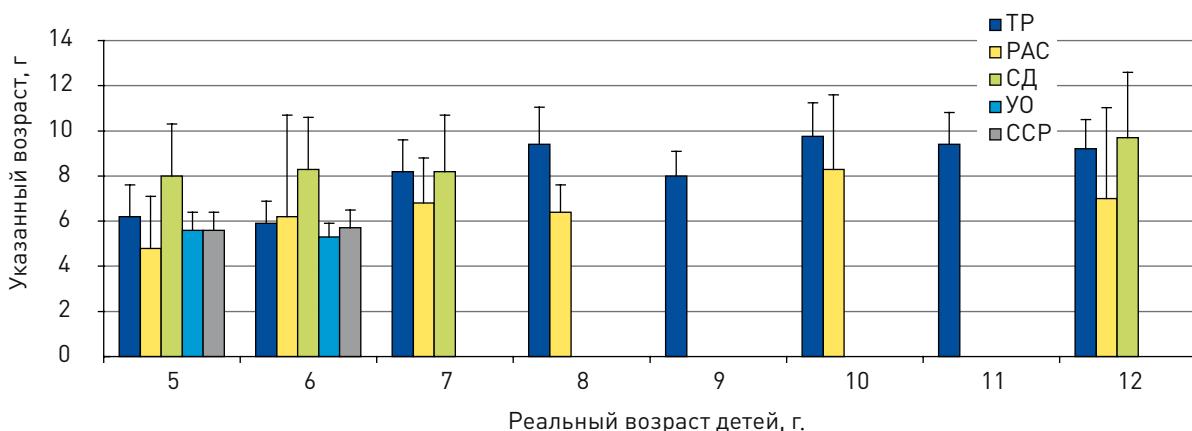
Показана возможность определения взрослыми пола и возраста детей при прослушивании их речевых сигналов. Для данной выборки возраст де-



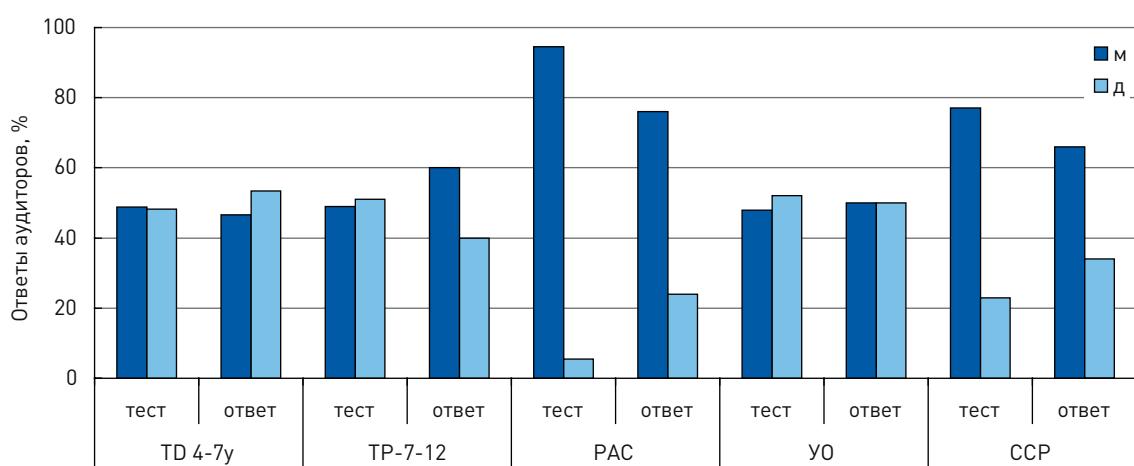
**Рис. 6.** Распознавание взрослыми значения слов детей при повторении (а) и из спонтанной речи (б). По горизонтальной оси — группы детей, по вертикальной — ответы аудиторов, %

тей с РАС аудиторы определяют ниже реального, для детей с СД — выше, детей с УО и CCP — в пределах возрастного диапазона (рис. 7). Предиктором распознавания взрослыми возраста ТР детей являются средние значения ЧОТ:  $F(4, 272) = 4,077$   $p < 0,0000$  ( $\text{Beta} = -0,5712$ ,  $R^2 = 0,043$ ) — мультирегрессионный анализ. Распознавание возраста детей с СД (большего, чем реальный) связано со значениями ЧОТ:  $F(4, 96) = 40,417$   $p < 0,0000$ : ЧОТ по слову ( $\text{Beta} = -1,149$   $R^2 = 0,612$ ), ЧОТ ударного гласного в слове  $p < 0,008$  ( $\text{Beta} = 0,610$   $R^2 = 0,612$ ), диапазоном ЧОТ  $p < 0,03$  ( $\text{Beta} = -1,161$ ). Предиктором для распознавания возраста детей с РАС являются длительность слова —  $F(8, 24) = 27,851$   $p < 0,000$  ( $\text{Beta} = 0,2503$   $R^2 = 0,8703$ ) и минимальные значения ЧОТ по гласному ( $\text{Beta} = -0,769$ ,  $R^2 = 0,8703$ ).

При определении пола ребенка в тестах, содержащих речь ТР в возрасте 7–12 лет, в ответах указывают на большее число мальчиков (60, 49% — ответ, тест соответственно), в тестах РАС указывают на большее число девочек (24, 5,5% — ответ, тест); для тестов, содержащих речь других детей, значимые различия отсутствуют (рис. 8).



**Рис. 7.** Возраст детей, указанный аудиторами при прослушивании речевого материала. По горизонтальной оси обозначен реальный возраст детей, г.



**Рис. 8.** Определение пола детей взрослыми при прослушивании тестовых последовательностей, содержащих их слова



Определение пола ТР ребенка связано со значениями ЧОТ F(5, 271) = 11,2 p<0,0000 (Beta = 0,3081, R<sup>2</sup> = 0,1712), значениями F1 (Beta = -0,2087, p<0,0004), F2 (Beta = 0,2573, p<0,0011), F3 (Beta = 0,1920, p<0,02) — мультирегрессионный анализ.

## ЗАКЛЮЧЕНИЕ

В проведенном исследовании выявлены отличия между детьми ТР и детьми с атипичным развитием по характеристикам голоса и речи. Акустические характеристики речи — значения ЧОТ, значения третьей форманты и ее интенсивность, длительность гласных — в совокупности могут быть использованы в качестве диагностических признаков нарушения развития ребенка. Способность взрослых к определению пола и возраста может быть использована при подборе персонала для работы с детьми с атипичным развитием. Дальнейшие исследования будут направлены на выявление специфики акустических характеристик речи для каждого из нарушений развития ребенка.

Работа выполнена при поддержке РФФИ (№№18-013-01133а, 16-06-00024а),  
РФФИ — огонь (№ 17-06-00503а).

## ЛИТЕРАТУРА

1. Schopler E., Mesibov G. B. Communication problems in autism / E. Schopler, G. B. Mesibov. — New York, US: Plenum Press, 1985. — 335 p.
2. МКБ 10 — Международная классификация болезней 10-го пересмотра (версия: 2016, текущая версия) [Электронный ресурс] // URL: <http://mkb-10.com>
3. Nakai Y., Takashima R., Takiguchi T. et al. Speech intonation in children with autism spectrum disorder // Brain and Development, 2014. — Vol. 36. — № 6. — Pp. 516–522.
4. Bonneh Y.S., Levanon Y., Dean-Pardo O. et al. Abnormal speech spectrum and increased pitch variability in young autistic children // Frontiers in Human Neuroscience, 2011. — Vol. 4. — P. 237. — doi:10.3389/fnhum.2010.00237
5. Sharda M., Subhadra T.P., Sahay S. et al. Sounds of melody — Pitch patterns of speech in autism // Neuroscience Letters, 2010. — Vol. 478. — № 1. — Pp. 42-45
6. Scharfstein L. A., Beidel D. C., Sims V. K. et al. Social skills deficits and vocal characteristics of children with social phobia or Asperger's disorder: A comparative study // Journal of abnormal child psychology, 2011. — Vol. 39. — № 6. — Pp. 865-875
7. Grossman R. B., Bemis R. H., Skwerer D. P. et al. Lexical and affective prosody in children with high-functioning autism // Journal of Speech, Language, and Hearing Research, 2010. — Vol. 53. — № 3. — Pp. 778-793
8. Grossman R. B., Edelson L. R., Tager-Flusberg H. Emotional facial and vocal expressions during story retelling by children and adolescents with high-functioning autism // Journal of speech, language, and hearing research, 2013. — Vol. 56. — № 3. — Pp. 1035-1044.
9. Lyakso E., Frolova O., Grigorev A. A Comparison of Acoustic Features of Speech of Typically Developing Children and Children with Autism Spectrum Disorders // Lecture Notes in Computer Science, 2016. — Vol. 9811. — pp. 43-50. — doi:10.1007/978-3-319-43958-7\_4
10. Lyakso E., Frolova O., Grigorev A. et al. Reflection of the Emotional State in Verbal and Nonverbal Behavioral of Normally Developing Children and Children with

- Autism Spectrum Disorders // Proceedings of the 17th European Conference on Developmental Psychology (September 8-12, 2015, Braga, Portugal). — Medimond Publishing Company, 2016. — Pp. 93-98/
11. Ляксо Е.Е., Фролова О.В., Григорьев А.С. и др. Распознавание взрослыми эмоционального состояния типично развивающихся детей и детей с расстройствами аутистического спектра // Российский физиологический журнал им. И. М. Сеченова, 2016. — Т. 102. — № 6. — С. 729-741.
  12. Lyakso E., Frolova O., Grigorev A. Perception and Acoustic Features of Speech of Children with Autism Spectrum Disorders // Lecture Notes in Artificial Intelligence, 2017. — Vol. 10458. — pp. 602-612. doi:10.1007/978-3-319-66429-3\_60
  13. Kent R.D., Vorperian H.K. Speech Impairment in Down Syndrome: A Review // Journal of Speech, Language, and Hearing Research, 2013. — Vol. 56. — № 1. — Pp. 178-210.
  14. Polišenská K., Kapalková S. Language profiles in children with Down Syndrome and children with Language Impairment: Implications for early intervention // Research in Developmental Disabilities, 2014. — Vol. 35. — № 2. — Pp. 373-382.
  15. Dodd B., Thompson L. Speech disorder in children with Down's syndrome // Journal of Intellectual Disability Research, 2001. — Vol. 45. — № 4. — Pp. 308-316.
  16. Sommers R. K., Reinhart R.W., Sistrunk D.A. Traditional articulation measures of Down syndrome speakers, ages 13-22 // Journal of Childhood Communication Disorders, 1988. — Vol. 12. — № 1. — Pp. 93-108.
  17. Sokol S.B., Fey M.E. Consonant and syllable complexity of toddlers with Down syndrome and mixed-aetiology developmental delays // International journal of speech-language pathology, 2013. — Vol. 15. — № 6. — Pp. 575-585. — URL: <http://dx.doi.org/10.3109/17549507.2013.781676>
  18. Moura C.P., Cunha L.M., Vilarinho H. et al. Voice parameters in children with Down syndrome // Journal of Voice, 2008. — Vol. 22. — № 6. — Pp. 34-42.
  19. Bunton K., Leddy M. An evaluation of articulatory working space area in vowel production of adults with Down syndrome // Clinical linguistics and phonetics, 2011. — Vol. 25. — № 4. — Pp. 321-334. — doi:10.3109/02699206.2010.535647
  20. Белоусова Е.Г., Кубасов А.В. Особенности устной речи детей с аутистическими расстройствами и с выраженной умственной отсталостью // Логопедические технологии в условиях инклюзивного обучения детей с нарушениями речи. Ч. 1. — Екатеринбург, 2013. — С. 40-45.
  21. Schopler E., Reichler R.J., DeVellis R.F. et al. Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS) // Journal of autism and developmental disorders, 1980. — Vol. 10. — № 1. — Pp. 91-103.
  22. Ляксо Е.Е., Григорьев А.С. Динамика длительности и частотных характеристик гласных на протяжении первых семи лет жизни детей // Российский физиологический журнал им. И.М. Сеченова, 2013. — Т. 99. — № 9. — С. 1097-1110.
  23. Roy N., Nissen S.L., Dromey C. et al. Articulatory changes in muscle tension dysphonia: Evidence of vowel space expansion following manual circumlaryngeal therapy // Journal of communication disorders, 2009. — Vol. 42. — № 2. — Pp. 124-135.

## **COMPARATIVE ANALYSIS OF THE VOICE AND SPEECH FEATURES OF CHILDREN TYPICALLY DEVELOPING, WITH AUTISM SPECTRUM DISORDERS, DOWN SYNDROME AND MENTAL RETARDATION**

**Elena E. Lyakso,**  
doctor of biological Sciences, Professor, Professor of the Department of internal Affairs and psychophysiology of the faculty of biology SPBU, head of the group for the study of children's speech



**Olga V. Frolova,**

*candidate of biological Sciences, researcher of the Department of GNI and psychophysiology of biological faculty of St. Petersburg state University, group for the study of child language*

**Alexey S. Grigoriev ,**

*post-graduate student of the Department of internal Affairs and psychophysiology of the faculty of biology SPBU, group for the study of children's speech*

**Victor A. Gorodny ,**

*master's student of the Department of internal Affairs and psychophysiology of biological faculty St. Petersburg state U*

### **Abstract**

The goal of the study is to reveal the specific features of the voice and speech of children with autism spectrum disorders (ASD), Down syndrome (DS), mental retardation of varying severity (MR) compared to typical developing children (TD). The aim of the perceptual study is the review of listeners' (Russian native speakers, adults) recognition of words meaning, child age and gender on the base of speech samples. The spectrographic analysis of temporal and frequency features of child's vocalizations and words was included in speech analysis. The duration of utterances, words, stressed and unstressed vowels and their stationary parts; pitch values, pitch range of utterances and vowels; pitch, formant frequencies and their energy on the stationary part of vowels were estimated. The acoustic features of speech of children with Down syndrome, ASD, and MR vs. TD children were revealed.

**Keywords:** children's speech, spectrographic analysis, articulation index, autism spectrum disorders, Down syndrome, mental retardation

# Speech enhancement in a smartphone-based hearing aid

**Maksim I. Vashkevich,**

*Candidate of technical Sciences, associate Professor of the Belarusian state University of Informatics and Radioelectronics (BSUIR)*

**Iliy S. Azarov,**

*Doctor of technical Sciences, associate Professor, BSUIR*

**Aleksandr A. Petrovsky,**

*Doctor of technical Sciences, Professor of the chair of electronic computing BSUIR*

## Abstract

The paper presents speech enhancement techniques for advanced smartphone-based hearing aid which originates from our free smartphone application "Petralex" recently released for iOS and Android devices. In the present contribution we develop a new solution which overcomes limitations of full-band processing and introduces extended functionality. The new processing scheme decomposes the signal into perceptually matched sliding bands and implements spectral gain shaping for hearing loss compensation, dynamic range compression, noise reduction and acoustic feedback suppression. We propose an acoustic feedback suppression algorithm that is based on spectral subtraction rule. The algorithm is robust to rapid changes in acoustic feedback path and according to experiments allows to achieve added stable gain up to 24 dB. The paper contains theoretical background, description of the implemented techniques and some experimental results.

**Keywords:** hearing aid, noise reduction, acoustic feedback suppression

## INTRODUCTION

Qualitative improvement of hearing aids in the last decade occurred due to increase of computational power of portable devices, their power resources and improvement of analog-to-digital/digital-to-analog converters. There is a miniaturization tendency in hearing aid design which can be noticed in retrospection [1]: pocket hearing aids were superseded by aids inserted in spectacle frame then appeared devices placed behind the ears and now they become small enough to be hidden inside ear's channel. Recently a wide spread of mobile multi-media platforms (especially smartphones) gave new life to pocket hearing aids. A smartphone is capable of functioning as a hearing aid under special software which takes control over audio subsystem of the device. Recently a number of hearing aid applications have been introduced for portable multimedia devices. Although a smartphone cannot be considered as an adequate substitute for a small-sized hearing aid it still might be advantageous for the following reasons [2]:

- functionality of the device can be very flexible regarding both signal processing algorithms and user interfaces;
- large power and computing resource of a smartphone allows implementing sophisticated real-time processing algorithms;



- hearing loss compensation algorithm can be applied to various multimedia content such as music, audio books, movies etc.;
- personal fitting of the hearing aid can be carried out without assistance of audiologist using in situ audiometry;
- it is possible to use different external headsets for different life situations;
- using a smartphone is psychologically comfortable since it is not recognized as a hearing aid by surrounding people;
- for hearing impaired smartphone users there is no need to buy and wear an additional device.

As a pocket hearing aid smartphone has additional advantages:

- a large distance between microphone and speaker prevents occurring of acoustical feedback on considerably high gain levels;
- large physical dimensions of the device can be convenient for persons with con-strained motor function;
- using speakers with bone conduction does not lead to mechanical feedback.

Some time ago we released "Petralex" — a free application for hearing loss compensation with in situ audiometry [3]. The application proved to be helpful and for now it is considered as one of the useful hearing assistive technology<sup>1</sup> [4]. Recently we completed a survey with more than 1500 participants among "Petralex" users that clearly indicated applicability of a smartphone as a self-fitting hearing aid. Compared to conventional hearing aids the application provided the same average change in the hearing ability and turned out to be even more effective in noisy situations.

Considering significant social impact of smartphone-based hearing aids we redesigned "Petralex" in accordance with accumulated user experience. Designing of an original signal processing algorithm is rather difficult considering requirements of the target platform. One of the main problems is processing delay. It has been shown that long processing delays are undesirable due to the comb filter effect, which occurs when the processed sound and the un-processed sound are mixed at the eardrum [5]. It is known that even very short delays (4–8ms) can noticeably reduce sound quality [6].

Although smartphone-based hearing aids are not capable of reaching such short delays due platform limitations. However it is still important to minimize inherent delay of the algorithm. In "Petralex" this problem was solved by using full-band processing scheme [2]. However the scheme has a problem in achieving required sound pressure level which is limited by available dynamic range of the device. Full-band digital amplification leads to clipping effect while applying full-band limiters restrict maximum gain of perceptually important components. Considering that in the preposed solution we implemented a subband processing scheme that processes the sound in narrow frequency bands and controls amplitude each of them individually.

<sup>1</sup> "Petralex" downloaded more than 300 000 users; according to iTunes Connect App Analytics iOS users have more than 1'000 active sessions per day.

It is known that 52 % of hearing impaired people use hearing aids in noisy disturbing situations [7]. Many studies have shown that noise reduction increases hearing comfort and significantly reduces harmful impact on user's hearing [1]. Along with background noise a hearing aid user suffers from acoustic feedback, which occurs when the processed signal leaks from the speaker back to the microphone. In context of smartphone-based solution this becomes a serious problem because the user normally applies standard headphones with heavy sound leakage. Despite that the microphone and speaker are separated by each other acoustic feedback often arises at the desired amplification level. Adaptive feedback cancellation presented in diversity of least mean squares (LMS) techniques is currently the mainstream of acoustic feedback cancellation in hearing aids [8–12]. However practical experience shows that this approach is ineffective for smartphone implementation. When using smartphone as a hearing aid the feedback path is very unstable because of changing distance between microphone and speakers. In such conditions adaptive filtering cannot noticeably improve maximum stable gain: when using low adaptation rates the reaction to changes becomes unpredictable, when using sufficiently high adaptation rates the speech signal drastically degrades. It was shown that the room acoustic also makes a considerable contribution to feedback path [13], however robust modeling of room acoustic by means of adaptive filtering can hardly be done in real-life environment. Another problem is robustness: adaptive filtering can be applied only when the system is stable and once stability is lost it cannot be recovered. A known approach for robust feedback control is notch-filtering howling suppression [14–17] which is able to stabilize a system without reducing the broadband gain. The approach is suitable for a smartphone; however its weak side compared to adaptive feedback cancellation is a low maximum stable gain increase and signal distortion [18]. Considering that we propose an original algorithm based on spectral subtraction instead of notch-filtering. The algorithm applies a weighting rule derived specially for feedback and can be combined with noise reduction which attenuates both background noise and feedback residual. According to experimental results the proposed solution provides close performance to adaptive feedback cancellation in terms of maximum stable gain increase and speech quality, however and at the same time is very robust against changes in feedback path. Combination of noise reduction implies that both algorithms share the same analysis/synthesis framework which is advantageous regarding computational efficiency.

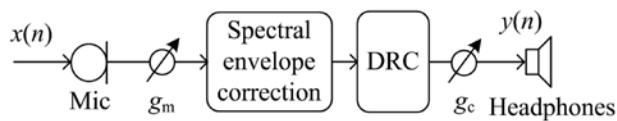
## **1. IMPLEMENTED PROCESSING SCHEME**

In modern hearing aids, signal processing is usually performed in frequency subbands introducing analysis-synthesis delays in the forward path. Many research efforts have been focused on this problem [19–20], however the delays of these solutions are still high (6–8 ms). Good low-delay filtering schemes based on peaking filters [6], cochlear filters [21] and side-branch processing [22] has been recently proposed. Some common frequency-dependent amplification schemes are shortly described below. Existing mobile platforms can process the signal in real-time by separate frames of 6 ms or longer, that requires block by block processing. It is impossible to eliminate delays introduced by analog-to-digital and digital-to-analog converters which can reach 0,4 to 2 ms depending on implementation [23]. Inherent hardware delay of a smartphone is much longer due to implementation of audio processing pipeline (10–20 ms for iPhone and 50–300 ms for Android).



### 1.1 Full-band processing

Considering constraints of the mobile platform it is possible to use full-band processing scheme that uses finite impulse response (FIR) filtering and dynamic range compression (DRC) for hearing loss compensation. The scheme is shown in Figure 1.



**Figure 1.** Full-band processing scheme

Spectral envelope correction is done using FIR filter which is designed using prescription gain formulas. There are two loudness controls: microphone sensitivity  $g_m$  and output level  $g_c$ , which the user can adjust according to the current acoustic conditions. The block of dynamic range compression applies time-varying gain for recruitment correction. Compression ratio is chosen according to the degree of hearing loss. Considering that smartphone uses a stereo headset it is possible to apply binaural hearing compensation processing left and right channels separately. In the previous version of "Petralex" we applied linear phase filter with group delay  $\approx 3$  ms, which is synthesized using the windowing method.

The full-band processing scheme has the following advantages: low processing delay (which consists of the group delay of the equalizer filter and platform delay), low computational cost and simplicity in design. However the scheme is not capable of controlling loudness of separate spectral components which requires time-frequency transform of the signal.

### 1.2 Sub-band processing

Functionality of the hearing aid can be significantly extended using sub-band decomposition of the signal into separate frequency components. Processing in this case can be carried out using individual time-varying amplification of each subband channel [6].

There are sub-band processing systems with reduced processing delay. In [22, 24] a scheme of sub-band amplification is proposed that does not require synthesis filter. Processing in the forward path is carried out using FIR filter, coefficients of which are updated for each processing frame according to amplifications gains derived from subband side branch. It is also possible to use parametric band-pass filters [6] or cochlear filters [21] summing outputs of the filters after amplification. Both cochlear filter bank and peaking filters decompose the signal into perceptually matched subband components and has a very low group delay. However these approaches are computationally more consuming compared to general sub-band processing scheme.

### 1.3 Proposed processing scheme

We assume that the input signal can be represented in the frequency domain as a sum of clean speech signal  $X(w)$ , acoustic feedback  $A(w)$  and background noise  $N(w)$ :

$$\bar{X}(w) = X(w) + A(w) + N(w) = \bar{X}(w) + N(w), \quad (1)$$

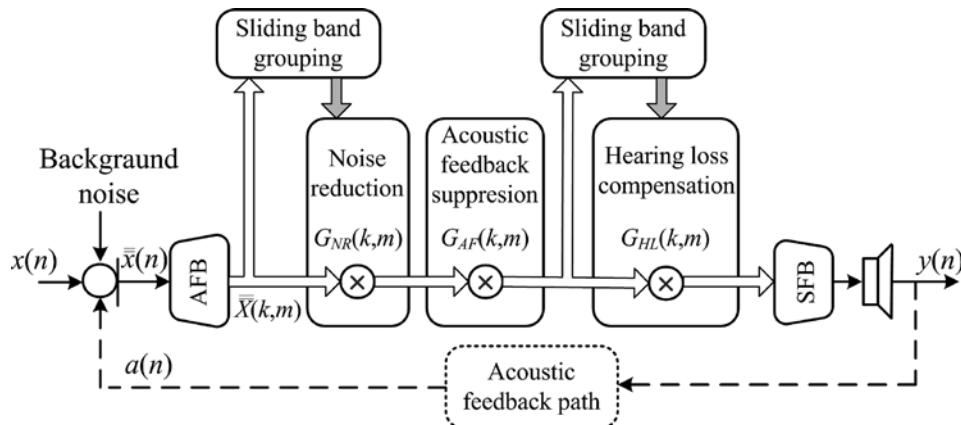
where  $\bar{X}(w) = X(w) + A(w)$  is the speech signal with acoustic feedback. Let  $R_{\bar{X}}(w)$ ,  $R_{\bar{X}}(w)$ ,  $R_n(w)$  are power spectral densities (PSD) of  $X(w)$ ,  $A(w)$  and  $N(w)$  respectively, then  $\bar{X}(w)$  can be estimated from  $\bar{X}(w)$  by using noise reduction factor

$$G_{NR}(\omega) = \sqrt{1 - \frac{R_n(\omega)}{R_{\bar{X}}(\omega)}}. \quad (2)$$

Feedback suppression factor can be estimated in the same way:

$$G_{AF}(\omega) = \sqrt{1 - \frac{R_a(\omega)}{R_{\bar{X}}(\omega)}}. \quad (3)$$

On the basis of described approach the following processing scheme is proposed (Figure 2). The signal processing includes three consecutive stages: 1) noise reduction; 2) acoustic feedback suppression and 3) hearing loss compensation.



**Figure 2.** Implemented processing scheme

The input signal  $\bar{x}(n)$  is decomposed into complex subbands  $\bar{X}(k,m)$  by the analysis filter bank (AFB), where  $k$  and  $m$  are frequency and time indices respectively, and the processed full-band signal  $y(n)$  is reconstructed by synthesis filter bank (SFB). For reasons of computational efficiency we use an oversampled DFT-modulated filter bank. Calculation of noise reduction coefficients  $G_{NR}(k,m)$  requires estimation of noise PSD. In order to make noise statistics more reliable subband signals are combined in a wide sliding bands. Acoustic feedback suppression coefficients  $G_{AF}(k,m)$  is calculated based on estimation of acoustic feedback signal PSD. At the last stage subband signals multiplied by the  $G_{NR}(k,m)$  and  $G_{AF}(k,m)$  are combined into sliding bands for determining required hearing compensations gains  $G_{HL}(k,m)$  which are calculated using to a desired prescription formula and DRC pro-file.



## 2. ANALYSIS-SYNTHESIS BASED ON DFT-MODULATED FILTER BANK

### 2.1. Analysis-synthesis framework

Filter banks are commonly used tool to organize subband signal processing in modern hearing instruments [6, 19–22, 24]. A DFT (or complex) — modulated filter bank with poly-phase implementation of FIR prototype filter is one of the most efficient and popular [19, 24]. For example in [25] for hearing aid system was used an oversampled, polyphase DFT filter bank with 16 frequency bands. Decimation of the subband signals reduces computational cost, however decimation/interpolation in this solution inevitably distorts the output signal. A well-known techniques such as aliasing compensation that used in perfect (near perfect) reconstruction filter banks are not suitable for hearing aid since gains applied to subbands are significantly different. For this reason a special procedure for FIR prototype design should be used [19].

Another different form for implementing DFT filter bank is weighted overlap-add (WOLA) structure [20, 26]. WOLA structure is more general than polyphase structure in which number of channels and decimation factor have the following restriction

$$K = MI. \quad (4)$$

where  $I$  — a positive integer ( $I = 1, 2, 3, \dots$ ) called oversampling ratio. In WOLA is unrelated to  $K$ .

The output signal for  $k$ -th channel of analysis DFT filter bank is expressed as follows

$$X(k, m) = \sum_{n=-\infty}^{\infty} h(mM - n)x(n)W_K^{-kn}, \quad k = 0, 1, \dots, K - 1. \quad (5)$$

where  $W_K = e^{j2\pi/K}$ ,  $x(n)$  — input signal,  $h(n)$  — filter prototype (of length  $L$ ) that is a sliding analysis window that selects and weights a frame of the input signal. Output signals  $X_k(m)$  are referred to as the short-time spectrum of the signal at time  $n = mM$ . Expression (5) can be rewritten as DFT

$$X(k, m) = \sum_{n=-\infty}^{\infty} y(n)W_K^{-kn} \quad (6)$$

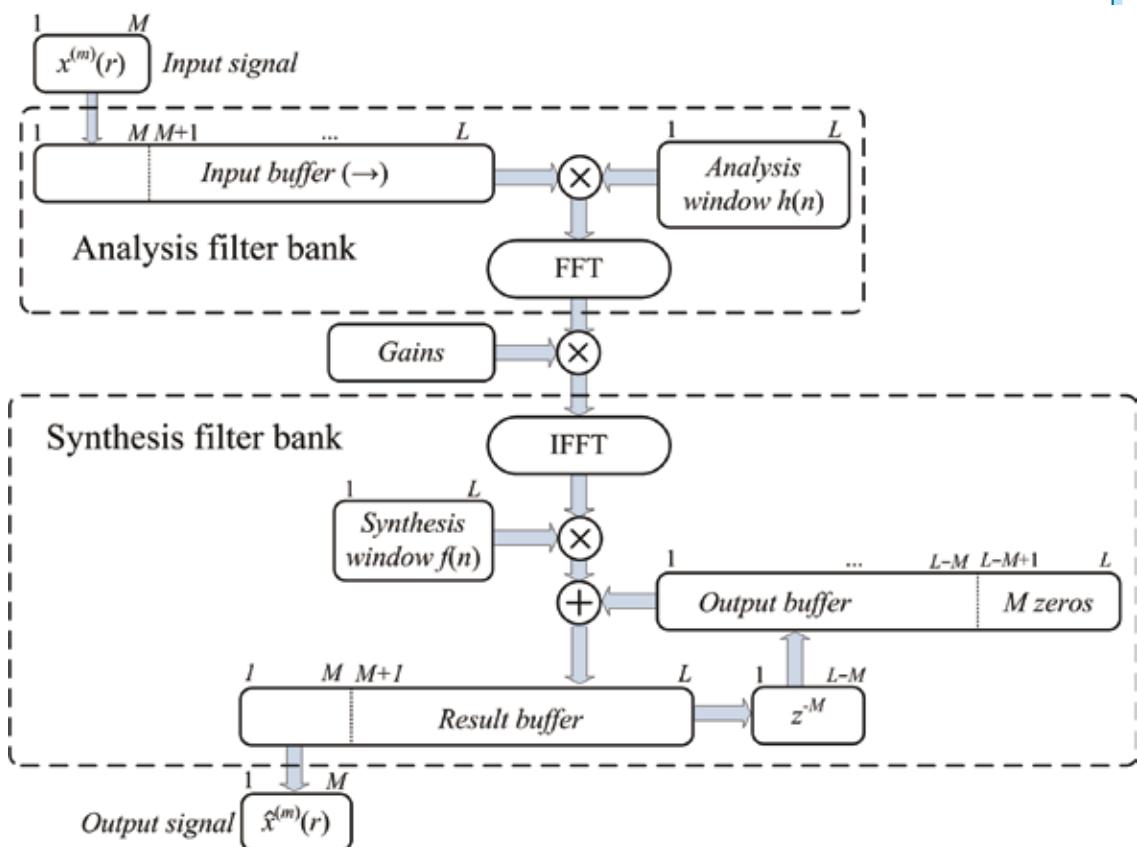
where  $y(n) = h(mM - n)x(n)$  — windowed input sequence.

The WOLA synthesis structure can be expressed in the form

$$\hat{x}(n) = \sum_{m=-\infty}^{\infty} f(mM - n) \frac{1}{K} \sum_{k=0}^{K-1} \hat{X}(k, m)W_K^{kn} \quad (7)$$

where  $f(n)$  — synthesis filter (or synthesis window). Simplified structure of WOLA filter bank (when length of analysis and synthesis windows  $L$  equals to  $K$ ) is given in Figure 3 where the following notation is used

$$x^{(m)}[r] = x(mM + r), \quad r = 0, 1, \dots, M - 1. \quad (8)$$



**Figure 3.** WOLA structure of DFT-modulated filter bank

We use a simple method of calculation  $h(n)$  and  $f(n)$  that allows to obtain good frequency resolution and low aliasing of the reconstructed signal. For analysis we use the Hamming window, which provides good trade-off between main-lobe width and side-lobes attenuation in short-time spectrum:

$$h(n) = h_{\text{hamm}}(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{L-1}\right). \quad (9)$$

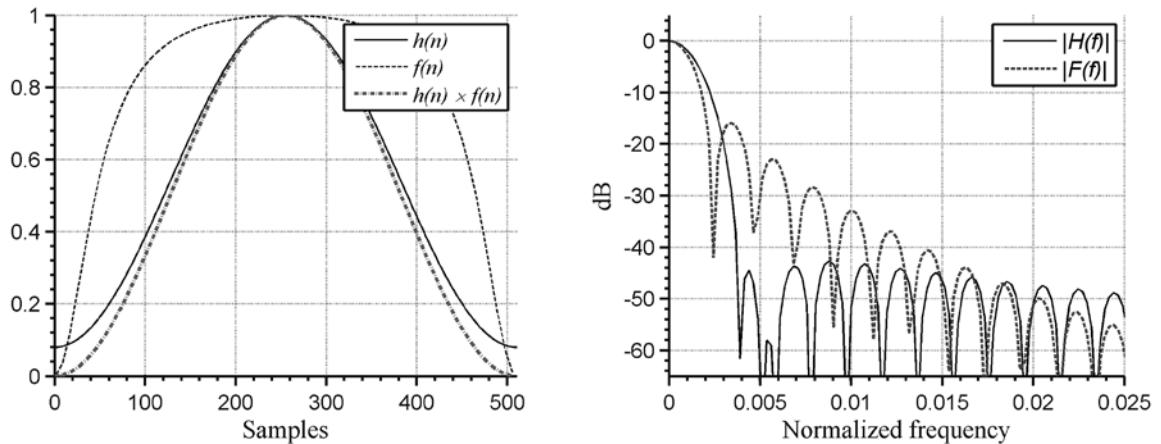
where  $n = 0 \dots L-1$ . Assuming that  $L$  and  $M$  are odd the synthesis window is defined as

$$f(n) = \begin{cases} \frac{h_{\text{hamm}}(n)}{h_{\text{hamm}}\left(n + \frac{L-1}{2} - M\right)} & \frac{0.5 - 0.5 \cos\left(\frac{2\pi n}{2M}\right)}{0.54 - 0.46 \cos\left(\frac{\pi(2n+L-1-2M)}{L-1}\right)}, \text{ for } n \in [0; 2M] \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

where numerator is the Hanning window of length  $2M+1$ . The synthesis window attenuates phase breaking effect between adjacent frames. According to (10) applying both windows  $h(n)$  and  $f(n)$  is equivalent to applying Hanning window that ensures perfect concatenation of the reconstructed signal since summation of two shifted version of Hanning windows gives one

$$h_{\text{hann}}(n) + h_{\text{hann}}(n+M) = 1. \quad (11)$$

Figure 4 shows windows calculated according to (9)–(10) that guarantee perfect signal reconstruction.



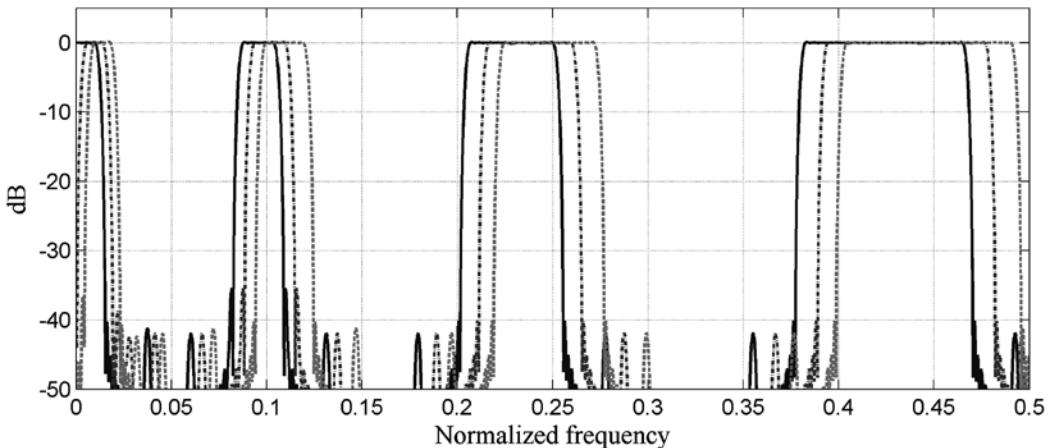
**Figure 4.** Impulse and magnitude frequency responses of  $h(n)$  and  $f(n$ ) ( $L = 511$  and  $M = 255$ )

## 2.2 Sliding band grouping

Using overlapping frequency bands is advantageous for estimating level of the sub-band signals [27] since it prevents from decrease of the spectral contrasts and modulation depths in speech signal. Subbands signals  $|X(k,m)|^2$  are grouped in sliding bands both in time and frequency domains

$$S(k, m) = \frac{1}{N_t} \sqrt{\sum_{i=-bw(k)}^{bw(k)} \sum_{j=0}^{N_t-1} |X(k+i, m-j)|^2}. \quad (12)$$

where  $bw(k)$  determines frequency bandwidth,  $N_t$  — number of summed frames. Following the psychoacoustic principle that the bandwidth is proportional to the central frequency  $bw(k)$  is calculated as



**Figure 5.** Overlapping frequency bands for spectral energy estimation (filter bank parameters:  $L = 511$ ,  $M = 40$ )

$$bw(k) = \max \left\{ \text{round} \left( \frac{k}{20} \right), bw_{\min} \right\}. \quad (13)$$

where  $bw$  is minimum bandwidth. Using such grouping reduces effect of musical artifacts in noise reduction algorithm since fragments of residual are not perceived as tones.

An illustration of using overlapping frequency bands is given in Figure 5, where partition of the frequency range into sliding bands is shown.

### 3. NOISE REDUCTION

Implemented noise estimation algorithm is generally based on the minima controlled recursive averaging (MCRA) [28], where noise spectrum is obtained by averaging past spectral values  $S(k,m)$ :

$$\tilde{D}(k,m) = \tilde{\alpha}_d(k,m) \tilde{D}(k,m-1) + (1 - \tilde{\alpha}_d(k,m)) S(k,m). \quad (14)$$

where

$$\tilde{\alpha}_d(k,m) = \alpha_d + (1 - \alpha_d)p(k,m) \quad (15)$$

is a time-varying smoothing parameters that depends on estimation of conditional signal presence probability  $p(k,m)$ . In (15)  $\alpha_d$  ( $0 < \alpha_d < 1$ ) is a smoothing parameter determines averaging time in the absence of speech.  $p(k,m)$  is obtained by recursive averaging

$$p(k,m) = \alpha_p p(k,m-1) + (1 - \alpha_p) I(k,m)$$

where

$$I(k,m) = \begin{cases} 1 & \text{if } S_r(k,m) > \delta, \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

denotes function that indicates the presence of speech component. In (16) the hard decision is done based on the ratio between the local noisy speech spectrum estimation and its derived minimum  $S_{\min}(k,m)$ :

$$S_r(k,m) \triangleq \frac{S(k,m)}{S_{\min}(k,m)}. \quad (17)$$

$S_{\min}(k,m)$  is updated using temporary minimum  $S_{\text{tmp}}(k,m)$ . Initially  $S_{\min}(k,0)$  and  $S_{\text{tmp}}(k,0)$  are set to  $S(k,0)$  and for each frame they are updated in following way

$$S_{\min}(k,m) = \min\{S_{\min}(k,m-1), S(k,m)\}. \quad (18)$$

$$S_{\text{tmp}}(k,m) = \min\{S_{\text{tmp}}(k,m-1), S(k,m)\}. \quad (19)$$

Every  $L$  frames the following update rule is used

$$S_{\min}(k,m) = \min\{S_{\text{tmp}}(k,m-1), S(k,m)\}. \quad (20)$$

$$S_{\text{tmp}}(k,m) = S(k,m) \quad (21)$$



Parameter  $L$  determines the resolution of the local minima search. The local minimum is searched on a window of at least  $L$  frames, but not more than  $2L$  frames. A good result is obtained for window duration of 0.5–1.5 s [28].

Spectral gain  $G_{NR}(k,m)$  is determined as

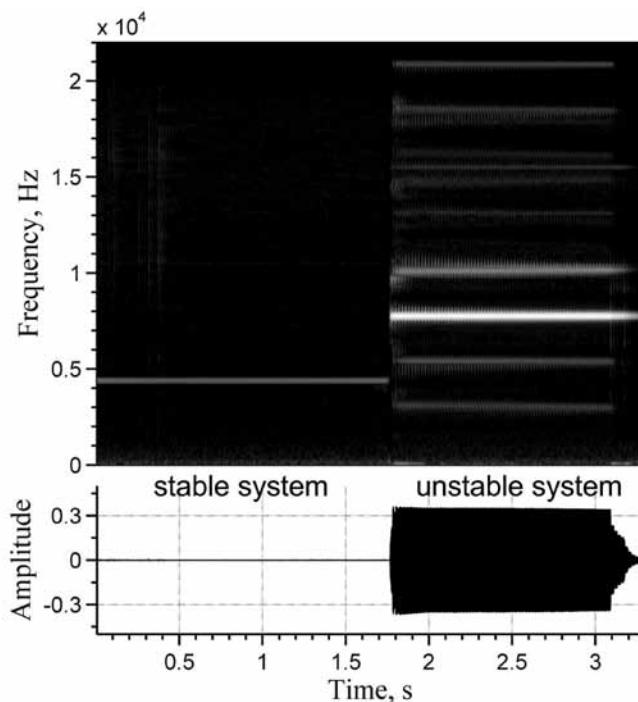
$$G_{NR}(k,m) = \max \left\{ \sqrt{\frac{S^2(k,m) - v\tilde{D}^2(k,m)}{S^2(k,m)}}, 10^{-RL/20} \right\}. \quad (22)$$

where  $v$  — subtraction factor ( $1 < v < 6$ ),  $RL$  — adjustable parameter that defines the desired residual noise level in dB.

#### 4. ACOUSTIC FEEDBACK SUPPRESSION

When the system is stable the feedback loop can be considered as linear and the feed-back signal typically occurs as a single sine wave. However when the system is unstable the loop becomes non-linear and feedback signal appears as a number of harmonics with unstable parameters as shown in Figure 6.

In both cases acoustic feedback occurs as a quasi periodical signal which is generated by recursive summation of the output with a time offset  $t_o$ . Periodicity of acoustic feedback ensures that its spectral components



**Figure 6.** Acoustic feedback in quiet conditions (recorded on iPhone using built-in microphone and standard headphones)

are spaced in frequency domain by fundamental frequency  $f_0 = 1 / t_0$  and therefore feedback affects only a subset of  $X(k, m)$ :

$$\bar{X}(k, m) = \begin{cases} X(k, m) + A(k, m), & \min_v \left| k - \frac{f_0 v}{f_s/K} \right| \leq d \\ X(k, m), & \min_v \left| k - \frac{f_0 v}{f_s/K} \right| > d \end{cases} \quad (23)$$

where  $v$  — number of feedback harmonics,  $f_s$  — sampling frequency and  $d$  — frequency offset, specified by the main lobe of frequency response of the analysis window.

For clean speech expected value of spectral amplitude can be roughly estimated from adjacent frequency components as  $E[|X(k, m)|] \approx E[|X(k \pm d, m)|]$  for any sufficiently small frequency offset  $d$ . Using (23) and assuming that acoustic feedback increases mean spectral amplitude i.e.  $E[|\bar{X}(k, m)|] > E[|X(k, m)|]$  we get

$$E[|X(k, m)|] \approx \min_{-d < i < d} [E[|\bar{X}(k \pm i, m)|]]. \quad (24)$$

$$E[|A(k, m)|] \approx E[|\bar{X}(k, m)|] - \min_{-d < i < d} [E[|\bar{X}_{k+i}(m)|]] \quad (25)$$

According to (23) and (24) expected feedback gain  $E[|\bar{X}(k, m)|] / E[|X(k, m)|]$  can be estimated from short-time spectral amplitudes close to the corresponding sample  $\bar{X}(k, m)$ . We introduce the following measure of feedback gain  $\chi(k, m)$  based on  $l$  previous frames and  $2d$  neighboring frequency bins:

$$\chi(k, m) \triangleq \frac{\min_{-l+1 \leq j \leq 0} |\bar{X}(k, m+j)|}{\min_{-d \leq i \leq d} \left[ \max_{-l+1 \leq j \leq 0} |\bar{X}(k+i, m+j)| \right]} \quad (26)$$

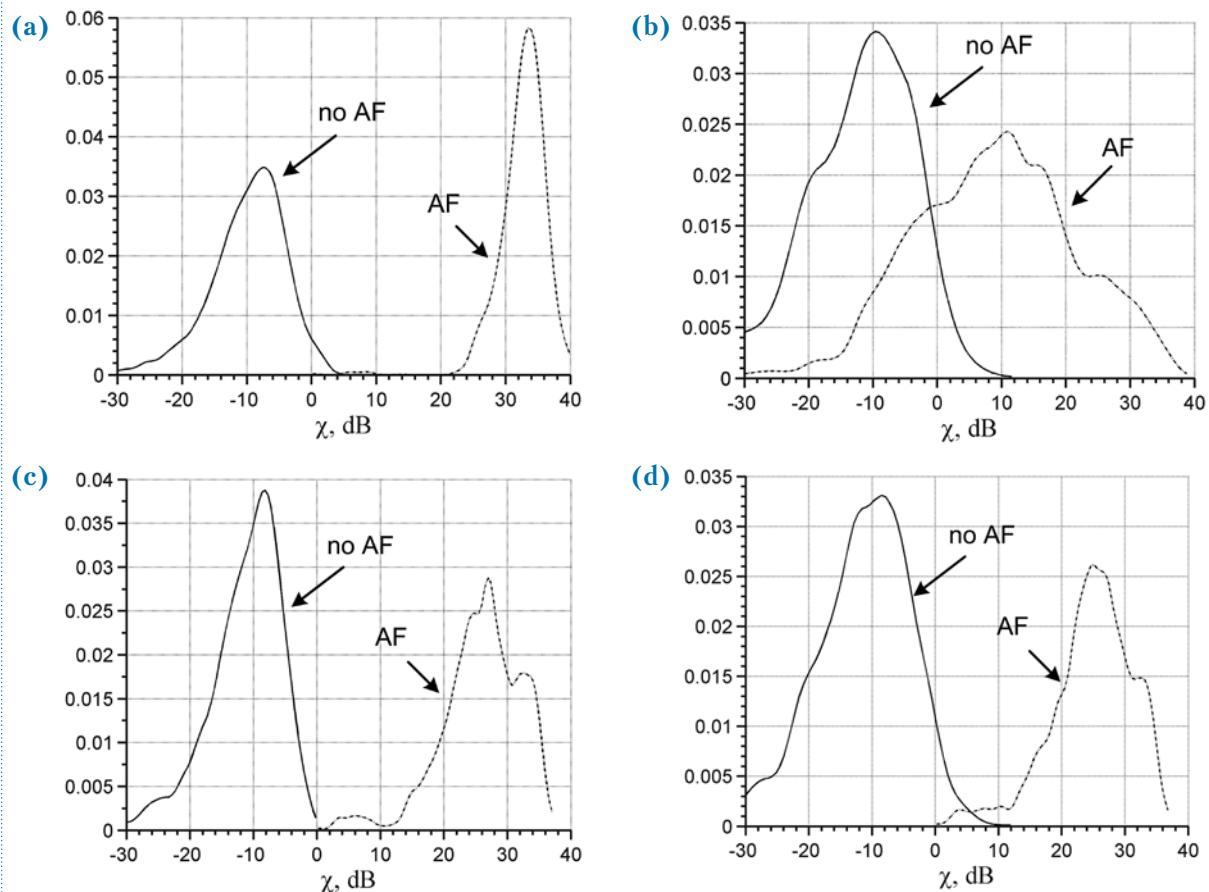
In order to avoid overrating of feedback level we use local minima over previous time samples for estimating  $E[\bar{X}(k, m)]$  and local maxima for estimating  $E[X(k, m)]$ . Figure 7 shows probability density function  $\rho(\chi)$  obtained experimentally in quiet conditions and during loud speaking for different signal-to-feedback ratios. According to experimental data  $\chi > 1$  indicates that acoustic feedback is present with probability around 95%. Feedback is not detected when the speech signal is very loud compared to feedback.

Feedback is smoothly controlled, using fixed smoothing value  $\alpha_{AF}$  ( $0 < \alpha_{AF} < 1$ ) that specifies averaging time and a time-varying smoothing parameter  $\tilde{\alpha}_{AF}$  that depends on feedback intensity

$$\tilde{\alpha}_{AF} = \alpha_{AF} + (1 - \alpha_{AF}) \left( \frac{\chi(k, m)}{\chi_{th}} \right)^\beta \quad (27)$$

where  $\beta$  is equalizing parameter that balances reaction to quiet and loud feedback and  $\chi_{th}$  is threshold value for hard decision. Suppression gain  $G_{AF}(k, m)$  is updated using the following expression:

$$G_{AF}(k, m) = \begin{cases} \tilde{\alpha}_{AF} G_{AF}(k, m-1) + \frac{(1 - \tilde{\alpha}_{AF})}{\max(\chi(k, m), 1)}, & \chi(k, m) < \chi_{th} \\ 1/\chi(k, m), & \chi(k, m) \geq \chi_{th} \end{cases} \quad (28)$$



**Figure 7.** Probability density functions  $\rho(\chi)$  for acoustic feedback presence and absence in different conditions: **(a)** stable system in quiet (signal-to-feedback ratio  $-7\text{dB}$ ), **(b)** stable system, loud speaking (signal-to-feedback ratio  $35\text{dB}$ ), **(c)** unstable system in quiet (signal-to-feedback ratio  $-38\text{dB}$ ), **(d)** unstable system, loud speaking (signal-to-feedback ratio  $-7\text{dB}$ )

When  $\chi(k, m)$  exceeds  $\chi_{th}$  an intense acoustic feedback is detected. In this case suppression gain is updated instantaneously and then slowly released.

## 5. HEARING LOSS COMPENSATION

In order to determine personal target amplification gains we use in situ audiology. Hearing threshold levels in quiet are measured using increasing tonal sounds with frequencies 125, 250, 500, 1000, 2000, 4000 and 8000 Hz. We calculate required amplification using conventional formulas: Berger [29], POGO (Prescription of gain and output) [30] and NAL-R (National Acoustic Laboratories, Australia) [31]. Correspondent calculation rules were implemented as described in [1]. Recruitment correction is implemented using subband compressors with compression ratio derived from hearing loss profile. For each channel the hearing loss compensation gain is calculated according to a given prescription formula and current compression gain.

## 6. EXPERIMENTAL RESULTS

### 6.1 Design aspects and implementation

According to Figure 2 the processing of incoming block of samples is carried out using the following steps:

1. New  $M$  samples block moved to input buffer, multiplied by analysis window  $h(n)$  followed by FFT (see Figure 3 and eq. (5)) to obtain  $\bar{X}(k, m)$ ;
2. In block "Sliding band grouping" (Figure 5)  $S(k, m)$  is calculated using eq. (12);
3. The estimated spectrum  $S(k, m)$  passed to "Noise reduction" block, where coefficients  $G_{NR}(k, m)$  calculated using eq. (14)–(22);
4. Modified filter bank outputs  $\bar{X}(k, m)G_{NR}(k, m)$  are passes to "acoustic feedback suppression" block, where  $G_{AF}(k, m)$  are calculated using eq. (26)–(28);
5. Modified filter bank outputs  $\bar{X}(k, m)G_{NR}(k, m)G_{AF}(k, m)$  are passes to "Sliding band grouping" block, where smoothed estimation of clean speech spectrum is obtained;
6. Based on estimation of clean speech spectrum obtained in step 5, prescription gains and DRC settings hearing loss compensation coefficients  $G_{HL}(k, m)$  are calculated in hearing loss compensation block. Filter bank outputs modified as

$$\hat{X}(k, m) = X(k, m)G_{NR}(k, m)G_{AF}(k, m)G_{HL}(k, m);$$

7. Subband signals  $\hat{X}(k, m)$  are sent to synthesis filter bank, where output block of sample of length  $M$  is obtained (see Figure 3).

The proposed signal processing system was implemented and tested using iPhone-5s and personal computer. The sampling frequency is 44.1 kHz and frame size  $L = 511$ , signal is captured by blocks of  $M = 255$  (50% overlap) that corresponds to 5.8 ms processing delay. In order to apply efficient radix-2 FFT we used zero padding. The program was written using combination of C++ and objective C languages, with Apple's IDE "Xcode, ver.8.2". The processing algorithm insignificantly reduces discharge time of a smartphone (on iPhone-5s the algorithm can continuously work more than 24 hours).

### 6.2 Noise reduction

Three types of different noises were added to create noisy signals with segmental SNRs in range  $[-5, 10]$  dB. The segmental signal-to-noise ratio (SSNR) is defined by [32]

$$SSNR = \frac{10}{|M|} \sum_{m \in M} \log \frac{\sum_{k=0}^{K/2} |X_k(m)|^2}{\sum_{k=0}^{K/2} |D_k(m)|^2}, \quad (29)$$

where  $M$  represents the set of frames that contain speech and  $|M|$  its cardinality.

Five male and five female speech samples of duration over 40 s were used in the experiment. The following parameters of noise reduction algorithm were used:  $bw_{min} = 3$ ,



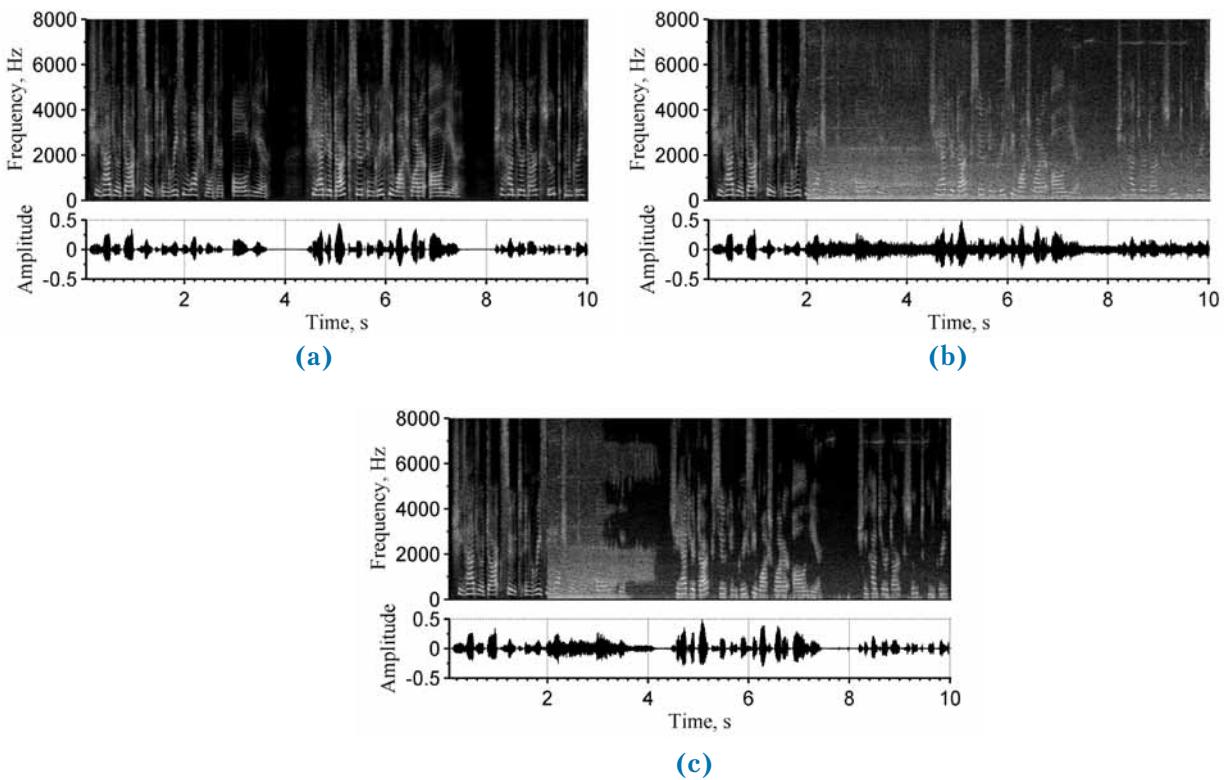
$\alpha_d = 0,95$ ,  $\alpha_p = 0,2$ ,  $L = 172$  {minimum search window is about 1 s},  $\delta \sqrt{5}$ ,  $v = 2$ ,  $RL = 9$ ,  $N_f = 4$ . Table 2 shows the average SSNR improvement obtained for different noises.

**Table 2**  
**Segmental SNR improvement for various noise types and levels**

Input SegSNR, dB	White noise	Cafeteria noise	Traffic noise
-5	9,81	6,04	6,64
0	7,87	4,03	4,43
5	6,13	2,36	2,68
10	4,59	0,91	1,31

Figure 8 shows the response of the algorithm to rapid change of noise intensity.

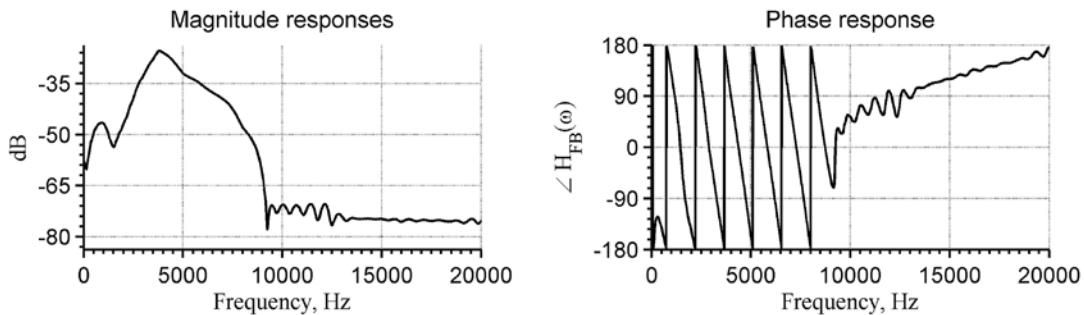
Figure 8 (b) shows a speech signal corrupted by the traffic noise which starts at 2 s. The algorithm reacts in less than 2 s after appearance of noise.



**Figure 8.** Performance of the noise reduction algorithm: (a) clean speech; (b) noisy speech signal (noise appears at 2 s); (c) processed noisy speech

### 6.3. Acoustic feedback suppression

The proposed combined noise and acoustic feedback reduction algorithm was evaluated using a feedback path model similar to [10]. Feedback path was modeled as a FIR filter with 279 coefficients, frequency response of the filter is shown in Figure 9. The hearing loss compensation gain was constant for all subbands. The following parameters of the feed-back suppression algorithm were used:  $\alpha_{AF} = 0,997$ ,  $\beta = 0,15$  and  $X_{th} = 10$ .



**Figure 9.** Frequency response of the acoustic feedback path

At first the maximum stable gain of the system was determined [9] that can be applied to signal without feedback control. Then we evaluated performance of the system at different added stable gains  $\Delta G$  using the proposed feedback suppression algorithm and the LMS adaptive filtering algorithm (279 coefficients) [8]. Table 3 shows the obtained SSNR values obtained in the experiment ('US' means unstable system). The noise signal was obtained as the difference between output signal (with feedback loop and suppression) and the output signal in ideal conditions (without feedback loop and without suppression).

**Table 3**  
**SSNR for different added stable gains  $\Delta G$**

$\Delta G, \text{dB}$	No AFR, dB	LMS, dB	Proposed AFR, dB
0	8,12	17,66	12,27
4	US	5,57	11,94
8	US	5,35	11,10
12	US	3,18	10,22
16	US	1,72	9,25
20	US	US	7,59
24	US	US	4,56

The proposed suppression algorithm provided much higher SSNR than LMS in all cases and keeps the system stable even at high added stable gain of 24 dB.

We also evaluated performance of combined feedback and noise suppression in noisy conditions. A speech signal mixed with pink noise at different SSNR levels was used as input to the system with  $\Delta G = 12\text{dB}$ . Table 4 presents obtained SSNR measures.

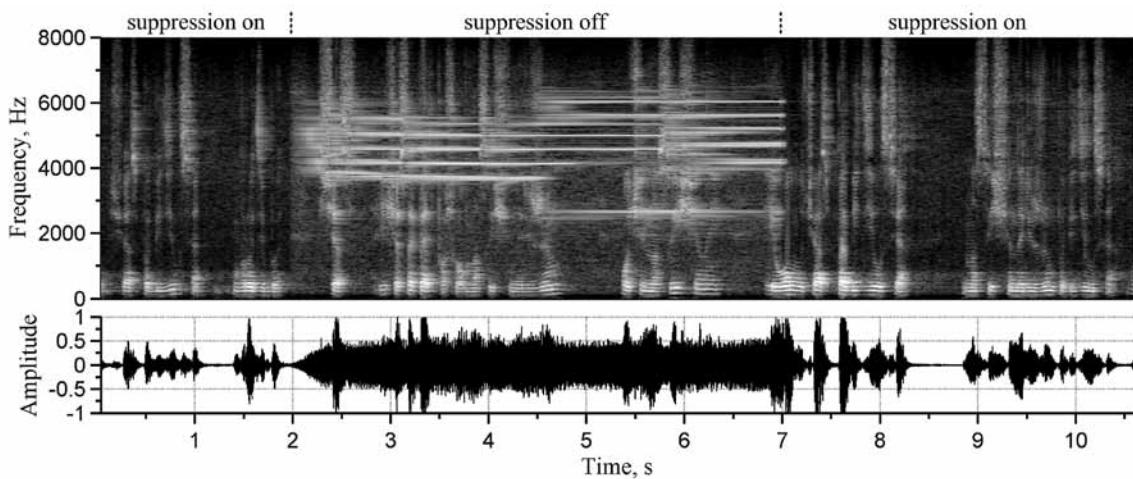


Table 4  
**SSNR in noisy condition, 12dB**

Input , dB	Proposed feedback suppression, dB	Proposed feedback suppression and noise reduction, dB	LMS, dB	LMS and noise reduction, dB
15	10,54	11,29	5,83	7,78
10	7,07	9,34	4,53	5,62
5	3,51	6,57	2,31	5,00
0	-0,73	3,12	-1,17	2,42
-5	-5,41	-1,24	-5,01	-0,93

The implemented noise reduction algorithm considerably improves SSNR, suppressing both feedback residual and background noise. Suppression of feedback residual significantly improves subjective perception of the processed speech, removing audible tonal components.

In order to evaluate performance of the algorithm in real-life environment we used a PC-based real-time mockup and standard multimedia headset with large headphones. The mockup was placed in a big reverberant room. During the test we changed orientation and location of the headset in order to model time-varying feedback path. The proposed algorithm showed similar performance to previous modeling experiments and never became unstable. An example of performance of the proposed algorithm is given in Figure 10. When feedback suppression is off the system quickly becomes unstable and feedback emerges as multiple tonal components,



**Figure 10.** Output signal of real-time mockup: feedback suppression is turned-on at 2 s (distance between microphone and headphones is approximately 30 cm, is approximately 12 dB)

turning on feedback suppression stabilizes the system and eliminates feed-back components completely. The response of the algorithm is very short due to derived short-time weighting rule. In the same conditions the LMS algorithm was unable noticeably increase the maximum stable gain.

## **CONCLUSION**

The paper presents speech enhancement techniques for a smartphone-based hearing aid.

The processing of the signal is performed using DFT-modulated filter bank and include noise reduction and acoustic feedback suppression. The paper introduces an acoustic feedback suppression algorithm based on spectral subtraction that is robust to rapid changes in feedback path. According to experimental results the technique provides high additional gain and high quality of processed speech.

## **ACKNOWLEDGEMENT**

This work was supported by ITForYou company (Moscow, Russian Federation).

## **REFERENCES**

1. A. Vonlanthen, H. Arndt. *Hearing instrument technology for the hearing health care professional 3rd Edition*, New York: Thomson Delmar Learning, Clifton Park, 2006.
2. E.S. Azarov, M.I. Vashkevich, S.V. Kozlova, A.A. Petrovsky. "Hearing correction system based on mobile computing platform," *Informatics*, vol.42, no. 2, pp. 7–25, April 2014. (in Russian).
3. IT ForYou. (2014). "Petralex hearing aid v1.4.3." [online] Available: [play.google.com/store/apps/details?id=com.it4you.petralex](http://play.google.com/store/apps/details?id=com.it4you.petralex), [itunes.apple.com/us/app/petralex-hearing-aid/id816133779?mt=8](http://itunes.apple.com/us/app/petralex-hearing-aid/id816133779?mt=8)
4. J. Ismaili, El H.O. Ibrahimi. "Mobile learning as alternative to assistive technology devices for special needs students," *Educ. and Inform. Technol.*, no.1, pp. 1–17, Jan. 2016.
5. R. W. Bäuml and W. Sörgel, "Uniform polyphase filter banks for use in hearing aids: design and constraints," in Proc. of *Proc. European Signal Process. Conf.*, Lausanne, Switzerland, Aug. 2008.
6. A. Pandey and V.J. Mathews "Low-delay signal processing for digital hearing aids," *IEEE Trans.s on Audio, Speech, and Lang. Process.*, vol. 19, no. 4, pp. 699–710, May 2011.
7. S. Bertoli, K. Staehelin, E. Zemp, C. Schindler, D. Bodmer, R. Probst "Survey on hearing aid use and satisfaction in Switzerland and their determinants," *Intern. journal of audiology*, vol. 48, no.4, pp. 183–195, 2009.
8. J.A. Maxwell and P.M. Zurek, "Reducing acoustic feedback in hearing aids," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 304–313, 1995.
9. R. Vicen-Bueno, A. Martinez-Leira, R. Gil-Pita, and M. Rosa-Zurera, "Modified LMS-based feed-back-reduction subsystems in digital hearing aids based on WOLA filter bank," *IEEE Trans. on Instrumentation and Measurement*, vol. 58, no. 9, pp. 3177–3190, May 2009.
10. H. Schepker, and S. Doclo, "A semidefinite programming approach to min-max estimation of the common part of acoustic feedback paths in hearing aids," *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 24, issue. 2, pp. 366–377, Feb. 2016.
11. H. Schepker, L.Tran, S. Nordholm, and S. Doclo, "Improving adaptive feedback cancellation in hearing aids using an affine combination," in Proc. *IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, Shanghai, China. March 2016, pp. 231–235.



12. F. Strasser and H. Puder, "Correlation detection for adaptive feedback cancellation in hearing aid," *IEEE Signal Processing Letters*, vol. 23, no. 7, pp. 979–983, June 2016.
13. J.M. Kates, "Room reverberation effects in hearing aid feedback cancellation," *Journal Acoust. Soc. Am.*, vol. 109, no. 1, pp. 367–378, Jan. 2001.
14. A. F. Rocha and A. J. S. Ferreira "An accurate method of detection and cancellation of multiple acoustic feedbacks," in *Preprints AES 118th Conv.*, Barcelona, Spain, May 2005, AES Preprint 6335.
15. T. van Waterschoot and M. Moonen "Comparative evaluation of howling detection criteria in notch-filter-based howling suppression," *J. Audio Eng. Soc.*, vol. 58, no. 11, pp. 923–940, Nov. 2010.
16. S. M. Kuo and J. Chen "New adaptive IIR notch filter and its application to howling control in speakerphone system," *IEE Electron. Lett.*, vol. 28, no. 8, pp. 764–766, Aug. 2002.
17. K. Ngo, T. van Waterschoot, M. G. Christensen, M. Moonen, S. H. Jensen and J. Wouters "Prediction-error-method-based adaptive feedback cancellation in hearing aids using pitch estimation," in *Proc. European Signal Process. Conf.*, Aalborg, Denmark, Aug. 2010, pp. 40–44.
18. T. Van Waterschoot and M. Moonen "Fifty years of acoustic feedback control: state of the art and future challenges," *Proc. of the IEEE*, vol. 99, no. 2, pp. 288–327, Feb. 2011.
19. D. Alfsmann, H. G. Göckler and T. Kurbiel "Filter bank for hearing aids applying subband amplification: a comparison of different specification and design approaches," in *Proc. European Signal Process. Conf.*, Glasgow, UK, Aug. 2009, pp. 2663 — 2667.
20. M. Rosa-Zurera, R. Gil-Pita, E. Alexandre Cortizo, M. Utrilla Manso, L. Cuadra-Rodriguez, "WOLA filter bank design requirements in hearing aids," in *Proc. Intern. Conf. on Pattern Recogn. and Inform. Process.*, Minsk, Belarus, May 2009, pp. 215–218.
21. M. Vashkevich, E. Azarov and A. Petrovsky "Low-delay hearing aid based on cochlear model with nonuniform subband acoustic feedback cancellation," in *Proc. European Signal Process. Conf.*, Bucharest, Romania, Aug., 2012, pp. 514–518.
22. K. M. Kates and K. H. Arehart "Multichannel dynamic-range compression using digital frequency warping," *EURASIP J. Appl. Signal Process.*, vol. 18, no. 1, pp. 3003–3014, Dec. 2005.
23. J. Ryan, S. Tewari "A digital signal processor for musicians and audiophiles," *Hearing Reveiw*, vol. 16, no. 2, pp. 38–41, Feb. 2009.
24. H. W. Lollmann and P. Vary "Generalized filter-bank equalizer for noise reduction with reduced signal delay," in *Proc. Interspeech*, Lisbon, Portugal, Sept. 2005, pp. 2105–2108.
25. T. Schneider and R. Brennan "A multichannel compression strategy for a digital hearing aid," *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, Munich, Germany, Apr. 1997, pp. 411–414.
26. R. E. Crochiere and L. R. Rabiner *Multirate digital signal processing*, New Jersey: Prentice-Hall Inc., 1983.
27. N. Tiwari, P. C. Pandey and P. N. Kulkarni "Real-time implementation of multi-band frequency compression for listeners with moderate sensorineural impairment," in *Proc. Interspeech*, Portland, USA, Sept. 2012, pp. 1860–1863.
28. I. Cohen and B. Berdugo Noise "Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement," *IEEE Signal Processing Letters*, vol 9, no. 1, pp.12–15, Jan. 2002.

29. K.W. Berger, E.N. Hagberg and R.L. Rane "Determining hearing aid gain," *Hearing Instruments*. vol. 30, no.4, pp. 26–44, 1980.
30. G.A. McCandless and P.E. Lyregaard "Prescription of gain/output (POGO) for hearing aids," *Hearing Instruments*, vol. 35, no. 1, — pp. 16–21, 1983.
31. D. Byrne and H. Dillon "The national acoustic laboratories (NAL) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and Hearing*, vol. 7, no.7, pp. 257–265, 1986.
32. S. R. Quackenbush, T. P. Barnwell, and M. A. Clements *Objective Measures of Speech Quality*, Englewood Cliffs, New Jersey: Prentice Hall, 1988.

## **ОБРАБОТКА РЕЧЕВОГО СИГНАЛА В СЛУХОВЫХ АППАРАТАХ НА ОСНОВЕ СМАРТФОНА**

**Максим Иосифович Вашкевич,**  
кандидат технических наук, доцент Белорусского государственного  
университета информатики и радиоэлектроники (БГУИР)

**Илья Сергеевич Азаров,**  
доктор технических наук, доцент БГУИР

**Александр Александрович Петровский,**  
доктор технических наук, профессор кафедры электронных  
вычислительных средств БГУИР

### **Аннотация**

В статье предложены методы обработки речевого сигнала для усовершенствованного слухового аппарата на основе смартфона, в основу которого положено наше бесплатное, недавно представленное, приложение «Petralex» для устройств iOS и Android. В данной работе показано новое решение, в котором преодолеваются ограничения обработки в широкополосном частотном диапазоне и расширяется функциональность аппарата. В новой схеме обработки осуществляется декомпозиция речевого сигнала на перцептуально согласованные частотные полосы и осуществляется спектральное усиление для компенсации потери слуха, сжатие динамического диапазона, снижение шума и подавление акустической обратной связи. Мы предлагаем алгоритм подавления акустической обратной связи, основанный на правиле спектрального вычитания. Алгоритм устойчив к быстрым изменениям путем акустической обратной связи и, согласно экспериментам, позволяет добиться стабильного усиления до 24 дБ. Статья состоит из теоретического обзора, описания реализованных методов и некоторых экспериментальных результатов.

**Ключевые слова:** слуховой аппарат, редактирование шума, подавление обратной акустической связи



# Особенности речевого развития детей дошкольного возраста с нарушениями развития, воспитывающихся в детском доме

**Ольга Владимировна Фролова,**  
научный сотрудник кафедры высшей нервной деятельности  
и психофизиологии биологического факультета Санкт-Петербургского  
государственного университета, группа по изучению детской речи

**Шанбиги Габибулаховна Бедалова,**  
студентка 2 курса магистратуры кафедры высшей нервной  
деятельности и психофизиологии биологического факультета  
Санкт-Петербургского государственного университета, группа  
по изучению детской речи

**Елена Евгеньевна Ляксо,**  
профессор кафедры высшей нервной деятельности и психофизиологии  
биологического факультета Санкт-Петербургского государственного  
университета, руководитель группы по изучению детской речи

## Аннотация:

Цель исследования — выявить специфику перцептивных и акустических характеристик речевых высказываний детей дошкольного возраста с нарушениями развития, воспитывающихся в детском доме, по сравнению с типично развивающимися детьми, воспитывающимися в условиях семьи. Показано, что в диалогах с взрослыми дети из детского дома с диагнозами умственная отсталость и смешанные специфические расстройства психологического развития используют менее сложные реплики, чем типично развивающиеся дети; описаны особенности лексикона, установлены перцептивные и акустические характеристики слов воспитанников детского дома. Показаны высокие значения частоты основного тона и длительности ударных гласных из слов детей с диагнозом умственная отсталость. Результаты исследования могут быть использованы для построения стратегий взаимодействия с детьми с нарушениями развития, уточнения существующих методик обучения таких детей, в том числе методик альтернативной коммуникации.

**Ключевые слова:** детская речь, лексикон ребенка, спектрографический анализ, умственная отсталость, смешанные специфические расстройства психологического развития, детский дом

## ВВЕДЕНИЕ

В настоящее время хорошо известно, что материнская депривация приводит к нарушению развития интеллектуальной сферы и речевого развития [1,

2, 3]. Специалисты-практики, работающие с детьми, воспитывающимися в детском доме, отмечают незрелость эмоционально-волевой сферы детей, что обуславливает нарушения в сфере коммуникации. Типично развивающиеся дети уже к четырем годам имеют сформированные речевые навыки, которые обеспечивают адекватную реализацию коммуникативной функции речи. У детей, растущих в доме ребенка (где воспитываются дети до четырехлетнего возраста), навыки речевого общения недостаточно сформированы, что затрудняет их общение с взрослыми [4]. Большая часть работ, посвященных развитию детей из детских домов, имеет практическую, логопедическую направленность. Описываются особенности речевой коммуникации детей-сирот [5], их общение со сверстниками и взрослыми [6].

В зарубежной литературе существует ряд актуальных исследований, направленных на изучение акустических характеристик речи детей с нарушениями развития, например, синдромом Дауна [7], расстройствами аутистического спектра [8]. На материале русского языка такие исследования проводятся в группе по изучению детской речи Санкт-Петербургского государственного университета, где получены данные об акустических характеристиках речи типично развивающихся детей и детей с расстройствами аутистического спектра [9, 10, 11].

Цель данного исследования — выявить специфику перцептивных и акустических характеристик речевых высказываний детей дошкольного возраста с нарушениями развития, воспитывающихся в детском доме, по сравнению с типично развивающимися детьми, воспитывающимися в условиях семьи.

## **МЕТОДИКА**

В исследовании приняли участие дети 4–7 лет трёх групп:

- 1) Дети из детского дома, с диагнозом смешанные специфические расстройства психического развития (CCP, F 83 по Международной классификации болезней — МКБ 10 пересмотра [12], 1998 г.; n = 15). Дети, входящие в данную группу, не имели органических поражений головного мозга и имели самый «лёгкий» диагноз из всех воспитанников коррекционного детского дома;
- 2) Дети из детского дома с диагнозом умственная отсталость лёгкой степени (УО, F 70 по МКБ-10; n = 10); для исследования выбраны только дети, не имеющие генетических синдромов (например, синдром Дауна), тяжелых неврологических заболеваний (детский церебральный паралич — ДЦП);
- 3) Типично развивающиеся дети, воспитывающиеся в условиях семьи, посещающие детский сад общеразвивающего вида (n = 50).

В исследовании использована методика, разработанная в группе по изучению детской речи Санкт-Петербургского государственного университета [10, 13]. Исследование проводили в условиях детского сада для ТР детей и детского дома для детей с УО и CCP. Аудио- и видеозапись речи и поведения детей осуществляли в модельных ситуациях: 1) диалог с экспериментатором с определённым списком вопросов, задаваемых ребенку; 2) повторение детьми набора слов с ударными гласными русского языка /а/, /у/, /и/ за экспериментатором («зайка», «киска» — слова, употребляемые детьми в спонтанной речи). При взаимодействии с каждым ребенком (не зависимо от группы детей), модельные ситуации были реализованы



последовательно, экспериментатор стандартизировал своё поведение и ситуацию взаимодействия. Для записи речи и поведения использована аппаратура: видеокамера SONY HDR-CX560E; видеокамера SONY HDR-CX330E, цифровой магнитофон Marantz PMD222 с выносным микрофоном SENNHEIZER e835S.

На основании анализа аудио- и видеозаписей модельной ситуации 1-диалог созданы тексты диалогов детей с взрослыми, определена сложность ответных реплик ребенка: реплика представлена одним словом, простой фразой, двумя или несколькими простыми фразами, реплика содержит сложноподчинённое предложение, представляет собой ответ «да/нет».

С использованием программы для семантического анализа текста [14] проанализирован лексикон, используемый детьми в диалогах с взрослым, по частоте употребления различных частей речи, наиболее частотным словам.

Произведены перцептивные эксперименты:

1. Перцептивный эксперимент, направленный на оценку использования детьми трёх групп вербальных и невербальных средств коммуникации в ситуации 1-диалог с экспериментатором. На основании анализа видеозаписей взаимодействия экспериментатора и ребенка двумя экспертами созданы 25 видеотестов, содержащих фрагмент диалога ребенка и экспериментатора, длительностью 1 минута, тема диалога — прогулки и общение с друзьями. Видеотесты просматривали эксперты, имеющие профессиональный опыт работы с детьми в Группе по изучению детской речи СПбГУ ( $n = 6$ , возраст  $24,7 \pm 6,6$  лет, 3 — мужского, 3 — женского пола). В специально разработанной анкете эксперты отвечали на следующие вопросы, характеризующие поведение ребенка («да», «нет», дополняли ответы комментариями): 1) Реплики ребенка соответствуют вопросам взрослого по смыслу; 2) Смысл высказываний ребенка понятен; 3) Ребенок демонстрирует желание отвечать и интерес к беседе; 4) Ребенок общается эмоционально; 5) Мимика ребенка выразительна; 6) Ребёнок использует жесты, дополняющие речевое высказывание; 7) Ребёнок использует жесты вместо речевого высказывания;
2. Перцептивный эксперимент, направленный на сравнение эмоциональности спонтанной речи детей трёх исследуемых групп и определение возможности распознавания взрослыми смысла фраз детей. Носителям языка — аудиторам (45 человек, возраст  $18,7 \pm 2,6$  лет, с бытовым опытом взаимодействия с детьми — 35 человек, без опыта — 10 человек, 14 мужчин, 31 женщина) предъявляли 3 тестовые последовательности, включающие высказывания детей (ответные реплики в диалоге с экспериментатором). Каждая тестовая последовательность включала 20 высказываний детей трёх групп, замешанных в свободном порядке. Пауза между высказываниями — 10 секунд. В специальных анкетах аудиторы отмечали, понятен ли им смысл высказывания ребенка, эмоциональна ли речь ребенка;
3. Перцептивный эксперимент, цель которого — оценка возможности определения взрослыми носителями языка значений слов детей групп ССР и УО, вырезанных из контекста фразы. Созданы 4 тестовые по-

следовательности, каждая включала 25 слов детей, воспитывающихся в детском доме, вырезанных из контекста фразы (ответной реплики в диалоге с взрослым). Каждое слово в тестовой последовательности повторялось 3 раза с интервалами в 5 секунд, интервал между разными словами — 10 секунд. Группам аудиторов ( $n = 85$ , возраст —  $19 \pm 2$  лет, с бытовым опытом взаимодействия с детьми — 59 человек, 17 мужчин, 68 женщин) предлагали записать лексическое значение услышанных слов детей;

4. Перцептивный эксперимент с целью определения соответствия между словом взрослого и повторяемым вслед за ним, словом ребенка (ситуация 2-повторение). Созданы 4 тестовые последовательности, каждая содержала слова взрослого и повторяемые слова детей трёх групп ( $n=35$  слов в каждой тестовой последовательности). Каждый образец — пара слов взрослый — ребенок — была включена в тестовую последовательность 1 раз, пауза между предъявляемыми образцами — 10 секунд. Аудиторы ( $n = 72$ ; возраст  $19 \pm 1,8$ ; с бытовым опытом взаимодействия с детьми — 61, без опыта — 11 мужского пола, 67 — женского пола) при прослушивании тестовых последовательностей отмечали, соответствует ли то, что произнес ребенок, повторяемому слову взрослого: 1) по значению, 2) по интонации.

Перцептивные эксперименты (2, 3, 4) проводили в открытом поле. Поскольку не было выявлено различий в распознавании значений слов детей в связи с полом и возрастом аудиторов, представлены объединённые данные по всем группам аудиторов.

Произведён спектрографический анализ слов из спонтанной речи детей (ситуация 1-диалог) и набора слов, повторенных за экспериментатором (ситуация 2-повторение). В звуковом редакторе "Cool Edit Pro". Определены длительность слов и ударных гласных, значения частоты основного тона (ЧОТ, F0), диапазона частоты основного тона (разница между максимальным и минимальным значением ЧОТ — [F0max-F0min]) ударных гласных слов из спонтанной речи детей. Выбранные акустические характеристики значимы для оценки эмоциональной речи детей дошкольного возраста [15]. Для повторяемых слов определяли значение длительности, ЧОТ ударного гласного; на стационарном участке гласного — ЧОТ, значения первых двух формантных частот и их интенсивности. На двухформантной плоскости строили треугольники с вершинами, соответствующими значениям первых двух формант ударных гласных /а/, /у/, /и/, определяли площади формантных треугольников [16, 17].

*Статистическая обработка данных проведена в программе «Statistica 10.0».*

*Исследование проведено при поддержке Этического комитета Санкт-Петербургского государственного университета № 00003875.*

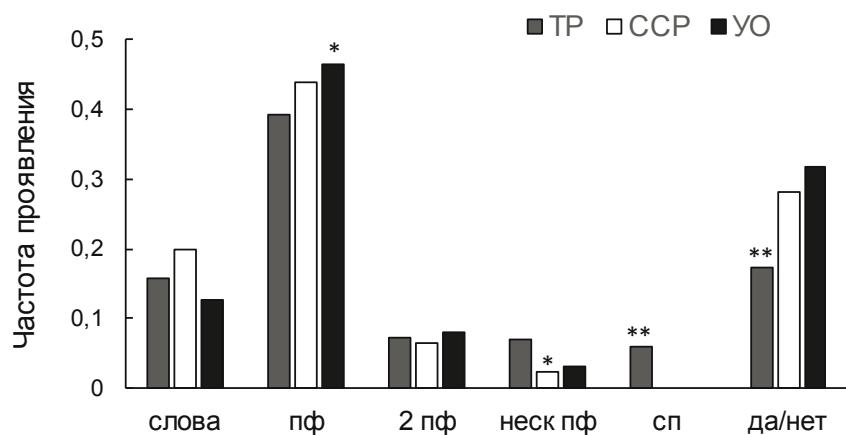
## **РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ**

### **1. Анализ текстов диалогов экспериментатор — ребенок**

На основании анализа текстов диалогов экспериментатор-ребенок установлено, что преобладающий тип ответных реплик детей трёх изучаемых групп — простые фразы и реплики «да/нет» (рис. 1). В меньшей степени дети используют реплики, включающие две или несколько простых фраз и сложноподчинённые предложения. Дети со смешанными специфическими расстройствами психологического



развития и умственной отсталостью, воспитывающиеся в детском доме, использовали в диалоге с экспериментатором меньше реплик, представленных несколькими простыми фразами ( $p < 0,05$  — критерий Манн-Уитни), и больше реплик — ответов «да/нет» ( $p < 0,01$ ), чем типично развивающиеся дети. ТР дети используют реплики, включающие сложноподчинённые предложения. В группах детей с ССР и УО зарегистрированы единичные реплики, включающие сложноподчинённые предложения (у отдельных детей — медиана для группы равна нулю). Значимых различий между детьми групп ССР и УО по сложности их ответных реплик в диалоге не выявлено.



Типы реплик: слова — реплика представлена одним словом; пф — простой фразой; 2 пф — двумя простыми фразами; неск пф — несколькими простыми фразами; сп — включает сложноподчинённое предложение; да/нет — ответ «да» или «нет»

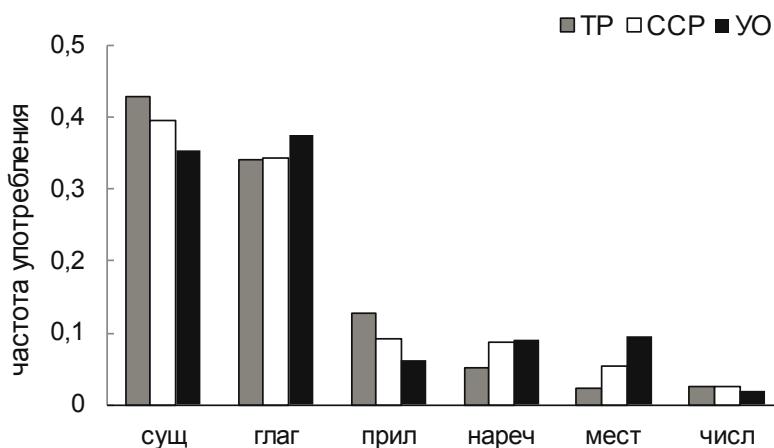
\* —  $p < 0,05$ , \*\* —  $p < 0,01$  различия между группами ТР, ССР и УО, критерий Манн — Уитни

**Рис. 1.** Частота проявления различных типов реплик у ТР детей и детей с диагнозами ССР и УО в диалогах экспериментатор — ребенок (медианы для группы)

Анализ лексикона детей трёх групп показал, что разнообразие используемых в диалогах существительных у детей групп ТР и ССР выше, чем других частей речи (рис. 2), дети с УО используют больше разных глаголов, чем существительных. Установлено, что частота употребления прилагательных максимальна у детей ТР группы, по сравнению с детьми с ССР и в особенности детьми с УО, при этом дети с УО чаще, чем дети других групп, используют местоимения.

В репликах детей с умственной отсталостью присутствовали вокализации, звукосочетания, сложные для интерпретации даже при учёте контекста ситуации.

Выявлены наиболее частотные слова в лексиконе детей трёх групп. Для ТР детей наиболее частотными были существительные — /мама, бабушка/; глаголы — /быть, ходить, звать, любить/; прилагательные — /ма-



Части речи: сущ — существительные, глаголы — глаголы, прил — прилагательные, нареч — наречия, мест — местоимения, числ — числительные

**Рис. 2.** Частота употребления разных частей речи ТР детьми и детьми с диагнозами CCP и YO в диалогах

ленький, большой/, названия цветов; местоимения — /я, она, он/; числительное — /один/; для детей с диагнозом CCP — /дом (в сочетании «детский дом»), дача/, наименования животных; глаголы — /быть, играть, хотеть/; прилагательные — /большой/; местоимения — /я, мы/; детей с YO существительное — /дом/; глаголы — /быть, ходить, играть/; местоимения — /я, мы, все/.

Полученные данные свидетельствуют о влиянии факторов заболевания и социальной среды на сложность реплик в диалоге и активный лексикон ребенка и согласуются с результатами исследования об особенностях речевых навыков детей младшего возраста, воспитывающихся в доме ребенка [4], и данными психологической литературы [2].

## 2. Перцептивный анализ

На основании экспериментального анализа видеофрагментов диалогов детей трёх групп с экспериментатором установлено, что реплики ТР детей и детей с диагнозом CCP соответствуют вопросам взрослого по смыслу (98 и 100% ответов «да» экспертов в анкетах). Дети с YO не всегда отвечают адекватно на вопросы (77% ответов экспертов). Смысл высказываний ТР детей чаще понятен экспертам (94% ответов), чем детей с CCP и YO. Желание отвечать на вопросы экспериментатора и интерес к беседе у детей трёх групп значимо не различается, однако детей группы CCP эксперты оценили как менее эмоциональных собеседников, чем их ТР и YO сверстников (60% в группе CCP и 76, 73% — в группе ТР и YO, соответственно). Эксперты отметили, что ТР дети и дети с YO используют выразительную мимику и множество жестов, дополняющих речевое высказывание. Дети с CCP, напротив, используют жесты, заменяющие речевое высказывание (табл. 1).

Данные перцептивного анализа (перцептивный эксперимент 2) позволили заключить, что носители языка на основании слухового восприятия однозначно (с вероятно-



Таблица 1

**Количество ответов экспертов («да» %) в перцептивном эксперименте по определению особенностей поведения детей в ситуации диалога с экспериментатором**

Поведение ребенка	Группы детей		
	TP	CCP	УО
Реплики ребенка соответствуют вопросам взрослого по смыслу	98	100	77
Смысл высказываний ребенка понятен	94	70	50
Ребенок демонстрирует желание отвечать и интерес к беседе	89	80	87
Ребенок общается эмоционально	76	60	73
Мимика ребенка выразительна	72	50	60
Ребёнок использует жесты, дополняющие речевое высказывание	73	53	77
Ребёнок использует жесты вместо речевого высказывания	20	57	40

стью  $p > 0,75\%$ ) относят к эмоциональным высказываниям только 10% ответных реплик TP детей и детей с CCP, 25% реплик детей с УО.

Смысл высказываний TP детей аудиторы распознают с большей вероятностью, чем детей с CCP и УО: с высокой вероятностью ( $p > 0,75$ ) носители языка распознают смысл 55% высказываний TP детей, 30% высказываний детей с CCP и 25% детей с УО. При этом лексическое значение слов детей групп CCP и УО, вырезанных из контекста фразы (перцептивный эксперимент 3), носители языка однозначно ( $p > 0,75$ ) распознавали в 37% случаев. (По данным литературы — аудиторы с высокой вероятностью ( $p > 0,75$ ) распознают 46,8% слов здоровых детей 5 лет; 50,7% слов детей 6 лет, 54,2% слов здоровых детей 7 лет, воспитывающихся в условиях семьи [16]).

На основании перцептивного анализа ситуации 2- повторение установлено, что TP дети повторяют слова за экспериментатором точнее, чем дети с CCP и УО. Аудиторы отметили соответствие произнесённого ребенком слова слову взрослого по значению в 95% случаев в группе TP детей, 66% случаев в группе CCP и 59% случаев — в группе детей с УО. Соответствие по интонации — в 53% случаев в группах TP и CCP и 47% — в группе детей с УО.

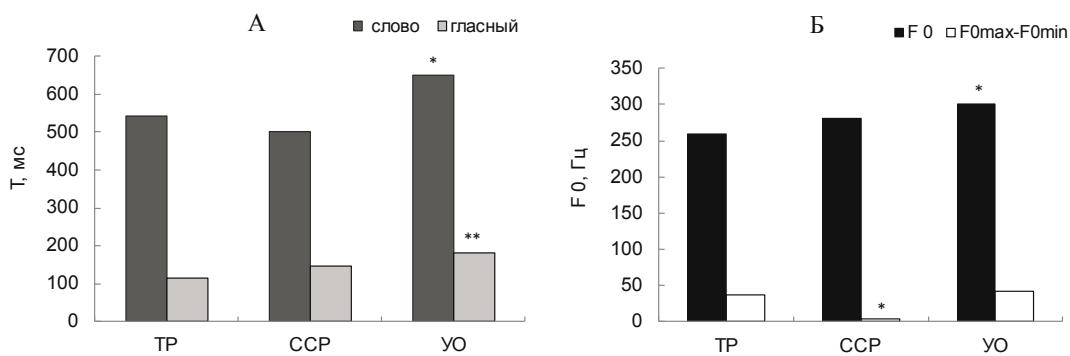
### 3. Спектрографический анализ

Анализ акустических характеристик слова из высказываний детей (ответных реплик в диалоге с экспериментатором) показал, что значения длительности слов и ударных гласных детей с УО значимо выше, чем соответствующие параметры TP детей ( $p < 0,05$ ,  $p < 0,01$  — для слов и ударных гласных, соответственно). Значения частоты основного тона ударных гласных слов детей с УО выше, чем TP детей. Значения диа-

пазона ЧОТ ударных гласных слов детей с CCP значимо ниже, чем ТР детей и детей с УО ( $p < 0,05$ ) (рис. 3).

Согласно полученным нами ранее данным [15, 18], высказывания детей, произнесённые в состоянии дискомфорта и комфорта, характеризуются высокими значениями длительности, ЧОТ и диапазона ЧОТ ударных гласных. Таким образом, данные перцептивного анализа — большая вероятность отнесения высказываний детей с диагнозом УО к эмоциональным — подтверждаются данными акустического анализа.

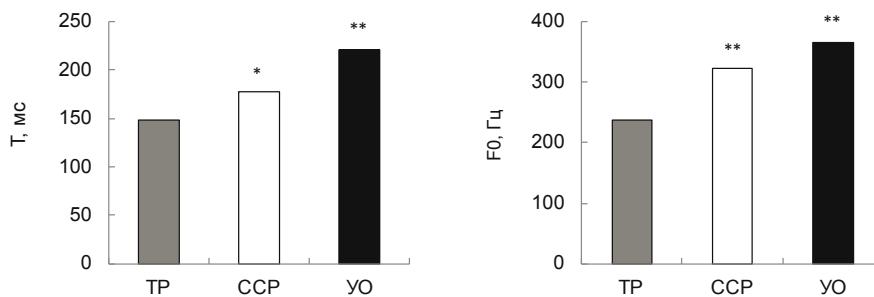
Акустический анализ слов детей, повторенных за экспериментатором, позволил выявить те же закономерности, что и в спонтанной речи детей с диагнозом умственная отсталость: значения длительности и ЧОТ ударных гласных детей с УО и CCP выше, чем ТР детей ( $p < 0,01$  и  $p < 0,05$  для УО и CCP, соответственно) (рис. 4).



А — длительность слов и ударных гласных слов детей; Б — значения частоты основного тона и диапазона частоты основного тона ударных гласных слов детей

\* —  $p < 0,05$ , \*\* —  $p < 0,01$  различия между группами ТР, CCP и УО, критерий Манн — Уитни

**Рис. 3.** Акустические характеристики слов из высказываний ТР детей и детей с диагнозами CCP и УО в диалоге с экспериментатором (медианы для группы)



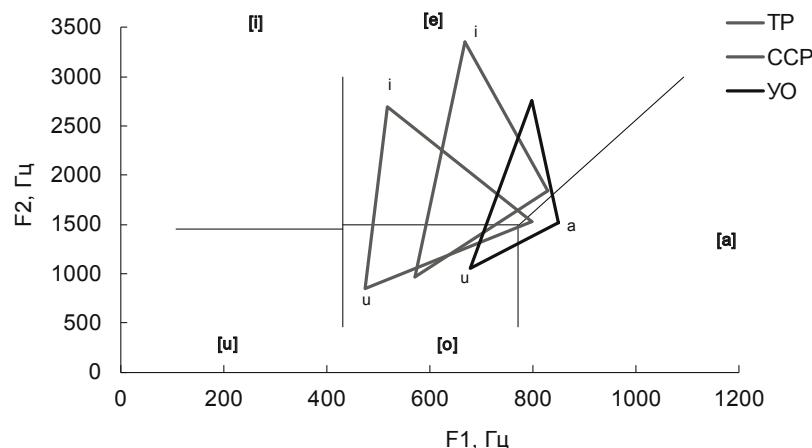
А — длительность ударных гласных; Б — значения частоты основного тона ударных гласных

\* —  $p < 0,05$ , \*\* —  $p < 0,01$  различия между группами ТР, CCP и УО, критерий Манн — Уитни

**Рис. 4.** Акустические характеристики слов ТР детей и детей с диагнозами CCP и УО в ситуации повторение за экспериментатором (медианы для группы)



Формантные треугольники ударных кардиальных гласных из слов детей с CCP и УО в ситуации повторение отличаются от формантных треугольников ударных гласных ТР детей по форме и расположению на двухформантной плоскости (рис. 5).

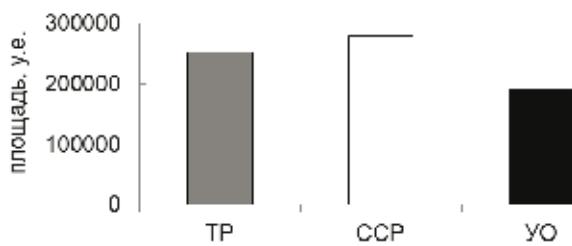


По горизонтальной оси – значения первой форманты, по вертикальной оси – вторая форманты. Серые линии – данные для ТР детей, пунктир – дети с CCP, чёрная жирная линия – дети с УО

**Рис. 5.** Формантные треугольники ударных гласных /а/, /у/, /и/ из слова детей трёх групп в ситуации повторение за экспериментатором

Формантный треугольник ТР детей – в большей степени приближается к соответствующему треугольнику кардиальных гласных взрослой речи [17], формантные треугольники детей с CCP и в особенности с УО смешены в более высокочастотную область двухформантной плоскости, что характерно для детей младшего возраста [19].

Площадь формантного треугольника детей с УО минимальна, по сравнению с другими группами детей (рис. 5, 6). Это может свидетельствовать о недостаточной чёткости артикуляции детей с УО, что также подтверждает данные перцептивного анализа.



**Рис. 6.** Площадь формантных треугольников ударных гласных /а/, /у/, /и/ из слов детей трёх групп в ситуации повторение за экспериментатором

## **ЗАКЛЮЧЕНИЕ**

В исследовании описана специфика ответных реплик в диалоге с взрослым детей, воспитывающихся в детском доме, имеющих диагноз умственная отсталость и смешанные специфические расстройства психологического развития, по сравнению с типично развивающимися детьми. Показано, что дети из детского дома используют менее сложные реплики, чем типично развивающиеся дети, установлено, что факторы заболевания и условий воспитания находят отражение в лексиконе детей.

Данные перцептивных экспериментов позволяют заключить, что спонтанная речь детей с УО воспринимается носителями языка как более эмоциональная, чем речь детей с ССР и типично развивающихся детей. Значение слов и фраз детей с умственной отсталостью распознаётся носителями языка с меньшей вероятностью, чем ТР детей. Установлены акустические характеристики слов детей с нарушениями развития, воспитывающихся в условиях детского дома. Показаны высокие значения частоты основного тона и большая длительность ударных гласных из слов детей с диагнозом умственная отсталость.

Полученные в ходе исследования данные имеют практическое значение для оценки речевых навыков детей, воспитывающихся в условиях детского дома, что необходимо для правильного построения стратегий взаимодействия с такими дошкольниками. Результаты исследования могут быть использованы для уточнения существующих методик обучения детей, в том числе — методик альтернативной коммуникации с детьми с нарушениями развития.

*Работа выполнена при финансовой поддержке фонда РФФИ (проекты №18-013-01133, №16-06-00024а, РФФИ — огон № 17-06-00503а).*

## **ЛИТЕРАТУРА**

1. Шпиц Р. А., Коблинер У.Г. Первый год жизни. М.: Академический проект. 2006. 352 с.
2. Лангмайер Й., Матейчек З. Психическая депривация в раннем возрасте. Прага: Авиценум. 1984. 335с.
3. Финашина Т.А. Развитие речи у детей раннего возраста, воспитывающихся в разных условиях: Автореф. канд. дис., Калуга. 2000.
4. Ляксо Е.Е., Столярова Э.И. Специфика реализации речевых навыков 4-5-летних детей в диалоге со взрослыми // Психологический журнал. 2008. Т. 29. № 3. С. 48-57.
5. Янчева С.В. Развитие речевой коммуникации у детей-сирот дошкольного возраста с интеллектуальной недостаточностью: Автореф. канд. дис., М. 2010.
6. Виноградова Н.В., Рычкова Л.С. Особенности общения детей-сирот в процессе межличностного взаимодействия со сверстниками // Вестник ЮУрГУ. Серия «Психология», выпуск 13. 2011. № 18. С. 86-88.
7. Kent R.D., Vorperian H.K. Speech Impairment in Down Syndrome: A Review // Journ. Speech Lang Hear Res. 2013. Vol. 56 (1), pp. 178–210.
8. Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D.M., Gaigg, S.B.: Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis // Autism Res. 2017. Vol. 10(3), pp. 384–407.



9. Lyakso E., Frolova O., Grigorev A. A Comparison of Acoustic Features of Speech of Typically Developing Children and Children with Autism Spectrum Disorders // Lecture Notes in Computer Science. 2016. Vol. 9811, pp 43–50.
10. Ляксо Е.Е., Фролова О.В., Григорьев А.С., Соколова В.Д., Яроцкая К.А. Распознавание взрослыми эмоционального состояния типично развивающихся детей и детей с расстройствами аутистического спектра // Рос. Физиол. журнал им. И.М. Сеченова. 2016. Т. 102, № 6. С. 729–741.
11. Lyakso E., Frolova O., Grigorev A. Perception and Acoustic Features of Speech of Children with Autism Spectrum Disorders // Lecture Notes in Computer Science. 2017. Vol. 10458, pp. 602–612.
12. Международная классификация болезней 10-го пересмотра (МКБ-10) // <http://mkb-10.com>
13. Ляксо Е.Е., Фролова О.В., Смирнов А.Г., Куражова А.В., Гайкова Ю.С., Бедная Е.Д., Григорьев А.С. Уровень речевого развития детей на этапе формирования навыка чтения // Психологический журнал. 2012. Т. 33, № 1. С. 73–87.
14. <https://advego.ru/text>
15. Kaya H., Ali Salah A., Karpov A., Frolova O., Grigorev A., Lyakso E. Emotion, age, and gender classification in children's speech by humans and machines // Computer Speech and Language. 2017. Vol. 46, pp. 268–283.
16. Григорьев А.С., Ляксо Е.Е. Слуховое восприятие слов детей 5–8 лет // Сенсорные системы. 2014. Т. 28, № 3. С. 28–35.
17. Lyakso E.E., Grigor'ev A.S. Dynamics of the duration and frequency characteristics of vowels during the first seven years of life in children // Neuroscience and Behavioral Physiology. 2015. Vol. 45(5), pp. 558–567.
18. Lyakso E., Frolova O., Dmitrieva E., Grigorev A., Kaya H., Salah A. A., Karpov A. EmoChildRu: emotional child Russian speech corpus // Lecture Notes in Computer Science. 2015. Vol. 9319, pp. 144–152.
19. Lyakso E.E., Frolova O.V., Grigor'ev A.S. The acoustic characteristics of Russian vowels in children of 6 and 7 years of age. Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009. Brighton, pp. 1739–1742.

## **SPEECH DEVELOPMENT OF PRESCHOOL CHILDREN WITH DEVELOPMENTAL DISORDERS GROWING UP IN AN ORPHANAGE**

**Olga V. Frolova ,**

*researcher, Department of higher nervous activity and psychophysiology of biological faculty, St. Petersburg state University, group for the study of child language*

**Shanbigi G. Bedalova,**

*2nd year master's student of the Department of higher nervous activity and psychophysiology of the biological faculty of St. Petersburg state University, group for the study of child language*

**Elena E. Lyakso,**

*Professor of the Department of higher nervous activity and psychophysiology biological faculty of St. Petersburg state University, head of the group for the study of children's speech*

### **Abstract**

The aim of the study is to reveal the specificity of the perceptual and acoustic characteristics of the speech utterances of preschool children with developmental disorders brought up in the orphanage, in comparison with the typically developing children brought up in the family. Children from an orphanage with mental retardation and mixed specific psychological development disorders use less complex dialogue replicas than typical developing children; the features of their lexicon are described. Perceptual and acoustic features of the words of orphans are revealed. The high frequencies of the pitch and the longer duration of the stressed vowels from the words of children with mental retardation are shown. The results of the research could be used to build the strategies for interaction with children with developmental disabilities, for elaboration of existing methods of teaching children, including alternative communication techniques.

**Keywords:** Child's speech, child's lexicon, spectrographic analysis, mental retardation, mixed specific psychological development disorders, orphanage



# Задача автоматической расстановки знаков пунктуации в распознанной спонтанной русской речи

**Дмитрий Анатольевич Бирин,**  
генеральный директор филиала ФГУП «НИИ «Квант» ,  
Санкт-Петербург

**Александр Евгеньевич Булашевич,**  
кандидат технических наук, научный сотрудник филиала  
ФГУП «НИИ «Квант», Санкт-Петербург

**Марианна Юрьевна Грекис,**  
инженер ФГУП «НИИ «Квант», Санкт-Петербург

## Аннотация:

Основная цель процесса распознавания речи — получение на выходе удобочитаемого, ясного текста. В русском языке это практически невозможно без знаков препинания. Проблема в том, что существующая система правил пунктуации была разработана для письменного языка. В спонтанной речи эти правила часто не соблюдаются и даже нарушаются. Кроме того, для спонтанной речи характерны такие явления, которые не описаны в правилах, сформулированных для литературного (письменного) языка, поскольку эти явления там практически отсутствуют (например, хезитационный поиск, самоисправления и т.д.). Таким образом, задача заключается в том, чтобы адаптировать классические правила для спонтанной речи и разработать систему автоматической пунктуации, которая сможет превратить последовательность распознанных слов спонтанной речи в понятный письменный текст. На данном этапе наша система позволяет в большинстве случаев однозначно определять границы предложения и с определённой точностью ставить внутренние знаки препинания.

**Ключевые слова:** распознавание спонтанной речи, пунктуация в спонтанной речи, автоматическая расстановка знаков препинания.

## ВВЕДЕНИЕ

Русский язык, как в устной, так и в письменной форме, представляет собой весьма сложную систему, функционирующую на основе набора некоторых принятых стандартов, так называемых правил, в соответствии с которыми в письменной речи употребляется та или иная буква (графема) или проставляется тот или иной знак препинания. Эти стандарты — результат труда многих квалифицированных специалистов, и они изложены в многочисленных авторитетных изданиях. Однако в

реальной жизни человек использует выработанные каноны лишь в определённой мере, которые зависят прежде всего от его так называемого социолингвистического статуса, т.е. уровня образования, культурного уровня, окружающего социума и т.п. Не все русские владеют правилами русского языка в совершенстве. «Среднестатистический» гражданин на практике достигает лишь определённого уровня грамотности, то есть использует лишь тот объём знаний, который получен им в процессе обучения, например в объёме средней школы. Это касается владения как устной, так и письменной речью.

Значительную сложность в письменной речи представляет собой система пунктуации.

Правила пунктуации весьма сложны, простановка знаков препинания часто неоднозначна и зависит от определённых исходных условий. Совершенно очевидно, что разработка алгоритмов для автоматизации процесса расстановки знаков препинания даже в «нормативном» тексте представляется весьма сложной задачей. Еще более трудной оказывается задача расстановки знаков препинания в спонтанной, т.е. заранее не продуманной, не «запограммированной» речи, когда фразы составляются «на ходу», выражаясь простым языком, когда слово опережает мысль. В этом случае говорящий не будет выстраивать правильные синтаксические конструкции, подбирать соответствующую лексику и т.п. Даже без какого-либо научного анализа понятно, что в этом случае языковые нормы «перестают работать» — разрушаются синтаксические связи, предложения часто неполные, слова могут не согласоваться друг с другом и т.п.

Расстановка знаков препинания в тексте, представляющем собой задокументированную спонтанную речь, является непростой задачей даже для лингвиста, что позволяет представить всю сложность попытки автоматизации решения данной задачи.

## 1. ПУНКТУАЦИЯ В СПОНТАННОЙ РЕЧИ: ПРОБЛЕМАТИКА

Очевидно, что системы пунктуации письменной речи недостаточно для транскрибирования устной речи. Например, невозможно средствами пунктуации передать выделенность слова, невозможно различить фразы «*Петр приехал?*» и «*Петр приехал?*», выражющие различные по смыслу вопросы.

Возникающие здесь трудности объясняются существенным различием между двумя системами национального языка — разговорной речью и кодифицированным литературным языком, на который преимущественно рассчитаны действующие правила пунктуации [1].

Многие явления, которые характерны для спонтанной речи, в письменной речи практически не встречаются. Например, хезитация, самоперебивы, повторы, неоформленность синтаксических и семантических связей в предложении (исключение — фрагменты художественной литературы, передающие устную речь персонажей). Проблема в том, что правила пунктуации, рассчитанные на литературный язык, эти явления никак не описывают.

Трудности возникают уже на этапе определения границ предложения.

Очень просто выделить предложение в письменном тексте. Знаки препинания — точка, вопросительный, восклицательный знаки, многоточие — почти однозначно отмечают конец предложения. И дальше можно определять его тип: сложное или про-



стое, выявлять связи между членами предложения и т.д. Казалось бы, те же процедуры легко проделает любой человек, а тем более лингвист, и со спонтанной звучащей речью. Однако, как показывает практика и как пишут исследователи, занимающиеся спонтанной (живой) речью, выделить в ней предложение — задача весьма сложная, а может быть, даже невыполнимая [2].

На филологическом факультете СПбГУ был проведен перцептивный эксперимент по членению спонтанных монологов на предложения. Предполагалось, что наличие звука позволит экспертам-аудиторам точно и единодушно определить границы предложений. Но это не так. Мнения экспертов во многом разошлись. Как описано в работе А.И. Рыко — С.Б. Степановой [2], аудиторы реализуют различные стратегии: «максималистскую» и «минималистскую». При реализации «максималистской» стратегии в качестве предложений выделяются длинные, многосintагменные структуры. В этом варианте членения текста в большом количестве представлены бессоюзные предложения, сложносочиненные и сложноподчиненные предложения. «Минималистская» стратегия позволяет провести границу везде, где только это можно сделать: два и более простых предложений вместо одного бессоюзного сложного и т.п. Фактически любая завершающая пауза или затянувшаяся межсintагменная пауза в рамках такой стратегии — повод для проведения границы между предложениями. Кроме этого «минималистская» стратегия предполагает ориентацию на интонацию текста. Но самое интересное, что в зависимости от темпа речи, качества звучания, ораторского мастерства диктора аудиторы не последовательны и придерживается то одной, то другой стратегии.

Суть в том, что единственным надежным пограничным сигналом, разбивающим речевой поток на отрезки с ограниченным лексическим составом являются паузы разной природы (вдохи и вздохи, молчание, хезитация, ларингализация). Образующиеся при этом интервалы непрерывного «говорения» существенно различаются по длительности и структуре [3].

В частности, в интервале между двумя паузами может быть произнесено только одно слово из длинной фразы, а несколько слов, разделенных запятыми, — без каких-либо перерывов.

В то же время отрезок речевого акустического сигнала между двумя последовательными паузами может содержать несколько синтагм в их общепринятом значении, причем на их стыках не обнаруживается никаких разрывов в мелодическом контуре, которые предположительно могли бы маркировать границы между синтагмами и одновременно являлись бы маркерами границ между словами. Более того, на стыке двух синтагм может происходить стяжение гласных с полной утратой какой-либо физической границы между ними.

Таким образом, оказывается, что анализ мелодической составляющей присодических признаков (аудиторский или инструментальный) не обеспечивает адекватного синтагматического членения. Поэтому А.В. Венцов приходит к выводу, что в акустическом сигнале, представляющем

спонтанную речь, не существует никаких физических признаков границ между лексическими единицами, кроме пауз [4].

Большинство исследователей сегментируют речь на элементарные дискурсивные единицы (ЭДЕ) [5], соответствующие клаузе (элементарному предложению) или синтагме. Такое членение удобно для исследования речи и процессов речепорождения, но не совсем подходит для сохранения смысловой целостности распознанного текста.

## **2. ОПИСАНИЕ РАЗРАБАТЫВАЕМОЙ СИСТЕМЫ ПУНКТУАЦИИ**

В нашей лаборатории ведётся разработка системы пунктуации (СП), которая должна расставить знаки препинания в потоке слов, являющимся результатом автоматического распознавания звучащей речи. Заметим, задача автоматической расстановки знаков препинания осмыслена только в связке с автоматическим распознаванием речи, так как в противном случае текст уже появляется со знаками препинания. При этом точность расстановки знаков препинания напрямую зависит от точности результатов распознавания.

Задача автоматической пунктуации естественно распадается на три подзадачи: определять границы предложения, выбирать завершающий знак препинания и ставить внутренние знаки препинания. При этом внутренние знаки можно ставить в соответствии с правилами русского языка чисто по тексту, но определение границ предложений и классификация завершающего знака чисто по тексту невозможно (собственно, именно вследствие этой невозможности и возникли различные завершающие знаки).

Соответственно, работа была разделена на несколько направлений:

- Создание речевых корпусов;
- Акустическое направление: детектор границы фраз и классификатор завершающего знака;
- Лингвистическое направление: формализация и программная реализация грамматических норм русского языка с целью расстановки внутренних знаков препинания;
- Разработка методики оценки качества пунктуации.

### **2.1. Создание речевых корпусов**

В рамках этого направления был подготовлен речевой корпус «Пунктуация». Общий объём корпуса — около 27 часов. Объём эталонной части — 8,5 часа.

На данном этапе не исследовались и не анализировались спонтанные монологи. Нашей задачей была разработка системы пунктуации для спонтанной диалогической речи.

Эталонная база проаннотирована в двух видах:

- по акустике — знаки проставлены только там, где они выражены диктором акустически. Обычно в таких местах присутствует пауза (различного характера);
- по правилам русского языка — в соответствии с нормами письменной речи со ссылкой на правило, по которому поставлен тот или иной знак.



### 2.2.1. Детектор границы

Задача детектора границы — определить места постановки завершающих знаков препинания, т.е. границы предложений распознанного текста. Именно текста, а не речи.

Как уже говорилось ранее, среди лингвистов существует точка зрения, что в спонтанной речи вообще отсутствует достаточно четкое членение на фразы. В случае спонтанного диалога в отличие от спонтанного монолога задача детектора границ существенно отличается наличием длинных пауз на месте речи собеседника.

Длительность и структура пауз прежде всего связаны с уровнем «ораторского мастерства» диктора, умением формулировать мысль и чётко ее излагать. Так называемые «публичные» люди или люди, часто говорящие вслух, разделяют свои мысли более короткими паузами ввиду того, что им не требуется длительного обдумывания. Длительные паузы чаще всего говорят о смене диктора или завершении высказывания с ожиданием ответной реакции. Как правило, такие отрезки речи заканчиваются «конечным» знаком препинания или обрывом высказывания (перебив собеседником). Короткие паузы чаще говорят о наличии «внутреннего» знака препинания (запятой, тире, двоеточия и т.п.), но вполне могут быть и «завершающими».

В текущей версии СП используется только один признак — длительность пауз, понимаемой как интервал от конца слова до начала следующего. Это позволяет с приемлемой точностью определять границы предложений. Чисто по длительности паузы в спонтанной диалогической речи (из нашего РК) удается обнаружить границу со спутыванием (SER) порядка 15%. Длительность пороговой паузы — обучаемый параметр. Интересным является факт довольно слабой эластичности погрешности обнаружения границы от величины порога. Улучшить этот результат вспомогательными мерами (адаптация порога по длительности к темпу речи и т.п.) не удалось. Причина в том, что пауза — необходимый, но нестабильный признак границы фразы. В потоке речи достаточно часто встречаются внутрифразовые паузы, более длинные, чем межфразовые паузы в этой же аудиозаписи. На рисунке 1 приведен пример пауз в конкретном аудиофайле.

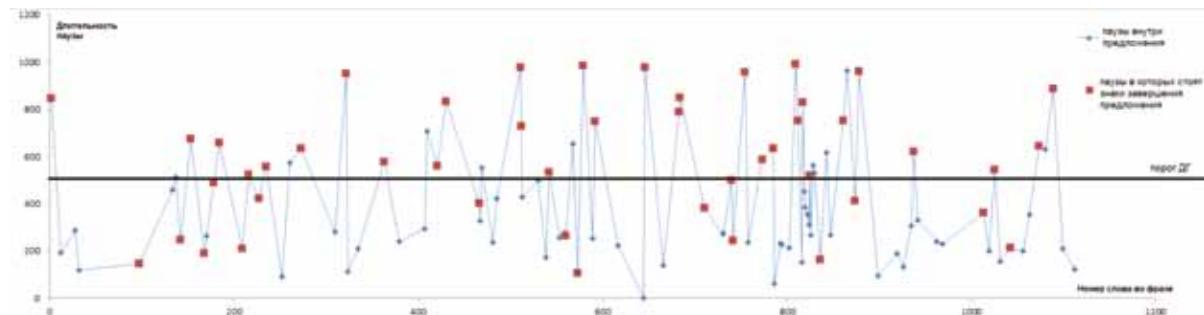
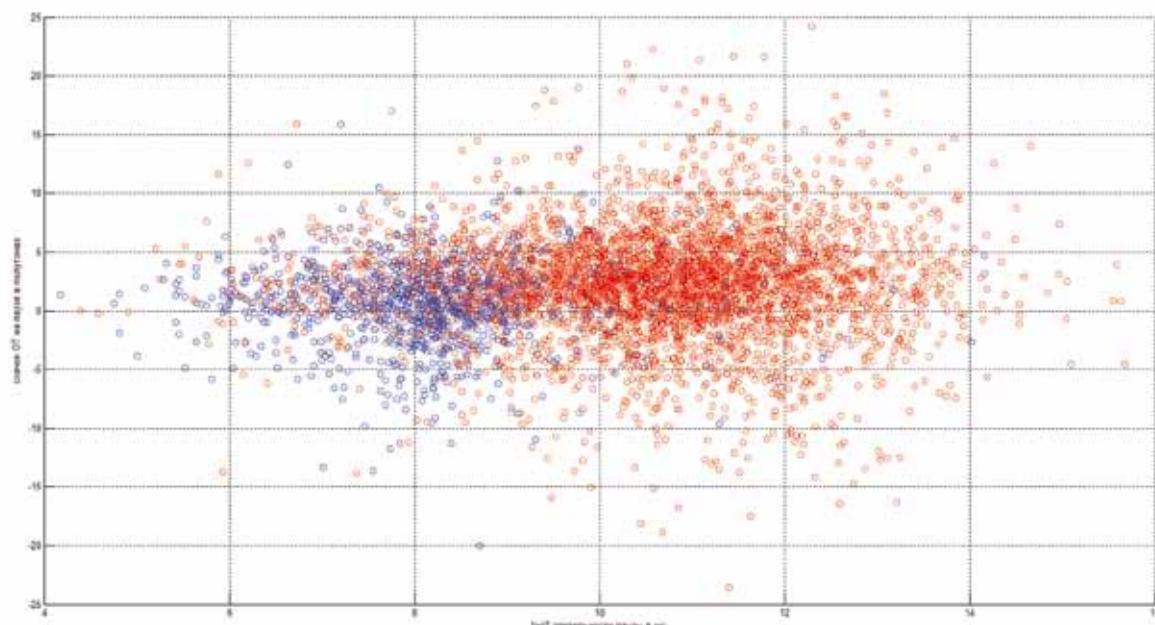


Рис. 1. Длительность межфразовых и внутренних пауз

Большинство фраз завершается понижением основного тона. Поэтому для обнаружения границ фраз был использован признак «скачок основного тона на паузе» в предположении, что большой положительный скачок коррелирует с наличием границы фразы. Однако количественное улучшение точности обнаружения границ составило менее 1%, что неожиданно. Добавление этого признака почти не улучшило разделения (рис. 2).

Быть может, надо использовать более сложные признаки, основанные на основном тоне.



**Рис. 2.** Длительность пауз и скачки основного тона

### 2.2.2. Классификатор завершающего знака

Классификатор знака нуждается в серьезном совершенствовании. Текущая версия классификатора использует признаки, характеризующие форму трека основного тона. Для коротких фраз (одно-два слова) используются признаки, вычисленные относительно конца фразы. Для длинных (более трёх слов) фраз используются признаки, характеризующие форму трека основного тона около участка с наиболее быстрым изменением основного тона — предполагается, что там расположен интонационный центр фразы. Основная нерешенная проблема — правильное определение интонационного центра, необходимое для отделения релевантных изменений основного тона от нерелевантных. SER по завершающим знакам в настоящий момент 42%, что явно недостаточно для практического использования.

### 2.3. Лингвистическое направление

На этом направлении разработан программный механизм расстановки знаков препинания, реализующий правила, записанные на формальном языке YAML.



При разработке этого направления авторы опирались на правила русской пунктуации, изложенные в работах Д.Э. Розенталя [1], и справочник, подготовленный Институтом русского языка им. В.В. Виноградова РАН и Орфографической комиссией при Отделении историко-филологических наук Российской академии наук под редакцией В.В. Лопатина [6].

Для нашей задачи классические правила были формализованы и структурированы. Из таблицы правил номера правил были введены в эталонную базу «Пунктуация». Была подсчитана статистика частотности правил пунктуации, на основании которой ведутся работы по реализации правил. Кроме статистики при последовательности реализации правил учитывается не только частотность, но и возможность реализации данного правила без привлечения дополнительных программных средств.

Для тестирования лингвистического блока СП созданы отдельные тесты, полученные из эталонных аннотаций путём удаления внутренних знаков препинания. В настоящий момент реализованы основные правила русского литературного языка. Например, такие, как:

- Запятые в сложноподчинённых предложениях;
- Обособление вводных слов и выражений;
- Выделение междометий;
- Запятые при частицах и др.

В ходе разработки каждое правило проходит тестирование на примерах из эталонной части корпуса «Пунктуация».

Реализация каждого правила, которое кажется однозначным в письменной речи, сталкивается с большим количеством исключений в спонтанной речи.

Основная сложность возникает с теми явлениями, которые отсутствуют в письменном языке. Однако при этом они очень часто встречаются в спонтанной речи. Это такие явления, как:

- самоисправления;
- вербальная хезитация;
- неоформленность синтаксических и семантических связей;
- нарушение привычных связей в предложении.

Опора на интонацию и паузы различного рода (вдохи и вздохи, молчание, хезитация, ларингализация), как установлено, не может считаться надёжным критерием, поскольку между интонацией и пунктуацией нет полного соответствия (возможно: есть пауза, но нет знака препинания; или: есть знак, но нет паузы).

В некоторых случаях представляется возможным находить какие-то соотношения между конструкциями разговорной речи и конструкциями речи письменной (кодифицированного литературного языка), проводить аналогию между теми и другими. Но иногда такое сопоставление невозможно и приходится искать новые решения.

Например, как ставить запятые при маркерах хезитационного поиска? Это явление часто встречается в спонтанной речи, но никак не описано в правилах пунктуации, составленных для литературного языка, поскольку оно там отсутствует.

По сути, это выражения, которые диктор подставляет, когда не может сразу подобрать какое-то слово или словосочетание (это, это самое):

*А вы вагончик как, на колёсах привезли или на этой, на шаланде?*

Если диктор подобрал нужное слово, то оно появляется после вербального хезитатива, раскрывая смысл предложения, поэтому представляется уместным поставить между ними запятую. Если никаких знаков не ставить, то смысл изменится или будет потерян. Ср.: телефон этого, моряка; телефон этого моряка.

Если поиск оказывается неудачным и искомое слово не найдено, можно говорить о появлении своеобразного словосочетания с использованием вербального хезитатива:

*Потому что дорога очень это самое, в темноту нельзя ездить.*

Из данного примера видно, что запятая перед вербальным хезитативом была бы лишней, нарушая структуру синтагмы. А поскольку первая часть высказывания (до появления вербального хезитатива и включая его) строится независимо от того, найдено ли впоследствии искомое слово, представляется логичным и в случаях с удачным поиском ставить запятые аналогичным образом, то есть без запятой перед вербальным хезитативом, но с запятой между ним и искомым словом.

Другая особенность спонтанной устной речи — самоисправления говорящего.

*Да \*чи(чистим), снег чистим.*

*Я холодильник \*ща{сейчас}, холодец \*щас{сейчас} поставлю.*

По сути, дискурс прерывается и синтагма выстраивается заново. Очевидно, что при расстановке знаков препинания игнорировать подобные речевые сбои нельзя. Поэтому было принято решение в таких случаях ставить запятую.

К сожалению, не всегда всё бывает столь очевидно. При спонтанном порождении речи человек думает и говорит одновременно, поэтому связи между членами предложения часто размыты. Случается так, что фраза выстраивается до некого слова или словосочетания, а потом уже произнесённое слово или словосочетание становятся началом новой синтагмы — формируются две равнозначные части с общим центром. Яркий пример — дублирование подлежащего или сказуемого:

*Ну ты Людке сама-то ты скажи...*

*Есть опять же в любой ситуации есть но.*

По нормам литературного языка, с одной стороны, в предложении не может быть два сказуемых, не разделённых запятой, а с другой, нельзя разделять запятой непосредственно подлежащее и сказуемое (если между подлежащим и сказуемым стоит вводное слово или другой член предложения, требующий обособления, то это, по сути, не запятые, разделяющие подлежащее сказуемое, а запятые, выделяющие оборот). Кроме того, в случае дублирования подлежащего или сказуемого не



ясно, куда ставить запятую, поскольку фраза произносится как интонационно целостная, оба продублированных слова равнозначны и равновесны, и получается, нет оснований отделять одно из них запятой.

Более простой случай — нарушение привычных связей в предложении (инверсия, перестановка слов). В литературном языке инверсия тоже возможна, но не в таких масштабах. Например, придаточное предложение со смешённым союзным словом:

*Вот это вот, я чем переболела, группу дают людям.  
Ходить там, \*ничё(ничего) нету когда.*

В данном случае вопрос о месте постановки запятой не стоит, поскольку выделяется придаточная часть сложноподчинённого предложения, независимо от позиции союзного слова, однако подобная перестановка слов серьёзно усложняет задачу автоматической расстановки знаков препинания.

#### 2.4. Разработка методики оценки качества пунктуации

При разработке методики оценки качества автоматической пунктуации участники проекта столкнулись с явлением, которое было не существенно при оценке качества распознавания: трудность создания ручного эталона вследствие низкой степени согласия разных аудиторов.

Данное явление обусловлено качественным отличием письменной речи от спонтанной устной речи и проявляется как на уровне членения потока слов на предложения, так и на уровне классификации знаков.

Упомянутый выше эксперимент с членением потока слов на предложения профессиональными лингвистами показал:

- При членении потока слов аудитор в значительной мере полагается на смысл сообщения (законченность мысли);
- Степень согласия аудиторов достаточно низка.

Исходя из этого, сейчас проводится работа по введению в эталоны вариативных знаков для маркировки мест, где имеется существенная неоднозначность. Для этого надо как уточнить сами эталоны, так и изменить программу оценки качества пунктуации для поддержки вариативных знаков.

### ЗАКЛЮЧЕНИЕ

На настоящий момент достигнуты следующие результаты: SER по завершающим знакам 42%, по внутренним знакам — 42%, по всем знакам препинания — 39%. Парадокса тут нет, так как при раздельном учете внутренних и завершающих знаков ошибка замены точки на запятую (весьма частая ошибка) учитывается как две ошибки (удаление точки и вставка запятой), а при общем учете — как одна ошибка замены.

Основная проблема — сложность самой задачи. Расстановка знаков пунктуации в потоке распознанных слов есть не задача распознавания (пунктуации в звучащей речи нет, и «распознать» её там невозможно), а задача перевода — перевода с устного на письменный. Главное отличие задачи перевода от задачи распознавания — различие выразительных средств входного и выходного языков. Так как устная речь, вообще говоря, богаче письменной, выразить знаками препинания на письме то, что выражается в устной речи просодически, — очень нелегко, а иногда и принципиально невозможно.

## **ЛИТЕРАТУРА**

1. Розенталь Д.Э. Справочник по пунктуации. — М.: АСТ, 1997.
2. Стратегии членения спонтанной речи на синтаксические единицы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). — Вып. 8 (15). — М.: РГГУ, 2009. — С. 438–443.
3. Комовкина Е.П., Слепокурова Н.А. Анализ межпаузальных интервалов в спонтанном тексте: предварительные результаты // Череповецкие научные чтения-2009: Материалы Всероссийской научно-практической конференции, посвященной Дню города Череповца (2–3 ноября 2009 г.) / Ч. 1. Литературоведческие и лингвистические науки в начале XXI в. — Череповец: ГОУ ВПО ЧГУ, 2010. — С. 47–51.
4. Венцов А.В. Спонтанная речь: проблемы сегментации // IX выездная школа-семинар «Проблемы порождения и восприятия речи»: Материалы. — Череповец: ГОУ ВПО «Череповецкий государственный университет», 2010. — С. 96–101.
5. Кибрлик А.А., Подлесская В.И. (ред.) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. — М.: ЯСК, 2009.
6. Лопатин В.В. (ред.) Правила русской орфографии и пунктуации. Полный академический справочник. — М: АСТ, 2009.

## **A TASK OF AUTOMATIC PUNCTUATION IN RECOGNIZED RUSSIAN SPONTANEOUS SPEECH**

**Dmitry A. Birin,**  
the General Director of branch FGUP "scientific research Institute "Kvant",  
Saint-Petersburg

**Alexander E. Bulashevich,**  
candidate of technical Sciences, researcher of the branch of FSUE "scientific research,  
Institute "Kvant", Saint-Petersburg

**Marianna Y. Grekis,**  
engineer of FGUP "scientific research Institute "Kvant", Saint-Petersburg

### **Abstract**

The main purpose of speech recognition process is to produce readable, understandable text at output. In the Russian language it is hardly possible without punctuation marks. There is a very complicated system of punctuation rules for the Russian language. The problem is that these rules were developed for written language. Most of them are



not observed or are even broken in spontaneous speech. There are also some phenomena in spontaneous speech which are not described in the rules for literary (written) language simply because these phenomena do not meet in the written text (hesitation search, self-repairs etc.). Thus, the task is to adopt classic rules for spontaneous speech and to develop an automatic punctuation system that would be able to transform a sequence of recognized words received from spontaneous speech into a comprehensible written text. At this stage our system allows to detect sentence boundaries in most cases and placing some internal punctuation marks with a certain accuracy.

**Keywords:** spontaneous speech recognition, punctuation of recognized speech, automatic punctuation.

# **Фонетическая функция А.А. Пирогова и помехоустойчивость канала речевой коммуникации**

**Сергей Борисович Козлачков,**  
кандидат технических наук, старший преподаватель  
МГТУ им. Н.Э. Баумана, Москва

**Сергей Владимирович Дворянкин,**  
профессор, доктор технических наук, заместитель заведующего  
кафедрой информационной безопасности Финансового университета, Москва

**Надежда Валерьевна Василевская,**  
сотрудник, ФСТЭК России

## **Аннотация**

В статье рассмотрены некоторые механизмы повышения помехоустойчивости речи. Обращено внимание на определенную согласованность основных характеристик механизма слухового восприятия, параметров и характеристик речевых сигналов, а также физических процессов среды распространения акустических сигналов с сопутствующими искажениями и помехами. Показано, что помехоустойчивость речевых сигналов дополнительно обеспечивается как особенностями амплитудной модуляции, так и сложной широкополосной структурой РС, а также уникальными свойствами фонетической функции А.А. Пирогова. Приведено объяснение эффектам устойчивости разборчивости речи относительно линейных искажений.

**Ключевые слова:** разборчивость речи, речевой сигнал, искажения, помехоустойчивость, форманты модуляция, фонетическая функция, пороги слухового восприятия.

## **ВВЕДЕНИЕ**

Одной из важнейших характеристик канала связи, в т.ч. канала речевой коммуникации, является его помехоустойчивость, т.е. способность системы выполнять свои функции при наличии помех. В теории оптимального приема помехоустойчивость оценивают интенсивностью помех, при которых нарушение функций устройства ещё не превышает допустимых пределов. Чем сильнее помеха, при которой устройство остаётся работоспособным, тем выше его помехоустойчивость [1-3]. При наличии аддитивных помех в канале передачи информации помехоустойчивость может быть увеличена повышением мощности передаваемых сигналов. Однако при воздействии мультиплектирующих (неаддитивных) помех простым увеличением мощности сигнала существенного повышения помехоустойчивости достичь нельзя, и требуется радикальное изменение используемых методов, например применение помехоустойчивого кодирования либо самонастраивающегося (адаптивного) приема [1-3].



В процессе эволюционного развития канал речевой коммуникации также обрел определенные механизмы, способствующие повышению его помехоустойчивости и согласующиеся с положениями теории оптимального приема. В данной статье предпринята попытка анализа некоторых механизмов слухового восприятия по критерию помехоустойчивости.

### **НЕКОТОРЫЕ МЕХАНИЗМЫ ПОВЫШЕНИЯ ПОМЕХОУСТОЙЧИВОСТИ РЕЧИ**

Наиболее наглядно действие некоторых таких механизмов можно пояснить на примере эффекта Ломбарда — форсировании речи при воздействии на диктора аддитивных помех (например, шума) высокого уровня. В таких условиях речь диктора становится более звучной и разборчивой. При этом рост уровня (фонация) речевого сигнала (РС) составляет около 6 дБ. Также значительно возрастают и динамические показатели сигнала в т.ч. глубина модуляция спектральной огибающей РС, что проявляется в более четком выделении формант и антиформант в спектре РС.

Определенный вклад в повышение помехоустойчивости вносит и нелинейный характер ( $\log_2 x$ ) зависимости уровня выходного сигнала от входного значения, сформулированный в законе Вебера-Фехнера, что позволяет значительно расширить динамический диапазон воспринимаемых уровней сигналов, в т.ч. принимать сигналы малого уровня при высоком уровне маскирующих помех.

Существенно влияет на помехоустойчивость и нелинейная зависимость АЧХ механизма слухового восприятия (корректирующая характеристика — фильтр А), благодаря чему усиливается относительный вес среднечастотного диапазона РС. С одной стороны, это позволяет снизить маскирующее влияние разнообразных акустических низкочастотных помех, с другой — подчеркивает значимость диапазона средних частот, в пределах которого находится более 30% информативного объема речи, что обусловлено как статистическими параметрами частоты встречаемости формант, так и преобладающей динамикой дифонных переходов наиболее значимой второй формантами [4].

Отдельного внимания заслуживают иные механизмы повышения помехоустойчивости, реализующие адаптивные свойства слухового восприятия, к числу которых можно отнести акустический рефлекс, предохраняющий слух от повреждений при воздействии длительных сигналов высокого уровня [5]. Важные адаптивные свойства также проявляются и в согласованных изменениях параметров слуховых фильтров (критических полос слуха), что позволяет реализовать один из принципов оптимального приема — согласование характеристик фильтра приемника параметрам РС в значимых спектральных областях (формантах) [6].

В рамках данной статьи планировалось рассмотреть некоторые механизмы повышения помехоустойчивости, связанные с амплитудными характеристиками РС. Также хотелось обратить внимание на то, что пра-

тически все значимые характеристики слухового восприятия описываются нелинейными функциями, что с учетом адаптивных свойств слухового восприятия позволяет решать актуальные задачи оптимизации и таким образом достигать наилучших условий приема в канале речевой коммуникации.

## **ПОМЕХОУСТОЙЧИВЫЕ СВОЙСТВА ФОНЕТИЧЕСКОЙ ФУНКЦИИ А.А. ПИРОГОВА**

При распространении и приеме слуховой системой РС подвергается определенным трансформациям и искажениям, которые могут оказывать значительное влияние на его характеристики. По этой причине передавать информацию можно лишь с помощью каких-то неизменных — инвариантных структур, устойчивых по отношению к искажающим воздействиям. Информация, в т.ч. речевая, передается путем модуляции параметров сигнала — амплитуды, частоты, фазы. В речевом сигнале наиболее важными являются частотная и амплитудная модуляции, поскольку слуховое восприятие не воспринимает вариации значения фазы [7].

Чувствительность слухового восприятия характеризуется абсолютными и дифференциальными порогами. При этом физические значения абсолютных порогов чувствительности слуха почти на порядок выше, чем дифференциальных. Кроме того, слух на порядок лучше воспринимает частотную модуляцию, чем амплитудную (применительно к модуляциям сигналов гармонического типа), что обусловлено более высокой помехоустойчивостью частотной модуляции [7].

Распространение акустических колебаний (и речевых сигналов) в однородной (изотропной) среде из-за наличия различных препятствий сопровождается явлениями рассеяния, затухания и появлением отраженных волн, т.е. интерференцией. В этом случае среду распространения можно представить в виде полосного фильтра с неравномерной АЧХ. Если при этом передаточная характеристика среды распространения носит линейную зависимость, то соответствующие искажения акустического сигнала называют линейными, в ином случае — нелинейными. Из нелинейных наиболее негативное влияние на РР оказывают искажения модуляционного характера, поскольку они искажают саму модулирующую функцию, т.е. основной переносчик информации.

Наиболее распространенными и наименее значимыми (по критерию разборчивости речи) являются линейные искажения. Линейные искажения приводят только к изменению соотношения амплитуд и фаз спектральных составляющих сигнала. В литературе неоднократно отмечалось весьма низкое влияние линейных искажений на восприятие РС в целом [8, 9]. Так даже полное удаление из РС одной либо двух формант влияет только на тембр звука, однако словесная разборчивость остается высокой, что сложно объяснить с позиций формантной теории РР.

Помехоустойчивость амплитудной модуляции РС можно объяснить характером динамики изменения амплитудных и частотных параметров РС, что препятствует как формированию стоячих волн, так и образованию значительных неоднородностей плотности энергий акустического поля. Таким образом, нестационарный характер и относительно широкий диапазон полосы частот РС нивелируют влияние эффектов интерференции звуковых волн на искажение формантной структуры РС. В значительно большей степени эффекты интерференции, фазовых запаздываний и реверберации проявляются в акустике помещений большого объема и с высокой гулкостью, например концертных залах.



С другой стороны, устойчивость амплитудной модуляции РС можно объяснить, прибегнув к иной модели механизма слухового восприятия РС, разработанной А.А. Пироговым [9, 10]. По модели ученого, «каждая фонема отличается главным образом характерным для этой фонемы изменением спектрального распределения, а не самим спектральным распределением, сопутствующим данной фонеме». Исходя из этих соображений, он ввел понятие «фонетической функции речи» (или фонетическая функция А.А. Пирогова — ФФР), согласно которому фонетические элементы речи целиком определяются законом изменения спектров во времени. В качестве оценки спектральных изменений А.А. Пирогов предложил использовать разность логарифмов интенсивностей двух спектральных разрезов, взятых через интервалы, соответствующие разрешающей способности слуха во времени:

$$P(\omega, t) = \ln \left| \frac{S(\omega, t)}{S(\omega, t - \tau)} \right| \quad (1),$$

где:

$S(\omega, t)$  и  $S(\omega, t - \tau)$  — интенсивности спектральных отсчетов РС, взятые через интервал  $\tau$ , учитывающий разрешающую способность слуха во времени.

Рассмотрим характер влияния неравномерной АЧХ среды распространения РС через свойства ФФР.

В общем случае коэффициент передачи тракта распространения сигнала (полосного фильтра) можно представить в виде следующего выражения:

$$K = |K| \cdot e^{j\varphi} \quad (2),$$

где:

$|K|$  — модуль коэффициента передачи,  $\varphi$  — сдвиг фазы,  $j$  — комплексная единица.

Зависимость  $|K|$  от частоты представляет собой амплитудно-частотную характеристику (АЧХ), а  $j\varphi$  — фазочастотную характеристику (ФЧХ).

Спектр сигнала, прошедшего через такой фильтр, определяется выражением:

$$Si(\omega, t) = S(\omega, t) K \quad (3),$$

где:

$S(\omega, t)$  — спектр входного сигнала,  $K$  — коэффициент передачи электрического тракта, а  $Si(\omega, t)$  — спектр сигнала, прошедшего через электрический тракт.

При допущении, что  $|K|$  в период прохождения сигнала не претерпевает нелинейных трансформаций можно выражение (1) представить в следующем виде:

$$P(\omega, t) = \ln \left| \frac{S(\omega, t)K}{S(\omega, t - \tau)K} \right| \quad (4).$$

Из чего следует фактическая инвариантность ФФП относительно неравномерностей АЧХ ( $K$ ) передаточного тракта.

## ЗАКЛЮЧЕНИЕ

На основании изложенного можно сделать вывод о том, что человеческий слух в значительной степени инвариантен в отношении амплитудных линейных искажений, если, конечно, эти искажения не выходят за пределы артикуляторных модуляций и пределов слухового восприятия.

Помехоустойчивость амплитудной модуляции РС обеспечивается как свойствами ФФП, так и сложной широкополосной структурой РС, носящей нестационарный динамический характер, что позволяет нивелировать влияние линейных искажений среди распространения сигнала.

Представляется целесообразных проведение дальнейших исследований механизмов повышения помехоустойчивости речи.

## ЛИТЕРАТУРА

1. Гоноровский И.С. Радиотехнические цепи и сигналы. М.: Сов. радио, 1977 г. 245 с.
2. Гуткин Л.С. Теория оптимальных методов радиоприема при флуктуационных помехах. Изд. 2-е, доп. и перераб. — М.: Сов. радио, 1972. 448 с.
3. Стратонович Р.Л. Принципы адаптивного приёма. . М.: Советское радио, 1973. 144 с.
4. Покровский Н.Б. Расчет и измерение разборчивости речи. М.: Связьиздат, 1962. 392 с.
5. Алдошина И.А. Основы психоакустики // Звукорежиссер. 1999. <http://www.625-net.ru>
6. Moore B.C.J. An Introduction to the Psychology of Hearing, Sixth Edition. / B.C.J. Moore. — Leiden: Boston Brill. — 2012. 441 с.
7. Женило В.Р. Инварианты речевого сигнала и обертона голоса// Материалы Всероссийской научно-практической конференции «Естественнонаучные методы исследований в теории и практике производства судебных экономических и речеведческих экспертиз». Нижний Новгород, 2017. С. 70–81.
8. Журавлев В.Н., Архипова А.Е. Анализ противоречий теорий речеобразования и слуха с позиции идентификации информационных параметров и характеристик речевых сигналов // Информационные технологии и компьютерная инженерия. 2007. №2(9). С. 180–185.
9. Пирогов А.А. Основы Фонетической теории речи. Фонетическая функция как универсальный природный механизм кодирования-декодирования речевой информации любого происхождения // Научный Журнал Русского Физического Общества. 2001. №1–12. С. 15–28.
10. Акбулатов А.Ш., Баронин С.П., Куля В.И., Лейтес Р.Д., Муравьев В.Е., Пирогов А.А., Слуцкер Г.С., Соболев В.Н., Трофимов Ю.К. Вокодерная телефония. Методы и проблемы. М.: Связь, 1974. 536 с.



## **PHONETIC THE FUNCTION OF A.A. PIROGOV AND NOISE IMMUNITY OF THE CHANNEL SPEECH COMMUNICATION**

**Sergey B. Kozlachkov,**

*candidate of technical Sciences, senior lecturer at MSTU. N. Uh. Bauman,  
Moscow*

**Sergey V. Dvoryankin,**

*Deputy head of Department of information security of the Financial  
University, doctor of technical Sciences, Professor, Moscow*

**Nadezhda V. Vasilevskaya,**

*FSTEC employee, Moscow*

### **Abstract**

The article discusses some mechanisms to improve speech resistance of speech. Attention is drawn to a certain consistency the main characteristics of the mechanism of auditory perception, ferry-tours and characteristics of speech signals, as well as physical processes-owls of the medium of propagation of acoustic signals with accompanying distortion and interference. It is shown that noise immunity of speech signals are provided as further features of the amplitudes-Noah modulation and wideband complex structure of RS, as well as unique properties of the phonetic function of A. A. Pirogov. At-Vedeno explanation for the effects of the stability of the intelligibility of speech refers-particularly linear distortion.

**Keywords:** speech intelligibility, speech signal distortion, stability, modulation formants, phonetic function, thresholds auditory perception.



# Speech Technology

---

1-2/2017

**Chief editor:**

**Dr. Alexander A. Kharlamov**

**Composition of the editorial Board:**

**Dr., prof. Rodmonga K. Potapova**, rkpotapova@yandex.ru

**Dr., prof. Vladimir V. Golenkov Byelarus**, golen@bsuir.by, Беларусь

**Dr., prof. Valery R. Zhenilo**, zhenilo@yandex.ru

**Ph.D Yury N. Zhigulevtsev**, ynzh@mail.ru

**Dr. Alexey A. Karpov**, karpov@iias.spb.ru

**Dr., prof. Olga F. Krivnova**, okrivnova@mail.ru

**Saule A. Kudubaeva Kazakhstan**, saule\_58@mail.ru, Казахстан

**Ph. D Alexey M. Kushnir**, kushnir-narobr@yandex.ru

**Ph. D Dmitry A. Kushnir**, kushdal@yandex.ru

**Dr. Boris M. Lobanov Byelarus**, lobanov@newman.bus-net.by, Беларусь

**Dr. Elena E. Lyakso**, lyakso@gmail.com

**Dr. Eugeny M. Maximov**, maximovem@inbox.ru

**Dr. Yury N. Matveev**, matveev@mail.ifmo.ru"ru

**Dr., prof. Roman V. Myescheryakov**, mrv@ieee.org

**Dr., prof. Alexander A. Petrovsky Byelarus**, palex@bsuir.by

**Ph. D Yury N. Romashkin**, romayn@yandex.ru

**Dr., prof. Andrey L. Ronzhin**, ronzhin@iias.spb.su

**Nicolay N. Sazhok Ukraine**, sazhok@gmail.com, Украина

**Acad. TAS, Dr., prof. Djavdet. Sh. Suleymanov**, профессор, alsu\_73@list.ru

**Ph.D Vladimir Ya. Chuchupal**, v.chuchupal@gmail.com

**Milos Zelezny Czech Republic**, zelezny@kky.zcu.cz



# Памяти С.В. Ёлкина

*С глубоким прискорбием сообщаем о трагической гибели 17 февраля 2018 года Сергея Владимировича Ёлкина, талантливого отечественного учёного и преподавателя, посвятившего более тридцати лет научной и педагогической деятельности проблеме развития сильного мышления, в том числе посредством универсального языка междисциплинарного общения Диал.*



Сергей Владимирович Ёлкин родился в 1958 году в Москве, окончил факультет Экспериментальной и теоретической физики Московского инженерно-физического института, аспирантуру (1989–1992). Кандидат физико-математических наук. Работал и преподавал в МИФИ на разных должностях: ассистент, ст. преподаватель, доцент (1996–2012). Начальник лаборатории Института прикладной математики имени М.В. Келдыша РАН (2003–2010), зам. директора Института экономических стратегий РАН по научной работе (2011–2012).

Автор и соавтор десяти монографий по физике, машинной и теоретической лингвистике, экономике, 104 публикаций<sup>1</sup>, 5 авторских свидетельств на изобретения, 3 авторских свидетельства на компьютерные программы; 7 публикаций на английском языке. Создатель алгебры Y-чисел, основы которой изложены в одной из ранних работ автора — «К вопросу об информационной физике» в 1997 году, а подходы освещены уже в 1994 м, и развиты в последующих более строгих статьях и книгах<sup>2</sup>.

Вёл научные исследования по широкому спектру направлений и смежным областям: физика детекторов, ультразвук, искусственные языки, теоретическая лингвистика и искусственный интеллект, теория чисел, философия, экономика, «бионические нейронные сети». Руководитель и исполнитель четырёх грантов РФФИ. Один из разработчиков [с В.В. Куликовым и Д.А. Гавриловым] универсального языка междисциплинарного общения, языка-транслятора и классификатора изобретательских идей Диал<sup>3</sup> [с нач. 1980 х гг.]. Организатор и ведущий

<sup>1</sup> Страница в РИНЦ пока охватывает только половину всех научных публикаций С.В. Ёлкина и сейчас лишь отчасти отражает его многогранную научную деятельность: [https://elibrary.ru/author\\_items.asp?authorid=152641](https://elibrary.ru/author_items.asp?authorid=152641)

<sup>2</sup> Ёлкин С.В. Математическая онтология. Диалектика-симметрия-числа-семантический язык. — Saarbrucken: LAP LAMBERT Academic Publishing GmbH & Co. KG, 2012.

<sup>3</sup> Куликов В.В., Ёлкин С.В., Гаврилов Д.А. Универсальный межзвёздный язык диал как средство научного общения и производства открытых // Труды XXV Чтений, посвященных разработке научного наследия и развитию идей К.Э. Циолковского. Секция «К.Э. Циолковский и философские проблемы освоения космоса». Симпозиум «Проблемы поиска жизни во Вселенной». Калуга, 11–14 сентября 1990. М.: ИИЕТ АН СССР. 1991;

экспериментальных групп школьников, студентов и аспирантов по интеллектуальному тренингу (с 1986 г.). 8 лет вёл в Интернет занятия по развитию творческих способностей в рамках сетевого Университета русского альтруизма.

Несколько лет читал авторский курс по методологии развития творческого мышления на кафедре «Экономика и менеджмент в промышленности» НИЯУ МИФИ.

С декабря 2012 года по 2015 г. советник-эксперт ООО «ЛУКОЙЛ-Инжениринг». Соавтор дистанционно-очного курса «Инженерно-техническое творчество в нефтегазовой отрасли». В 2013–2014 гг. осуществлял подготовку и триггерное сопровождение команд молодых специалистов «ПечорНИПИнефть», «ПермНИПИнефть», «КогалымНИПИнефть» и др., читал курс избранных лекций для экспертов-аналитиков профильных центров «ЛУКОЙЛ» по методам морфологического анализа и эвристики. Проводил интеллектуальные мероприятия, направленные на творческое развитие и «апгрейд» ИТР и менеджеров.

Соавтор пяти монографий<sup>4</sup>, посвящённых разным аспектам теории творчества и развитию сильного мышления. Последняя из них увидела свет в начале февраля 2018 года, за неделю до гибели Сергея Владимировича Ёлкина. Он был полон планами на будущее

С 2017 года на общественных началах в качестве эксперта «Фонда содействия технологиям XXI века» работал над двумя широкомасштабными авторскими проектами – постановкой эксперимента по проверке гипотезы Сепира-Уорфа<sup>5</sup> с применением универсального языка Диал, а также проведением на постоянной основе всероссийской Олимпиады по развитию творческого мышления.

Выражаем искренние соболезнования родным, близким и друзьям покойного в связи с этой безвременной и невосполнимой утратой!

Мы надеемся, что научное наследие Сергея Владимировича Ёлкина будет сохранено последователями, и современники ещё оценят его по достоинству.

---

Ёлкин С.В., Гаврилов Д.А. Сильное мышление как результат диалектически выстроенного языка / Материалы IV Международной научно-практической конференции «Непознанное. Традиции и современность» 18 октября 2014 г. // Информационно-аналитический вестник «Аномалия». 2014. № 4. С. 25–29.

<sup>4</sup> Гаврилов Д.А., Ёлкин С.В. Избранные лекции по курсу «Начала сильного мышления». Часть 1: Эвристика и развитие творческого воображения / Фонд содействия технологиям XXI века. М.: Издатель Воробьёв А.В., 2018;

Ёлкин С.В., Гаврилов Д.А. Инженерно-техническое творчество в нефтегазовой отрасли. Избранные лекции курса и сборник задач. М.: Центр стратегической конъюнктуры, 2014;

Латыпов Н.Н., Гаврилов Д.А., Ёлкин С.В. Турбулентное мышление. Зарядка для интеллекта / Под ред. А.А. Вассермана. М.: АСТ, 2013;

Латыпов Н.Н., Ёлкин С.В., Гаврилов Д.А. Самоучитель игры на извилинах / Под ред. А.А. Вассермана. М.: АСТ, 2012;

Латыпов Н.Н., Ёлкин С.В., Гаврилов Д.А. Инженерная эвристика / Под ред. А.А. Вассермана. М.: Астрель, 2012.

<sup>5</sup> Ёлкин С.В., Гаврилов Д.А. Как проверить гипотезу Сепира-Уорфа: приглашаем к эксперименту // Информационно-аналитический вестник «Аномалия». 2015. №2. С.46–48.



## Contents

<i>Vladimir J. Chuchupal</i>	
<b>Implicit pronunciation variation model for automatic speech recognition .....</b>	<b>3</b>
<i>Maksim I. Vashkevich, Iliy S. Azarov, Aleksandr A. Petrovsky</i>	
<b>Estimation of instantaneous fundamental frequency of speech based on multirate signal processing .....</b>	<b>12</b>
<i>Stanislav A. Krejci, Olga F. Krivnova, Ekaterina A. Tikhonov</i>	
<b>Immunity of syllabic tables in perception of speech in noise ..</b>	<b>25</b>
<i>Ekaterina G. Solonina</i>	
<b>Acoustical and perceptual features of coarticulatory nasalization of russian vowels .....</b>	<b>37</b>
<i>Elena E. Lyakso, Olga V. Frolova, Alexey S. Grigoriev, Victor A. Gorodny</i>	
<b>Comparative analysis of the voice and speech features of children typically developing, with autism spectrum disorders, Down syndrome and mental retardation .....</b>	<b>50</b>
<i>Maksim I. Vashkevich, Iliy S. Azarov, Aleksandr A. Petrovsky</i>	
<b>Speech enhancement in a smartphone-based hearing aid .....</b>	<b>63</b>
<i>Olga V. Frolova , Shanbigi G. Bedalova, Elena E. Lyakso</i>	
<b>Speech development of preschool children with developmental disorders growing up in an orphanage .....</b>	<b>82</b>
<i>Dmitry A. Birin, Alexander E. Bulashevich, Marianna Y. Grekis</i>	
<b>A task of automatic punctuation in recognized russian spontaneous speech .....</b>	<b>94</b>
<i>Sergey B. Kozlachkov, Sergey V. Dvoryankin, Nadezhda V. Vasilevskaya</i>	
<b>Phonetic the function of A.A. Pirogov and noise immunity of the channel speech communication .....</b>	<b>105</b>



**Индекс: 62203  
ISSN 2305-8129**