УДК 004.934
DOI: 10.15593/2224-9397/2021.2.03

А.К. Алимурадов

Пензенский государственный университет, Пенза, Россия

ПОМЕХОУСТОЙЧИВЫЙ СПОСОБ СЕГМЕНТАЦИИ «РЕЧЬ/ПАУЗА» НА ОСНОВЕ МЕТОДА ДЕКОМПОЗИЦИИ НА ЭМПИРИЧЕСКИЕ МОДЫ

Актуальность и цели: сегментация речь/пауза является одной из важнейших задач обработки в речевых приложениях и представляет собой обнаружение границ начала и окончания вокализованной, невокализованной речи и пауз. Точность сегментации особенно важна при анализе зашумленной речи, так как от уровня окружающего шума зависит работоспособность практически всех речевых приложений. Целью работы является повышение эффективности сегментации речь/пауза зашумленных речевых сигналов на основе метода декомпозиции на эмпирические моды. **Материалы и методы:** в работе использовалась уникальная технология адаптивного разложения нестационарных сигналов – улучшенная полная множественная декомпозиция на эмпирические моды с адаптивным шумом. Программная реализация способа была выполнена в среде математического моделирования © MatLab (MathWorks). **Результаты:** разработан помехоустойчивый способ сегментации речь/пауза на основе применения декомпозиции на эмпирические моды на этапе предварительной обработки и последующего анализа количества пересечения сигнала через нулевую ось и кратковременной энергии речи. Особенностью разработанного сопосба является формирование набора новых исследуемых сигналов, содержащих максимально достоверную информацию о границах начала и окончания информативных участков зашумленной речи. **Выводы:** проведено исследование, в рамках которого оценивалось влияние метода декомпозиции и длительности исследуемых фрагментов на эффективность сегментации зашумленной речи. В соответствии с результатами исследований отмечается снижение ошибок 1-го и 2-го рода в среднем на 2,3 и 2,6 % соответственно при разных отношениях сигнал/шум (от 20 до –5 дБ с шагом 5 дБ) зашумленной речи. Помехоустойчивый способ сегментации речь/пауза может успешно тестироваться в интеллектуальных системах оценки психоэмоционального состояния человека в реальных условиях «агрессивной» шумовой обстановки.

Ключевые слова: обработка речи, зашумленные речевые сигналы, сегментация речи, вокализованная и невокализованная речь, декомпозиция на эмпирические моды.

A.K. Alimuradov

Penza State University, Penza, Russian Federation

EMD-BASED NOISE-ROBUST METHOD FOR "SPEECH/PAUSE" SEGMENTATION

Relevance and goals: Speech/pause segmentation is one of the most important tasks in speech applications, being detection of the boundaries of the beginning and the end of voiced and unvoiced speech, and pauses. Segmentation accuracy is especially important when analyzing noisy speech, since the performance of almost all speech applications depends on the level of ambient noise. The aim of this work is to improve the efficiency of speech/pause segmentation of noisy speech signals based on the method of empirical mode decomposition. **Materials and methods:** A unique technology for adaptive decomposition of non-stationary signals, namely, the improved complete ensemble empirical mode decomposition with adaptive noise, has been used in the work. The software implementation of the method was performed in © MATLAB (MathWorks) mathematical modeling environment. **Results:** A noise-robust method for speech/pause segmentation based on the empirical mode decomposition at the preprocessing stage, and subsequent analysis of zero-crossing rate and short-term speech energy, has been developed. A characteristic feature of the developed method is the formation of a set of new investigated signals containing the most reliable information about the boundaries of the beginning and the end of noisy speech informative sections. **Conclusions:** Research to assess the influence of the decomposition method, and the duration of the studied signal fragments on the efficiency of noisy speech segmentation has been done. In accordance with the research results, there is a decrease in the 1st and 2nd kind errors by an average of 2.3% and 2.6%, respectively, at different signal-to-noise ratios (from 20 to -5 dB with a step of 5 dB) of noisy speech. Noise-robust method for speech/pause segmentation can be tested under real conditions of "aggressive" noise environment in intelligent assessment systems of a human psycho-emotional state.

Keywords: speech processing, noisy speech signals, speech segmentation, voiced and unvoiced speech, empirical mode decomposition.

Введение

Сегментация «речь/пауза» представляет собой точное обнаружение границ начала и окончания информативных участков речи: пауз, вокализованной и невокализованной речи. Точность сегментации особенно важна при анализе зашумленных речевых сигналов, так как от уровня окружающего шума зависит работоспособность практически всех речевых приложений.

В зависимости от назначения речевых приложений информативные участки речи имеют разную значимость. В приложениях распознавания речи и голосового управления паузы в анализируемых речевых сигналах являются неинформативными и удаляются на этапе предварительной обработки. Объясняется это тем, что основной набор информативных параметров дикторов (амплитудно-частотных, спектрально-временных, кепстральных и др.) сосредоточен в вокализованной

и невокализованной речи. В приложениях, предназначенных для выявления патологии голоса человека, нарушений моторики речевого аппарата, паузы в речевых сигналах максимально информативны. Например, при анализе скорости, ускорения и энтропии распределения вокализованной и невокализованной речи, пауз, а также при анализе средней продолжительности пауз в слитной речи.

На сегодняшний день задача сегментации «речь/пауза» решается разными способами, которые можно разделить на частотные и временные [1]. Способы сегментации во временной области основаны на определении характерных точек речевого сигнала с последующим использованием их для анализа. В качестве характерных точек могут быть использованы экстремумы (максимумы, минимумы) и моменты пересечения нулевой оси времени функцией сигнала. Недостатком способов обработки во временной области является неоднозначность выделения характерных точек, вызванная шумами и смещениями нулевого уровня. К временным относятся способы на основе анализа количества пересечения сигнала через нулевую ось (Zero-Crossing Rate, ZCR) [2, 3], отклонения автокорреляционной функции (Autocorrelation Function, ACR) [4, 5], кратковременной энергии (Short Time Energy, STE) [6, 7], а также одномерного расстояния Махаланобиса (One Dimensional Mahalanobis Distance, ODMD) [8].

Способы сегментации в частотной области основаны на использовании всех отсчетов данных, зарегистрированных в речевом сигнале. Речевые сигналы имеют специфический частотный состав и занимают характерные спектральные области. Использование способов в частотной области позволяет обрабатывать речевые сигналы с достаточно высокой точностью. К недостаткам обработки в частотной области относятся низкая адаптивность к локальным свойствам сигналов, недостаточно высокое спектральное разрешение и сравнительно большие вычислительные затраты. К частотным относятся способы на основе анализа мел-частотных кепстральных коэффициентов (Mel-Frequency Cepstral Coefficients, MFCC) [9, 10] и линейно-частотных кепстральных коэффициентов (Linear-Frequency Cepstral Coefficients, LFCC) [11, 12].

Частотные способы сегментации «речь/пауза» в сравнении с временными обладают большей помехоустойчивостью. Однако наибольшую практическую популярность получили временные способы, так как для их реализации необходима меньшая вычислительная мощ-

ность. Поэтому перед исследователями всегда стоит выбор между помехоустойчивостью и быстродействием речевых приложений.

В данной статье представлен помехоустойчивый способ сегментации речь/пауза на основе метода декомпозиции на эмпирические моды (ДЭМ). ДЭМ представляет собой частотно-временной способ обработки, включающий в себя все преимущества временного и частотного анализов с минимальными проявлениями их недостатков. ДЭМ в помехоустойчивом способе применяется на этапе предварительной обработки исходного зашумленного речевого сигнала. По результатам предварительной обработки формируется набор новых исследуемых сигналов, содержащих максимально достоверную информацию о границах начала и окончания участков вокализованной, невокализованной речи и пауз. Разработанный помехоустойчивый способ сегментации речь/пауза предназначен для применения в интеллектуальных системах оценки психоэмоционального состояния человека в условиях зашумленной обстановки [13].

Данная статья является результатом научной работы, посвященной разработке эффективных алгоритмов и способов обработки речевых сигналов на основе новых частотно-временных методов анализа [14–16].

Структурно статья состоит из семи разделов. Второй и третий разделы посвящены краткому обзору наиболее популярных способов сегментации «речь/пауза», а также методу и разновидностям ДЭМ. Четвертый, пятый и шестой разделы посвящены описанию и исследованию помехоустойчивого способа, а также анализу результатов исследований. Последний раздел посвящен кратким выводам и перспективам дальнейших исследований.

1. Сегментация «речь/пауза»

Способы сегментации «речь/пауза» на основе анализа ZCR и STE применяются в речевых приложениях ограниченно. Ограничения связаны с выбором и обоснованием корректных пороговых значений, соответствующих вокализованной, невокализованной речи и паузам.

Функция ZCR основана на сравнении знаков соседних дискретных отсчетов времени и определяется по следующей формуле:

$$ZCR_s = 0{,}5 \sum_{n=-} |sgn(x(n)) - sgn(x(n-1))| w(s-n), \qquad (1)$$

где $x(n)$ – исследуемый сигнал; n – дискретный отсчет времени; s – номер фрагмента; $sgn(x)$ – знаковая функция ($sgn(x) = 1$ при $x \geq 0$ и $sgn(x) = -1$ при $x \leq 0$); w – функция анализируемого окна.

Для прямоугольного анализируемого окна формула (1) принимает следующий вид:

$$ZCR_s = 0{,}5 \sum_{n=1}^{N-1} |sgn(x(s-1)N + n + 1) - sgn(x(s-1)N + n)|, \quad (2)$$

где N – количество дискретных отсчетов в исследуемом фрагменте.

Функция STE представляет собой сумму квадратов амплитуд дискретных отсчетов времени и определяется по следующей формуле:

$$E_s = \sum_{n=-} [x(n)w(s-n)]^2. \quad (3)$$

Для прямоугольного анализируемого окна формула (3) принимает следующий вид:

$$E_s = \sum_{n=1}^{N} [x(s-1)N + n]^2. \quad (4)$$

Способ сегментации речь/пауза на основе анализа ZCR построен на предположении, что количество пересечений функции сигнала через нулевую ось для пауз с фоновым шумом больше по сравнению с вокализованной, невокализованной речью. Аналогично построен способ на основе анализа STE: энергия вокализованной, невокализованной речи больше, чем энергия пауз с фоновым шумом. Однако данные предположения не совсем корректные, так как остается нерешенным вопрос: насколько текущие значения ZCR и STE должны быть больше, чем пороговые, для корректной сегментации информативных участков. Кроме того, известно, что пороговые значения могут варьироваться для каждого конкретного анализируемого речевого сигнала. В работе [17] авторами была предпринята попытка выбрать и обосновать пороговые значения ZCR и STE, соответствующие вокализованной, невокализованной речи и паузам. В соответствии с выводами в работе [17] точность составила 65 % в сравнении с сегментацией, осуществленной вручную.

Способ сегментации «речь/пауза» на основе анализа ODMD построен на статистических свойствах фонового шума [8, 18]. В соответствии с физиологией воспроизведения речи человек перед произношением выдерживает вынужденную начальную паузу, длительностью не менее 200 мс, которая соответствует фоновому шуму. Предполагается, что фоновый шум, регистрируемый во время начальной паузы, имеет

Гауссовский характер, а остальные информативные участки вокализованной и невокализованной речи имеют другое распределение. В этом случае функция плотности вероятности распределения фонового шума является критерием сегментации «речь/пауза». Таким образом, решается проблема выбора и обоснования корректных пороговых значений.

В основе вычисления ODMD лежит функция плотности вероятности нормального распределения:

$$p(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, \quad (5)$$

где μ и σ – математическое ожидание и стандартное отклонение независимых случайных величин *y*.

Как известно, кривая функции плотности вероятности нормального распределения имеет форму симметричного колокообразного импульса. Независимые случайные величины имеют тенденцию группироваться около среднего значения. Пик нормального распределения соответствует $y = \mu$, а ширина пропорциональна стандартным отклонениям σ. Аналитическое выражение ODMD имеет следующий вид:

$$r = \frac{|y-\mu|}{\sigma}, \quad (6)$$

где выражение $|y - \mu|$ является естественной мерой расстояния от *y* к среднему значению μ.

В работе [19] представлен подробный сравнительный анализ результатов сегментации речь/пауза, полученных с помощью способов на основе анализа ZCR, STE и ODMD. В соответствии с выводами в работе [19] способ на основе анализа ODMD эффективнее для отдельных словосочетаний, чем способы на основе анализа ZCR и совместного анализа ZCR и STE на 5,6 и 13,18 % соответственно. Для слитной речи повышение эффективности составляет на 8,88 и 9,59 % соответственно.

В соответствии с вышеупомянутым математическим описанием проведены исследования способов, основанных на анализе ZCR, STE, совместном анализе ZCR и STE, а также на анализе ODMD. В табл. 1 представлены усредненные данные ошибок 1-го (α) и 2-го рода (β), полученные по результатам сегментации вышеупомянутыми способами.

Таблица 1

Усредненные данные ошибок 1-го и 2-го рода, полученные по результатам сегментации способами на основе анализа ZCR, STE, совместного анализа ZCR и STE, а также анализа ODMD

Способ сегментации речь/пауза	Ошибки 1-го и 2-го рода, %	
	α	β
Способ на основе анализа ODMD	21,97	0,89
Способ на основе анализа ZCR	23,11	3,02
Способ на основе анализа STE	10,53	3,2
Способ на основе совместного анализа ZCR и STE	7,32	5,33

Наилучший результат сегментации участков речи с ошибкой 1-го рода 7,32 % достигается способом на основе совместного анализа ZCR и STE. По отдельности способы на основе анализа ZCR и STE обеспечивают неудовлетворительный и средний результаты сегментации – 23,11 и 10,53 % соответственно. Повышение точности обнаружения границ начала и окончания вокализованной и невокализованной речи при совместном анализе ZCR и STE объясняется совокупной эффективностью кратковременного анализа фрагментов с отличающимися свойствами. Наихудший результат сегментации с ошибкой 1-го рода 23,11 % отмечается у способа на основе анализа ZCR. Объясняется это трудностью отличить фрагменты пауз с фоновым шумом и фрагменты, содержащие слабые шумные щелевые и фрикативные звуки. Например, у тихо произнесенных звуков «ш» и «с» значение ZCR близко к значению, соответствующему паузе с фоновым шумом.

Наилучший результат сегментации участков пауз с ошибкой 2-го рода 0,89 % достигается способом на основе анализа ODMD. Наихудший результат с ошибкой 2-го рода 5,33 % отмечается у способа на основе совместного анализа ZCR и STE. В соответствии с полученными данными ошибок 1-го и 2-го рода в табл. 1 сделан вывод, что целесообразной является разработка помехоустойчивого способа сегментации речь/пауза с применением ДЭМ на этапе предварительной обработки на основе совместного анализа ZCR и STE.

2. Декомпозиция на эмпирические моды

Среди наиболее известных способов обработки в частотно-временной области можно выделить способы, основанные на преобразовании Фурье и вейвлет-преобразовании [4].

Результатом преобразования Фурье является амплитудно-частотный спектр, в рамках которого можно определить присутствие определенной частоты в анализируемом сигнале. Недостаток преобразования заключается в том, что частотные составляющие не могут быть локализованы во времени, что накладывает ограничения на применимость к ряду задач по обработке речевых сигналов.

Вейвлет-преобразование обладает самонастраивающимся подвижным частотно-временным окном и одинаково хорошо выявляет как низкочастотные, так и высокочастотные характеристики сигнала на разных временных масштабах. В этом случае сигнал анализируется путем разложения по базисным функциям, полученным из некоторого прототипа путем сжатий, растяжений и сдвигов. Функция «прототип» называется анализирующим вейвлетом. Благодаря хорошей приспособленности к анализу нестационарных сигналов вейвлет-преобразование стало мощной альтернативой преобразованию Фурье. Недостатком вейвлет-преобразования является необходимость априорной информации об исследуемом сигнале для правильного подбора анализирующего вейвлета.

Подробный анализ перспективных методов частотно-временной обработки выявил способы на основе ДЭМ [20]. ДЭМ представляет собой адаптивный метод разложения нестационарных данных, основным преимуществом которого является полная адаптивность. Базисные функции, используемые для разложения, извлекаются непосредственно из исследуемого сигнала и позволяют учитывать только ему свойственные особенности. В рамках декомпозиции многократно осуществляется процесс просеивания, в результате которого исследуемый сигнал раскладывается на эмпирические моды (ЭМ) и конечный неделимый остаток. Процесс просеивания включает в себя обязательное выполнение следующих задач:

Задача 1. Определение среднего значения верхней и нижней огибающих исследуемого сигнала.

Задача 2. Вычитание среднего значения верхней и нижней огибающих из исследуемого сигнала.

Задача 3. Многократное повторение первой и второй задачи до тех пор, пока среднее значение не будет близко к нулю.

Задача 4. Выделение сигнала первой ЭМ, у которого среднее значение верхней и нижней огибающих максимально приблизилось к нулю в соответствии с критерием останова.

Задача 5. Вычитание первой ЭМ из исследуемого сигнала.

Задача 6. Повторение с первой по пятую задачи до тех пор, пока не будет получен монотонный сигнал (конечный неделимый остаток), из которого невозможно извлечь ни одну ЭМ.

Аналитическое выражение ДЭМ выглядит следующим образом:

$$x(n) = \sum_{i=1}^{I} IMF_i(n) + r_I(n), \quad (7)$$

где $IMF_i(n)$ – сигнал ЭМ; i – номер ЭМ; I – количество ЭМ; $r_I(n)$ – конечный неделимый остаток.

Метод ДЭМ впервые был представлен в 1998 г. [20]. На сегодняшний день известны различные методы декомпозиций: множественная ДЭМ – МДЭМ (2009 г.) [21], комплементарная МДЭМ – КМДЭМ (2010 г.) [22], полная МДЭМ с адаптивным шумом – ПМДЭМАШ (2011 г.) [23] и улучшенная ПМДЭМАШ (2014 г.) [24]. Наиболее адаптивным методом для обработки речи является метод улучшенной ПМДЭМАШ. Особенность улучшенной декомпозиции заключается в добавлении к исследуемому сигналу контролируемого белого шума малой амплитуды для создания новых нулей и экстремумов (локальных особенностей) функции сигнала. Создание новых локальных особенностей позволяет устранить известные недостатки предыдущих методов декомпозиции: эффект смешивания ЭМ; остаточный шум; неполное разложение; неинформативные «паразитные» ЭМ, выделяемые на ранних этапах декомпозиции.

Математический аппарат методов декомпозиции с добавлением шума выглядит следующим образом:

$$x_j(n) = x(n) + w_j(n), \quad (8)$$

где $x_j(n)$ – зашумленный речевой сигнал белым шумом; $w_j(n)$ – белый шум; $j = 1, 2, \ldots; J$ – реализации белого шума:

$$x_j(n) = \sum_{i=1}^{I} IMF_{j,i}(n) + r_{j,I}(n), \quad (9)$$

$$IMF_i(n) = \sum_{j=1}^{J} \frac{IMF_{j,i}(n)}{J}, \quad (10)$$

$$r_I(n) = \sum_{j=1}^{J} \frac{r_{j,I}(n)}{J}, \quad (11)$$

3. Описание помехоустойчивого способа сегментации «речь/пауза»

Помехоустойчивый способ сегментации «речь/пауза» разработан на основе совместного анализа ZCR и STE с применением ДЭМ на этапе предварительной обработки зашумленной речи. По результатам предва-

рительной обработки осуществляется формирование набора новых исследуемых сигналов, содержащих максимально достоверную информацию о границах начала и окончания информативных участков зашумленной речи. На рис. 1 структурно представлен помехоустойчивый способ сегментации «речь/пауза». Блоки 1–4 представляют этап предварительной обработки. Непосредственная сегментация «речь/пауза» на основе совместного анализа ZCR и STE реализована в блоках 5–10. Блоки 11 и 12 не относятся к помехоустойчивому способу и предназначены для постобработки ошибок сегментации «речь/пауза», а также сравнения результатов с сегментацией, осуществленной вручную. Рассмотрим подробнее некоторые этапы обработки предлагаемого помехоустойчивого способа.

Рис. 1. Структура помехоустойчивого способа сегментации «речь/пауза» на основе совместного анализа ZCR и STE с применением ДЭМ на этапе предварительной обработки

Блок 1. Фрагментирование представляет собой линейное разделение речевого сигнала на отрезки (фрагменты) равной длительности. Фрагментирование основано на кратковременном анализе, в рамках которого фрагменты обрабатываются так, как если бы они были короткими речевыми сигналами с отличающимися свойствами. В соответствии со структурой помехоустойчивого способа от длительности исследуемых фрагментов зависит результат последующей декомпозиции. В работе [25] авторами представлены результаты исследований влияния длительности анализируемых речевых сигналов на частотно-избирательные свойства различных методов декомпозиции. В соответствии с полученными результатами в работе [25] сделан вывод, что для корректного частотно-временного анализа длительность исследуемых фрагментов должна быть от 10 до 50 мс. Фрагментирование речевого сигнала осуществляется по следующим формулам:

$$S = \frac{x(n)}{L}, \tag{12}$$

где S – количество фрагментов в исследуемом речевом сигнале $x(n)$ (с округлением в меньшую сторону); L – количество дискретных отсчетов времени в одном фрагменте.

$$x_{s+1}(n) = x[(s \cdot L) + 1 : (s + 1) \cdot L], \tag{13}$$

Блок 2. В помехоустойчивом способе сегментации «речь/пауза» применяются следующие методы декомпозиции (блок 2): ДЭМ, МДЭМ и улучшенная ПМДЭМАШ. Использование только двух методов декомпозиции с добавлением шума объясняется тем, что МДЭМ и КМДЭМ, а также ПМДЭМАШ и улучшенная ПМДЭМАШ аналогичны с точки зрения процесса просеивания ЭМ. Отличительной особенностью улучшенной ПМДЭМАШ от других методов декомпозиции с добавлением шума является локальное разложение белого шума на шумовые моды параллельно с разложением исследуемого сигнала. Использование шумовых мод в качестве добавляемого контролируемого белого шума на каждом этапе процесса просеивания обеспечивает полноту разложения [23].

В помехоустойчивом способе применяются следующие настройки методов декомпозиции с добавлением шума: стандартное отклонение амплитуды шума от амплитуды сигнала – не более 20 %; количество реализаций белого шума – 100; допустимое максимальное количество просеивающих итераций – 50; отношение стандартных отклонений сигнала и шума на всех этапах процесса просеивания ЭМ неизменное (для метода улучшенной ПМДЭМАШ).

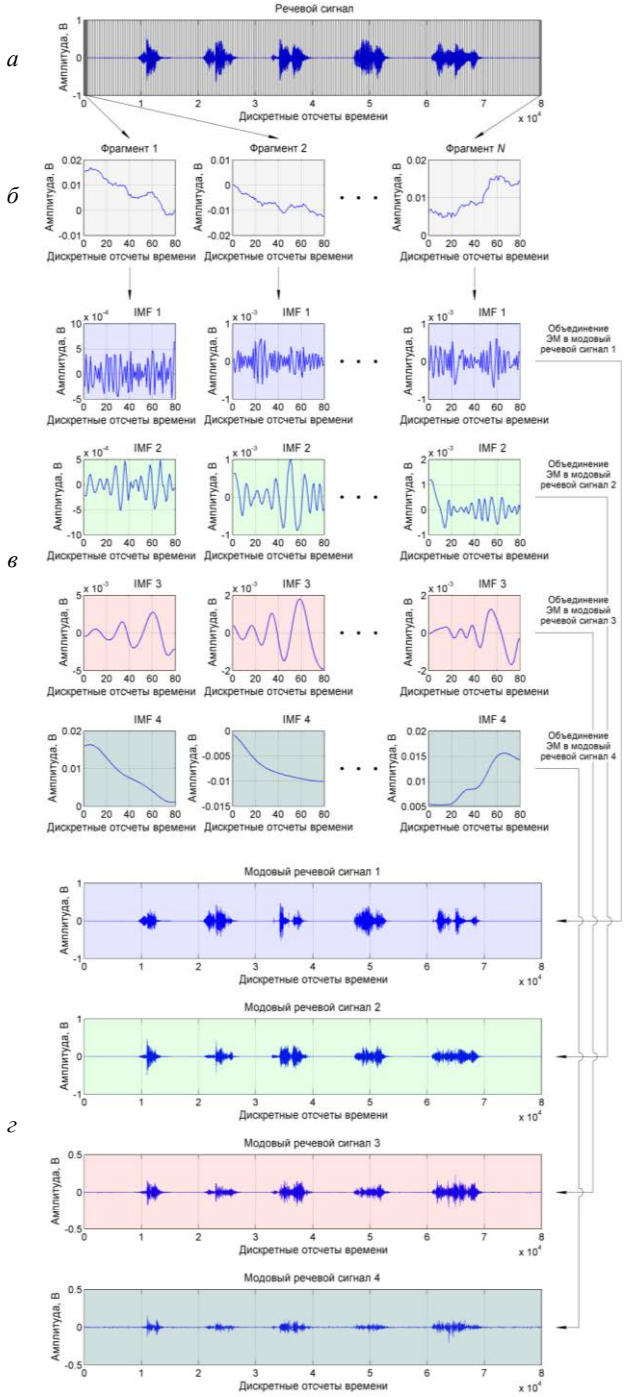

Рис. 2. Объединение ЭМ-фрагментов исходного речевого сигнала в новые модовые речевые сигналы: *а* – исходный речевой сигнал, *б* – фрагменты речевого сигнала, *в* – ЭМ фрагментов, *г* – новые модовые речевые сигналы

Блок 3. В соответствии с результатом декомпозиции каждый фрагмент исходного зашумленного речевого сигнала представлен набором ЭМ. Для оптимизации вычислительных затрат помехоустойчивого способа осуществляется объединение ЭМ-фрагментов исходного сигнала в новые модовые речевые сигналы:

$$xmode_i(n) = \sum_{s=1}^{S} IMF_{s,i}[(s \cdot L) + 1 : (s+1)L], \qquad (14)$$

где $xmode_i(n)$ – модовый речевой сигнал; $i = 1, 2, \ldots, I$ – количество ЭМ в наборах для каждого фрагмента s.

Количество сформированных модовых речевых сигналов зависит от количества ЭМ, полученных для каждого фрагмента. На рис. 2 визуально представлен процесс формирования четырех модовых речевых сигналов.

Блок 4. Формирование информативных сигналов на основе комбинирования четырех модовых речевых сигналов осуществляется по следующей формуле:

$$xinfo(n) = a \cdot x(n) + b \cdot xmode_1(n) + c \cdot xmode_2(n) + d \times \\ \times xmode_3(n) + e \cdot xmode_4(n), \qquad (15)$$

где a, b, c, d, e – коэффициенты, определяющие участие исходного и модовых речевых сигналов в формировании информативных сигналов (коэффициенты принимают только три значения: –1, 0, 1)

Суть формирования информативных сигналов заключается в поиске уникального сигнала, содержащего максимально достоверную информацию о границах начала и окончания участков вокализованной, невокализованной речи и пауз. В соответствие с проведенными исследованиями [26] в табл. 2 представлены оптимальные информативные сигналы.

Таблица 2

Оптимальные информативные сигналы

Информативный сигнал	Значение коэффициентов	Комбинирование модовых речевых сигналов
$xinfo_1$	a=0, b=1, c=1, d=0, e=0	$xmode_1 + xmode_2$
$xinfo_2$	a=0, b=1, c=1, d=1, e=0	$xmode_1 + xmode_2 + xmode_3$
$xinfo_3$	a=0, b=1, c=1, d=1, e=1	$xmode_1 + xmode_2 + xmode_3 + xmode_4$
$xinfo_4$	a=0, b=0, c=1, d=1, e=0	$xmode_2 + xmode_3$
$xinfo_5$	a=0, b=0, c=1, d=1, e=1	$xmode_2 + xmode_3 + xmode_4$
$xinfo_6$	a=0, b=0, c=1, d=0, e=1	$xmode_3 + xmode_4$
$xinfo_7$	a=0, b=1, c=0, d=1, e=1	$xmode_1 + xmode_3 + xmode_4$
$xinfo_8$	a=0, b=1, c=1, d=0, e=1	$xmode_1 + xmode_2 + xmode_4$
$xinfo_9$	a=1, b=–1, c=0, d=0, e=0	$x - xmode_1$
$xinfo_{10}$	a=1, b=0, c=–1, d=0, e=0	$x - xmode_2$

Окончание табл. 2

Информативный сигнал	Значение коэффициентов	Комбинирование модовых речевых сигналов
$xinfo_{11}$	a=1, b=0, c=0, d=–1, e=0	$x - xmode_3$
$xinfo_{12}$	a=1, b=0, c=0, d=0, e=–1	$x - xmode_4$
$xinfo_{13}$	a=1, b=–1, c=–1, d=0, e=0	$x - xmode_1 - xmode_2$
$xinfo_{14}$	a=1, b=0, c=-1, d=0, e=-1	$x - xmode_2 - xmode_4$
$xinfo_{15}$	a=1, b=0, c=–1, d=–1, e=0	$x - xmode_1 - xmode_3$
$xinfo_{16}$	a=1, b=–1, c=0, d=0, e=–1	$x - xmode_1 - xmode_4$

Блоки 5–10. Фрагментирование информативного сигнала осуществляется по формулам (12) и (13). Вычисление значений ZCR и STE фрагментов информативного сигнала осуществляется по формулам (2) и (4) соответственно.

В помехоустойчивом способе представлено решение проблемы выбора пороговых значений ZCR и STE. По аналогии со способом сегментации на основе анализа ODMD используется период начальной паузу (фоновый шум длительностью не менее 200 мс) в качестве данных для формирования пороговых значений ZCR и STE. Вычисляются математические ожидания и дисперсии значений ZCR и STE для начальных фрагментов, соответствующих вынужденной паузе:

$$\mu_{ZCR} = \frac{1}{S}\sum_{s=1}^{S} ZCRxinfo_s, \tag{16}$$

$$\sigma_{ZCR} = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(ZCRxinfo_s - \mu_{ZCR}\right)^2}, \tag{17}$$

$$\mu_E = \frac{1}{S}\sum_{s=1}^{S} Exinfo_s, \tag{18}$$

$$\sigma_E = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(Exinfo_s - \mu_E\right)^2}, \tag{19}$$

где $ZCRxinfo_s$, $Exinfo_s$ – значения ZCR и STE исследуемого фрагмента информативного сигнала.

Определение статуса «речь/пауза» фрагментов информативного сигнала заключается в проверке следующих условий:

$$\left|ZCRxinfo_s - \mu_{ZCR}\right| \geq \sigma_{ZCR}, \tag{20}$$

$$\left|Exinfo_s - \mu_E\right| \geq \sigma_E, \tag{21}$$

Если разница между текущим и средним значениями ZCR больше или равна стандартному отклонению, то фрагмент соответствует паузе. И наоборот, если условие не выполняется, то фрагмент соответствует речи. Аналогично, если разница между текущим и средним значениями

STE больше или равна стандартному отклонению, то фрагмент соответствует речи. И наоборот, если условие не выполняется, то фрагмент соответствует паузе.

Сегментация на основе совместного анализа ZCR и STE осуществляется с учетом сопоставления результатов, полученных по отдельности способами на основе анализа ZCR, STE.

4. Исследование помехоустойчивого способа сегментации «речь/пауза»

Для оценки помехоустойчивости предлагаемого способа сегментации речь/пауза сформирована база зашумленных речевых сигналов. Первоначально осуществлена запись чистых речевых сигналов посредством специально разработанной методики, в рамках которой дикторы зачитывали следующий текстовый материал:

– статья из рекламно-информационной газеты, включающая публицистический текст на русском языке (не менее 200 слов);

– краткое детское литературное произведение, включающее фонетически сбалансированный текст на русском языке (не менее 200 слов);

– счёт чисел от 0 до 99 на русском языке (100 слов).

Произношение диктором – размеренное. Параметры и пространственные характеристики микрофона не изменялись для всех регистрируемых дикторов. Количество дикторов – 20 человек, мужчины и женщины. Зашумление зарегистрированных чистых речевых сигналов осуществлялось программно посредством наложения сгенерированного белого шума. Таким образом, сформирована база зашумленных речевых сигналов с различными отношениями сигнал/шум (ОСШ): от 20 до –5 дБ с шагом 5 дБ.

Эффективность сегментации «речь/пауза» оценивалась посредством определения ошибок 1-го и 2-го рода. Основной задачей сегментации является точное обнаружение границ начала и окончания информативных участков вокализованной и невокализованной речи, поэтому ошибкой 1-го рода считалось ошибочное присваивание речевому фрагменту статуса «пауза». Ошибкой 2-го рода считалось ошибочное присваивание фрагменту паузы статуса «речь». Ошибки 1-го и 2-го рода определялись в сравнении с результатом сегментации, осуществленной вручную. Для упрощения анализа большого объема данных использовалось среднеарифметическое значение ошибок 1-го и 2-го рода.

Предполагается, что наименьшее среднеарифметическое значение соответствует оптимальным значениям ошибок 1-го и 2-го рода.

В табл. 3 представлены усредненные данные среднеарифметических значений ошибок 1-го и 2-го рода, полученные по результатам сегментации чистого и зашумленных речевых сигналов с помощью помехоустойчивого способа. Данные представлены для 3 модовых и 16 информативных сигналов. Зеленым цветом отмечены минимальные среднеарифметические значения ошибок 1-го и 2-го рода, полученные для каждого метода декомпозиции (ДЭМ1 – ДЭМ, ДЭМ2 – МДЭМ, ДЭМ3 – улучшенная ПМДЭМАШ) при разных длительностях исследуемых фрагментов (*T*) 10, 30 и 50 мс. Желтым цветом отмечены оптимальные результаты сегментации речь/пауза для каждого исследуемого сигнала.

5. Анализ результатов исследований

Наилучший результат сегментации зашумленной речи с ОСШ = 20 дБ со среднеарифметическим значением ошибок 1-го и 2-го рода 2,26 % достигается при исследовании информативного сигнала 3, представляющего собой сумму всех четырех модовых речевых сигналов. Среднеарифметическое значение 2,26 % соответствует оптимальным значениям 2,75 и 1,78 % ошибок 1-го и 2-го рода соответственно. Результат обеспечен при использовании метода МДЭМ с длительностью фрагментов 10 мс.

Наилучший результат сегментации зашумленной речи с ОСШ = 15 дБ со среднеарифметическим значением ошибок 1-го и 2-го рода 4,37 % достигается при исследовании информативного сигнала 3. Среднеарифметическое значение 4,37 % соответствует оптимальным значениям 5,72 и 3,02 % ошибок 1-го и 2-го рода соответственно. Результат обеспечен при использовании метода улучшенной ПМДЭМАШ с длительностью фрагментов 30 мс.

Наилучший результат сегментации зашумленной речи с ОСШ = 10 дБ со среднеарифметическим значением ошибок 1-го и 2-го рода 6,28 % достигается при исследовании информативного сигнала 7, представляющего собой сумму первого, второго и четвертого модовых речевых сигналов. Среднеарифметическое значение 6,28 % соответствует оптимальным значениям 10,07 и 2,49 % ошибок 1-го и 2-го рода соответственно. Результат обеспечен при использовании метода МДЭМ с длительностью фрагментов 10 мс.

Таблица 3

Усредненные данные среднеарифметических значений ошибок 1-го и 2-го рода, полученные по результатам сегментации чистого и зашумленных речевых сигналов предлагаемым помехоустойчивым способом

Сигнал	*T*, мс	Среднеарифметическое значение ошибок 1-го и 2-го рода, %																				
		Чистый сигнал			Зашумленный речевой сигнал																	
					ОСШ = 20 дБ			ОСШ = 15 дБ			ОСШ = 10 дБ			ОСШ = 15 дБ			ОСШ = 0 дБ			ОСШ = -5 дБ		
		ДЭМ1	ДЭМ2	ДЭМ3	ДЭМ1	ДЭМ2	ДЭМ3	ДЭМ1	ДЭМ2	ДЭМ3	ДЭМ1	ДЭМ2	ДЭМ3	ДЭМ1	ДЭМ2	ДЭМ3	ДЭМ1	ДЭМ2	ДЭМ3	ДЭМ1	ДЭМ2	ДЭМ3
Модовый речевой сигнал 1	10	13,96	13,60	13,34	6,84	6,63	6,41	16,32	16,16	15,65	18,28	17,82	17,36	21,76	22,22	21,99	37,29	36,50	37,80	42,92	43,15	43,38
	30	14,22	13,34	13,78	6,29	6,56	6,35	19,38	18,79	18,41	17,90	18,45	17,64	22,30	22,20	22,26	37,97	37,30	38,16	44,81	45,08	44,74
	50	13,34	13,42	14,55	6,66	6,27	6,35	6,66	16,99	16,15	18,64	18,64	17,53	22,16	23,36	22,91	38,27	37,93	37,62	44,04	44,06	44,27
Модовый речевой сигнал 2	10	4,38	5,84	8,50	3,52	3,61	3,32	10,97	10,51	10,74	14,24	13,95	13,89	14,75	14,52	14,98	29,49	29,63	27,97	31,16	30,57	29,72
	30	14,93	14,22	13,87	4,26	4,35	4,23	11,61	11,08	10,86	14,49	15,53	14,60	15,43	15,20	15,84	30,16	29,16	28,22	31,78	31,51	32,00
	50	9,02	9,61	10,07	4,10	4,00	3,87	16,78	11,31	11,66	14,49	14,26	13,89	14,29	14,98	15,32	30,03	29,16	28,56	31,99	31,95	31,25
Модовый речевой сигнал 3	10	15,19	14,92	22,78	5,74	7,08	5,50	7,86	7,83	7,86	11,88	11,08	11,43	13,85	15,68	14,69	18,64	19,14	19,59	34,59	32,18	31,33
	30	5,12	4,52	5,38	8,63	10,32	6,94	8,03	7,92	7,28	12,56	12,93	10,67	11,92	13,43	12,63	21,47	20,89	22,28	31,83	31,17	32,34
	50	12,25	8,96	5,36	6,38	11,00	10,53	12,00	7,88	7,77	11,67	9,95	11,04	12,21	12,80	12,69	19,47	19,55	18,60	35,24	31,87	30,00
Информативный сигнал 1	10	4,96	7,63	9,76	4,41	4,58	4,41	9,66	9,32	9,66	11,53	11,76	11,76	21,91	20,73	20,82	24,76	24,45	24,31	32,06	34,10	32,65
	30	17,22	16,60	16,15	3,56	3,27	3,38	10,45	10,52	10,30	12,13	11,95	11,63	26,65	20,49	26,33	23,10	24,13	23,55	33,68	34,52	33,56
	50	11,18	12,15	12,33	3,42	3,33	3,38	7,65	10,41	10,87	12,30	12,04	12,10	21,50	19,87	22,95	24,41	23,71	23,78	32,68	33,43	34,60
Информативный сигнал 2	10	16,68	21,48	27,16	6,50	6,50	6,76	5,38	5,49	5,52	9,25	8,56	8,16	9,50	10,81	8,96	15,25	15,19	15,18	22,33	21,28	22,37
	30	11,00	11,71	12,24	8,63	8,27	7,74	5,40	5,55	5,49	9,82	9,73	10,14	10,06	9,85	10,23	15,83	15,65	15,42	21,77	22,63	22,62
	50	19,70	19,79	12,07	14,99	8,88	8,42	10,71	5,14	5,02	11,49	10,11	11,16	11,85	11,85	11,75	16,20	15,37	15,37	22,15	23,06	23,68
Информативный сигнал 3	10	6,03	6,38	5,62	3,09	2,26	2,46	5,35	5,49	5,49	7,50	7,66	7,80	7,35	7,26	7,14	15,42	17,54	20,71	19,83	21,49	20,95
	30	5,52	4,53	3,69	4,68	4,77	4,68	4,75	4,93	4,37	7,62	7,83	7,71	11,05	10,54	11,87	18,72	17,45	14,81	19,57	18,03	19,33
	50	3,57	2,31	3,91	6,73	6,81	6,81	5,76	4,76	4,45	8,70	8,73	8,70	9,64	10,85	10,86	15,66	16,02	16,55	21,34	20,00	20,07
Информативный сигнал 4	10	16,37	18,46	19,08	3,75	3,75	3,66	9,03	6,96	7,42	8,75	8,75	8,63	10,64	10,40	9,95	19,18	19,82	19,53	25,07	24,97	25,33
	30	8,11	9,01	8,02	13,58	13,04	12,33	9,22	9,71	9,31	12,28	10,67	8,68	9,48	10,06	9,60	20,07	19,84	19,96	25,08	25,44	24,60
	50	14,29	13,75	7,93	9,15	10,13	7,96	4,92	8,71	8,66	9,52	8,97	10,63	10,00	10,40	9,83	19,39	19,48	19,07	26,02	26,27	25,31
Информативный сигнал 5	10	7,74	7,37	7,16	4,85	4,50	5,07	8,43	8,99	8,65	7,75	7,40	7,71	10,23	10,09	9,78	17,01	17,24	17,84	20,64	20,22	20,07
	30	5,54	4,05	4,91	3,29	3,93	3,72	6,17	5,47	5,74	8,11	8,11	8,11	9,24	8,91	9,42	17,31	16,86	16,88	21,06	20,58	20,69
	50	4,46	3,15	3,52	6,95	7,58	6,69	8,15	6,78	6,78	7,09	6,91	7,14	8,63	8,72	9,66	16,86	16,72	16,72	20,29	21,40	23,26
Информативный сигнал 6	10	11,63	11,01	11,27	7,10	12,77	6,80	6,14	5,95	6,20	8,68	8,77	8,41	12,89	12,28	12,56	17,29	17,28	17,67	25,70	21,66	22,21
	30	7,58	5,83	6,97	4,80	4,86	4,27	5,63	6,02	4,98	10,97	10,97	10,17	11,41	10,32	10,12	17,97	18,86	18,21	24,01	23,46	27,40
	50	6,41	4,97	4,60	5,72	8,90	8,81	6,74	5,67	5,64	9,14	9,03	8,22	11,84	11,55	11,56	15,76	15,90	15,53	25,83	23,02	24,89
Информативный сигнал 7	10	7,46	7,24	7,24	5,89	5,17	5,61	4,94	4,85	4,76	6,39	6,28	6,56	8,22	8,25	7,44	14,53	15,33	15,50	20,67	19,63	19,33
	30	6,59	5,51	6,05	8,44	8,59	7,61	5,06	4,82	4,55	6,72	6,93	6,63	17,02	17,93	19,08	15,01	14,33	16,15	19,08	22,03	21,19
	50	5,56	3,39	4,43	10,84	11,61	10,81	5,40	4,70	4,63	7,73	11,02	8,31	14,75	15,96	15,34	15,51	15,55	16,05	19,90	22,44	22,41
Информативный сигнал 8	10	7,09	6,24	6,72	4,35	4,44	4,79	5,06	5,23	5,21	8,65	11,31	12,13	8,63	8,43	8,57	17,23	16,65	18,59	23,84	25,87	23,62
	30	7,00	4,10	6,20	2,73	2,91	2,46	6,06	5,82	5,30	7,43	7,85	7,35	16,73	18,53	15,54	16,19	16,02	17,69	25,17	23,29	24,10
	50	3,72	3,53	2,62	3,17	2,91	3,71	5,34	5,49	5,58	7,61	7,38	7,96	15,08	15,64	15,60	15,91	17,42	18,01	31,06	28,81	26,00
Информативный сигнал 9	10	7,56	7,38	7,47	5,82	5,88	5,88	7,24	7,24	7,24	9,23	9,12	9,03	10,38	9,53	9,88	15,75	15,52	15,98	22,86	24,37	23,96
	30	7,07	7,35	6,91	6,03	5,73	5,97	7,16	7,07	7,33	9,82	9,73	9,68	9,79	9,56	9,56	15,98	15,86	15,75	22,76	23,53	24,43
	50	6,68	7,09	6,82	5,48	6,22	6,22	5,52	7,04	7,04	12,35	9,03	9,12	9,21	9,39	9,44	15,62	15,90	15,85	22,87	25,67	26,52
Информативный сигнал 10	10	7,34	7,25	7,80	6,80	6,80	6,57	9,17	8,93	8,93	6,98	6,51	6,60	9,02	9,49	9,60	13,69	14,41	14,05	21,18	19,43	19,68
	30	6,71	7,14	6,66	5,96	5,96	6,07	11,21	10,97	11,06	7,05	6,61	6,78	8,30	8,87	8,81	13,08	13,46	13,94	21,80	20,00	20,54
	50	7,23	7,62	7,62	5,73	5,96	5,84	7,13	7,86	8,13	7,33	6,71	6,80	8,39	9,01	8,98	15,56	15,55	16,86	20,94	20,58	21,03
Информативный сигнал 11	10	6,49	7,74	8,15	8,04	7,78	8,13	7,78	7,92	7,80	8,08	8,02	8,51	10,78	10,47	10,19	14,84	15,83	14,92	20,92	20,81	20,17
	30	7,35	7,08	7,39	7,29	7,93	7,65	7,48	8,16	8,00	13,85	11,87	10,05	9,90	9,50	9,73	14,54	14,88	15,04	19,11	20,61	20,49
	50	8,13	8,08	7,13	7,05	7,53	7,52	8,22	7,72	7,84	10,26	9,18	11,96	10,92	9,90	10,52	14,74	14,91	15,25	21,17	21,46	21,09
Информативный сигнал 12	10	10,34	9,76	9,59	8,00	7,47	7,56	7,27	7,46	7,66	9,69	9,87	10,04	9,23	8,96	9,10	18,52	16,44	13,92	21,73	23,86	21,81
	30	6,49	6,68	6,54	6,82	6,59	6,80	9,40	8,69	9,31	9,71	9,77	9,24	9,14	8,62	8,88	14,93	13,73	15,72	22,54	25,56	24,03
	50	5,54	6,73	5,74	6,15	6,15	5,63	7,64	9,02	8,80	9,19	8,96	9,48	8,44	8,88	8,35	18,66	18,54	19,08	23,55	20,94	24,34
Информативный сигнал 13	10	9,82	9,23	8,68	8,26	8,22	8,14	8,06	8,17	8,08	9,93	9,06	9,95	12,87	12,25	12,52	16,61	16,98	16,92	21,65	22,73	21,49
	30	9,51	9,53	8,78	8,72	8,23	8,17	8,73	8,82	8,73	10,19	10,10	9,79	12,01	12,37	12,32	15,08	14,46	14,01	21,20	21,57	21,18
	50	10,01	9,16	8,98	9,42	8,81	8,60	8,79	8,82	8,77	10,12	9,01	10,07	12,59	12,20	12,36	13,83	13,03	13,13	21,27	21,13	22,13
Информативный сигнал 14	10	13,33	8,86	12,02	9,73	9,07	9,38	9,06	9,26	9,03	11,34	13,53	15,28	13,35	8,96	12,19	15,90	15,95	20,96	25,98	26,62	27,12
	30	8,08	7,72	6,70	7,61	7,60	7,34	8,26	8,29	8,45	11,60	9,99	9,04	8,83	9,19	8,02	15,34	16,05	15,10	29,57	28,53	26,45
	50	6,45	7,80	7,95	7,13	7,14	6,68	9,17	8,27	8,19	7,83	8,78	8,26	8,83	9,10	8,37	15,84	16,47	16,54	29,62	27,26	27,93
Информативный сигнал 15	10	7,99	8,43	8,12	7,95	7,99	7,71	9,06	9,26	8,94	9,47	9,44	9,87	14,45	14,33	14,77	20,53	19,40	18,97	28,48	30,19	28,31
	30	7,72	8,48	7,48	8,56	8,41	8,94	8,60	9,26	8,98	11,28	11,53	10,77	11,20	11,83	11,31	19,90	22,30	21,73	25,93	27,10	28,71
	50	8,57	8,59	8,50	7,95	7,57	8,11	8,31	9,91	9,02	11,66	11,85	11,46	11,92	11,47	11,69	18,84	21,15	21,73	26,76	28,13	27,45
Информативный сигнал 16	10	6,67	6,06	7,22	8,35	8,11	8,46	8,63	8,65	8,65	12,05	12,35	11,46	11,10	11,45	11,19	17,30	17,52	17,39	24,79	25,91	26,77
	30	6,31	7,17	5,96	6,48	7,17	6,38	8,13	7,91	8,02	11,93	10,01	8,18	11,12	10,77	10,63	18,56	19,09	19,68	24,23	24,79	24,93
	50	6,16	6,72	6,58	8,61	8,49	8,95	9,31	8,69	8,69	12,45	12,19	12,37	11,03	11,47	10,23	18,47	17,99	18,62	24,42	24,86	25,64

Наилучший результат сегментации зашумленной речи с ОСШ = 5 дБ со среднеарифметическим значением ошибок 1-го и 2-го рода 7,14 % достигается при исследовании информативного сигнала 3. Среднеарифметическое значение 7,14 % соответствует оптимальным значениям 13,04 и 1,24 % ошибок 1-го и 2-го рода соответственно. Результат обеспечен при использовании метода улучшенной ПМДЭМАШ с длительностью фрагментов 10 мс.

Наилучший результат сегментации зашумленной речи с ОСШ = 0 дБ со среднеарифметическим значением ошибок 1-го и 2-го рода 13,03 % достигается при исследовании информативного сигнала 13, представляющего собой разность исходного зашумленного речевого сигнала и суммы 1-го и 2-го модовых сигналов. Среднеарифметическое значение 13,03 % соответствует оптимальным значениям 23,57 и 2,49 % ошибок 1-го и 2-го рода соответственно. Результат обеспечен при использовании метода МДЭМ с длительностью фрагментов 50 мс.

Наилучший результат сегментации зашумленной речи с ОСШ = –5 дБ со среднеарифметическим значением ошибок 1-го и 2-го рода 18,03 % достигается при исследовании информативного сигнала 3. Среднеарифметическое значение 18,03 % соответствует оптимальным значениям 27,00 и 9,06 % ошибок 1-го и 2-го рода соответственно. Результат обеспечен при использовании метода МДЭМ с длительностью фрагментов 30 мс.

В табл. 4 представлены усредненные данные значений ошибок 1-го и 2-го рода, полученные по результатам сегментации чистого и зашумленных речевых сигналов способами на основе анализа ZCR, STE, совместного анализа ZCR и STE, анализа ODMD, а также предлагаемого помехоустойчивого способа. В соответствии с данными, представленными в табл. 4, помехоустойчивый способ обеспечивает наилучшие показатели ошибок 1-го и 2-го рода для чистого и зашумленных речевых сигналов с ОСШ 20 и 15 дБ.

Для зашумленных речевых сигналов с ОСШ = 10 дБ достигнутое оптимальное значение ошибки 1-го рода (10,07 %) больше на 0,69 %, чем значения, полученные способом на основе совместного анализа ZCR и STE (9,38 %). Детализированный анализ данных в табл. 3 выявил лучшие значения ошибок 1-го и 2-го рода – 8,24 и 4,97 % соответственно, достигнутые помехоустойчивым способом при анализе информативного сигнала 10 (МДЭМ, 30 мс).

Таблица 4

Усредненные данные значений ошибок 1-го и 2-го рода, полученные по результатам сегментации чистого и зашумленных речевых сигналов способами на основе анализа ZCR, STE, совместного анализа ZCR и STE, анализа ODMD, а также предлагаемым помехоустойчивым способом

Способ сегментации речь/пауза	Среднеарифметическое значение ошибок 1-ого и 2-ого рода, %													
	Чистый сигнал		Зашумленный речевой сигнал											
			ОСШ = 20 дБ		ОСШ = 15 дБ		ОСШ = 10 дБ		ОСШ = 15 дБ		ОСШ = 0 дБ		ОСШ = -5 дБ	
	α	β	α	β	α	β	α	β	α	β	α	β	α	β
Способ на основе анализа ODMD	21,97	0,89	22,88	0,89	35,01	0,89	41,19	0,89	59,50	0,89	99,77	0,89	99,77	0,89
Способ на основе анализа ZCR	23,11	3,02	27,23	2,13	71,85	1,42	65,90	1,95	75,29	1,07	97,48	0,89	90,85	1,78
Способ на основе анализа STE	10,53	3,20	9,61	1,95	7,09	4,97	9,38	3,55	12,59	1,42	16,25	22,20	37,07	1,42
Способ на основе совместного анализа ZCR и STE	7,32	5,33	6,41	4,80	7,09	6,75	9,38	7,28	11,90	1,60	16,25	22,91	34,10	3,73
Помехоустойчивый способ	1,60	3,02	2,75	1,78	5,72	3,02	10,07	2,49	13,04	1,24	23,57	2,49	27,00	9,06

Для зашумленных речевых сигналов с ОСШ = 5 дБ ошибка 1-го рода (13,04 %) больше на 1,14 %, чем у способа на основе совместного анализа ZCR и STE (11,90 %). В рамках детализированного анализа выявлено лучшее значение ошибки 1-го рода 9,61 %, достигнутое предлагаемым способом при анализе информативного сигнала 8 (МДЭМ, 50 мс). Однако соответствующее значение ошибки 2-го рода (21,67 %) остается намного больше, чем у способа на основе совместного анализа ZCR и STE (1,60 %).

Для зашумленных речевых сигналов с ОСШ = 0 дБ ошибка 1-го рода (23,57 %) больше на 7,32 %, чем у способа на основе совместного анализа ZCR и STE (16,25 %). В рамках детализированного анализа выявлено наиболее близкое значение 18,08 %, достигнутое помехоустойчивым способом при анализе информативного сигнала 3 (улучшенная ПМДЭМАШ, 30 мс).

Для зашумленных речевых сигналов с ОСШ = –5 дБ ошибка 2-го рода (9,06 %) на 5,33 % больше, чем у способа на основе совместного анализа ZCR и STE (3,73 %).

На рис. 3 представлен пример, иллюстрирующий результаты сегментации речь/пауза сигнала длительностью 10 с с помощью предлагаемого помехоустойчивого способа. Сигнал представляет собой сочетание слов на русском языке: шанс, шар, баян, Лара, нормально. Слова подобраны таким образом, чтобы в них содержались разные по способу образования звуки: гласные, сонорные, шумные смычные (взрывные, фрикативные) и шумные щелевые.

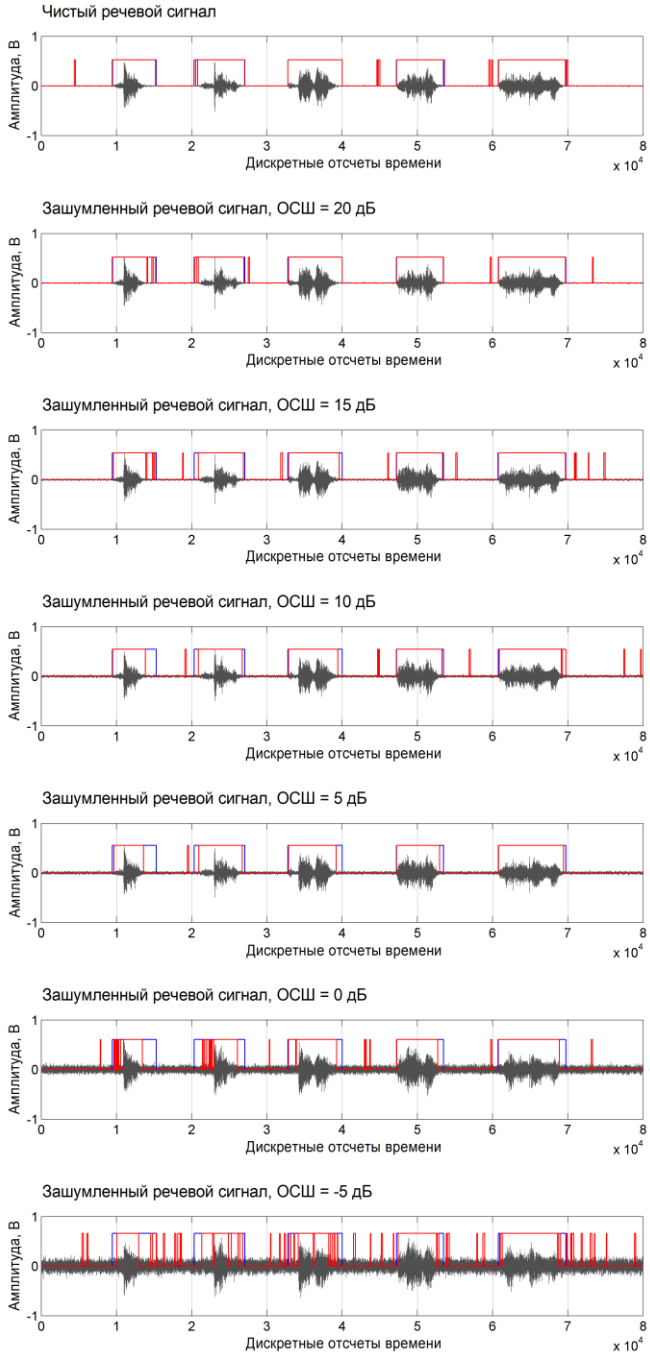

Рис. 3. Пример, иллюстрирующий результаты сегментации речь/пауза с помощью помехоустойчивого способа на основе ДЭМ (линией синего цвета обозначены результаты сегментации, осуществленной вручную, линией красного цвета обозначены достигнутые результаты сегментации)

Заключение

Подводя итоги анализа результатов исследований, можно сделать следующие выводы:

1. За счет применения метода ДЭМ на этапе предварительной обработки предлагаемый помехоустойчивый способ обеспечивает наилучшие результаты сегментации зашумленной речи на информативные участки. В сравнении с наиболее помехоустойчивым способом сегментации на основе совместного анализа ZCR и STE отмечаются следующие изменения значений ошибок 1-го и 2-го рода:

– ОСШ = 20 дБ улучшение на 3,66 и 3,02 %;
– ОСШ = 15 дБ улучшение на 1,37 и 3,73 %;
– ОСШ = 10 дБ улучшение на 1,14 и 2,31 %;
– ОСШ = 5 дБ улучшение на 2,29 и 0,36 %;
– ОСШ = 0 дБ ухудшение на 1,83 % и улучшение на 11,36 %;
– ОСШ = -5 дБ улучшение на 7,10 % и ухудшение на 5,33 %.

2. Наилучший результат сегментации речи на информативные участки достигается при использовании метода МДЭМ и длительности исследуемых фрагментов от 10 до 30 мс. Важно отметить, что применение методов ДЭМ и улучшенной ПМДЭМАШ, а также исследование фрагментов длительностью 50 мс также обеспечивает приемлемые результаты сегментации зашумленной речи.

3. В зависимости от приоритета решаемой задачи сегментации речь/пауза у исследователей имеется возможность выбирать между анализируемыми модовыми и информативными сигналами, обеспечивающими требуемые значения ошибок 1-го и 2-го рода.

4. В соответствии с результатами исследований разработанный помехоустойчивый способ сегментации «речь/пауза» может успешно тестироваться в интеллектуальных системах оценки психоэмоционального состояния человека в реальных условиях «агрессивной» шумовой обстановки.

В перспективе коллективом авторов планируется провести исследование быстродействия помехоустойчивого способа сегментации речь/пауза на основе метода ДЭМ, а также исследовать помехоустойчивый способ при зашумлении речи коричневым и розовым шумами.

Библиографический список

1. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов: пер. с англ. – М.: Радио и связь, 1981. – 496 с.

2. Atal B., Rabiner L.R. A pattern recognition approach to voiced unvoiced-silence classification with applications to speech recognition // IEEE Trans. Acoust. Speech Signal Process. – 1976. – Vol. 24, № 3. – P. 201–212.

3. Separation of Voiced and Unvoiced Using Zero Crossing Rate and Energy of the Speech Signal / R.G. Bachu, S. Kopparthi, B. Adapa, B.D. Barkana // American Society for Engineering Education (ASEE) Zone Conference Proceedings. – Pittsburgh, USA, 2008. – P. 1–7.

4. Huang X., Acero A., Hon H.-W. Spoken Language Processing. Guide to Algorithms and System Developmen. – New Jersey: Prentice Hall, 2001. – 980 p.

5. Yang J., Li Z., Su P. Review of speech segmentation and endpoint detection // Journal of Computer Applications. – 2020. – Vol. 40, № 1. – P. 1–7.

6. Childers D.G., Hand M., Larar J.M. Silent and voiced/unvoied/ mixed excitation (four-way), classification of speech // IEEE Transaction on ASSP. – 1989. – Vol. 37, № 11. – P. 1771–1774.

7. Reliable mute model and speech activity detection in speaker logs / D.-Z. Yang, J.-M. Xu, J. Liu, S.-H. Xia // Journal of Zhejiang University (Engineering Science). – 2016. – Vol. 50, № 1. – P. 151–157.

8. Duda R.O., Hart P.E., Strok D.G. Pattern Classification. – 2nd ed. – New Jersey: A Wiley-Interscience Publ. John Wiley & Sons, Inc., 2001. – 688 p.

9. Martin A., Charlet D., Mauuary L. Robust speech/non-speech detection using LDA applied to MFCC // 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221) (ICASSP2001) (May 7–11, 2001). – Salt Lake City, UT, USA. – Vol. 1. – P. 237–240.

10. Jiang N., Liu T. An improved speech segmentation and clustering algorithm based on SOM and k-means // Mathematical Problems in Engineering. – 2020. – Vol. 2020. – 19 p.

11. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder / J. Hlavnička, R. Čmejla, T. Tykalová, K. Šonka, E. Růžička, J. Rusz // Scientific Reports. – 2017. – Vol. 7(12). – 13 p.

12. Zheng G. Speech endpoint recognition algorithm based on wavelet coefficient variance // 2020 5th International Conference on Smart Grid and Electrical Automation (ICSGEA) (June 13–14, 2020). – Zhangjiajie, China. – P. 226–230.

13. Schuller B.W., Batliner A.M. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. – New York: Wiley, 2013. – P. 344.

14. Повышение точности измерения частоты основного тона на основе оптимизации процесса декомпозиции речевых сигналов на эмпирические моды / А.К. Алимурадов, Ю.С. Квитка, П.П. Чураков, А.Ю. Тычков // Измерение. Мониторинг. Управление. Контроль. – 2018. – № 4(26). – С. 53–65.

15. Алимурадов А.К. Исследование частотно-избирательных свойств методов декомпозиции на эмпирические моды для оценки частоты основного тона речевых сигналов // Труды МФТИ. – 2015. – Т. 7, № 3. – С. 56–68.

16. Алимурадов А.К., Тычков А.Ю. Применение метода декомпозиции на эмпирические моды для исследования вокализованной речи в задаче обнаружения стрессовых эмоций человека // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. – 2020. – № 3(35). – С. 7–29.

17. Фант Г.К. Акустическая теория речеобразования / пер. с англ. Л.А. Варшавского и В.И. Медведева; науч. ред. В.С. Григорьева. – М.: Наука, 1964. – 284 с.

18. Greenwood M.A., Kinghorn A. SUVing: automatic silence/unvoiced/voiced classification of speech. Undergraduate Coursework, Department of Computer Science, the University of Sheffield, UK, 1999. – 4 p.

19. Saha G., Chakroborty S., Senapat S. A new silence removal and endpoint detection algorithm for speech and speaker recognition applications // Eleventh National Conference on Communications (NCC-2005) (Jan. 28–30, 2005). – Kharagpur, India. – P. 51–61.

20. Huang, N.E., Zheng Sh., Steven R.L. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis // Proceedings of the Royal Society of London. – 1998. – A 454. – P. 903–995.

21. Zhaohua W., Huang N.E. Ensemble empirical mode decomposition: A noise-assisted data analysis method // Advances in Adaptive Data Analysis. – 2009. – № 1(1). – P. 1–41.

22. Yeh J.-R., Shieh J.-S., Huang N.E. Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method // Advances in Adaptive Data Analysis. – 2010. – № 2(2). – P. 135–156.

23. A complete Ensemble Empirical Mode decomposition with adaptive noise / M.E. Torres, M.A. Colominas, G. Schlotthauer, P. Flandrin // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-11) (May 22–27, 2011). – Prague, Czech Republic, 2011. – P. 4144–4147.

24. Colominasa M.A., Schlotthauera G., Torres M.E. Improved complete ensemble EMD: a suitable tool for biomedical signal processing // Biomed. Signal Proces. – 2014. – Vol. 14. – P. 19–29.

25. Алимурадов А.К., Фокина Е.А., Журина А.Е. Исследование влияния длительности анализируемых речевых сигналов на частотно-избирательные свойства декомпозиции на эмпирические моды // Новые информационные технологии и системы: сб. науч. ст. XVI Междунар. науч.-техн. конф. (г. Пенза, 27–29 ноября 2019 г.). – Пенза: Изд-во ПГУ, 2019. – С. 201–205.

26. Improvement of the Efficiency of Voice Control Based on the Complementary Ensemble Empirical Mode Decomposition / A.K. Alimuradov, P.P. Churakov, A.Yu. Tychkov, I.I. Artemov, A.V. Kuzmin // International Siberian Conference on Control and Communications (SIBCON 2016) (May 12–14, 2016). – Russia, Moscow, 2016. – 6 p.

Reference

1. Rabiner L.R., Shafer R.V. Tsifrovaia obrabotka rechevykh signalov [Digital speech signal protection: translation from English]. Moscow: Radio i sviaz', 1981, 496 p.

2. Atal B., Rabiner L.R. A pattern recognition approach to voiced unvoiced-silence classification with applications to speech recognition. *IEEE Trans. Acoust. Speech Signal Process*, 1976, vol. 24, no. 3, pp. 201-212.

3. Bachu R.G., Kopparthi S., Adapa B., Barkana B.D. Separation of Voiced and Unvoiced Using Zero Crossing Rate and Energy of the Speech Signal. *American Society for Engineering Education (ASEE) Zone Conference Proceedings*. Pittsburgh, USA, 2008, pp. 1-7.

4. Huang X., Acero A., Hon H.-W. Spoken Language Processing. Guide to Algorithms and System Developmen. New Jersey: Prentice Hall, 2001, 980 p.

5. Yang J., Li Z., Su P. Review of speech segmentation and endpoint detection. *Journal of Computer Applications*, 2020, vol. 40, no. 1, pp. 1-7.

6. Childers D.G., Hand M., Larar J.M. Silent and voiced/unvoied/ mixed excitation (four-way), classification of speech. *IEEE Transaction on ASSP*, 1989, vol. 37, no. 11, pp. 1771-1774.

7. Yang D.-Z., Xu J.-M., Liu J., Xia S.-H. Reliable mute model and speech activity detection in speaker logs. *Journal of Zhejiang University (Engineering Science)*, 2016, vol. 50, no. 1, pp. 151-157.

8. Duda R.O., Hart P.E., Strok D.G. Pattern Classification. 2nd ed. New Jersey: A Wiley-Interscience Publ. John Wiley & Sons, Inc., 2001, 688 p.

9. Martin A., Charlet D., Mauuary L. Robust speech/non-speech detection using LDA applied to MFCC. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221) (ICASSP2001) (May 7-11, 2001)*. Salt Lake City, UT, USA, vol. 1, pp. 237-240.

10. Jiang N., Liu T. An improved speech segmentation and clustering algorithm based on SOM and k-means. *Mathematical Problems in Engineering*, 2020, vol. 2020, 19 p.

11. Hlavnička J., Čmejla R., Tykalová T., Šonka K., Růžička E., Rusz J. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Scientific Reports*, 2017, vol. 7(12), 13 p.

12. Zheng G. Speech endpoint recognition algorithm based on wavelet coefficient variance. *2020 5th International Conference on Smart Grid and Electrical Automation (ICSGEA) (June 13-14, 2020)*. Zhangjiajie, China, pp. 226-230.

13. Schuller B.W., Batliner A.M. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. New York: Wiley, 2013, 344 p.

14. Alimuradov A.K., Kvitka Iu.S., Churakov P.P., Tychkov A.Iu. Povyshenie tochnosti izmereniia chastoty osnovnogo tona na osnove optimizatsii protsessa dekompozitsii rechevykh signalov na empiricheskie mody [Increasing the accuracy of measuring the pitch frequency based on the

optimization of the process of decomposition of speech signals on empirical modes]. *Izmerenie. Monitoring. Upravlenie. Kontrol'*, 2018, no. 4(26), pp. 53-65.

15. Alimuradov A.K. Issledovanie chastotno-izbiratel'nykh svoistv metodov dekompozitsii na empiricheskie mody dlia otsenki chastoty osnovnogo tona rechevykh signalov [Research of frequency-selective properties of empirical mode decomposition methods for speech signals 'pitch frequency estimation]. *Trudy Moskovskogo fiziko-tekhnicheskogo instituta*, 2015, vol. 7, no. 3, pp. 56-68.

16. Alimuradov A.K., Tychkov A.Iu. Primenenie metoda dekompozitsii na empiricheskie mody dlia issledovaniia vokalizovannoi rechi v zadache obnaruzheniia stressovykh emotsii cheloveka [Application of the method empirical mode decomposition for the study of voiced speech in the problem of detecting human stress emotions]. *Vestnik Permskogo natsional'nogo issledovatel'skogo politekhnicheskogo universiteta. Elektrotekhnika, informatsionnye tekhnologii, sistemy upravleniia*, 2020, no. 3(35), pp. 7-29.

17. Fant G.K. Akusticheskaia teoriia recheobrazovaniia [The acoustic theory of speech production]. Ed. V.S. Grigor'eva. Moscow: Nauka, 1964, 284 p.

18. Greenwood M.A., Kinghorn A. SUVing: automatic silence/unvoiced/voiced classification of speech. Undergraduate Coursework, Department of Computer Science, the University of Sheffield, UK, 1999, 4 p.

19. Saha G., Chakroborty S., Senapat S. A new silence removal and endpoint detection algorithm for speech and speaker recognition applications. *Eleventh National Conference on Communications (NCC-2005) (Jan. 28-30, 2005)*. Kharagpur, India, pp. 51-61.

20. Huang, N.E., Zheng Sh., Steven R.L. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London*, 1998, A 454, pp. 903-995.

21. Zhaohua W., Huang N.E. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, 2009, no. 1(1), pp. 1-41.

22. Yeh J.-R., Shieh J.-S., Huang N.E. Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method. *Advances in Adaptive Data Analysis*, 2010, no. 2(2), pp. 135-156.

23. Torres M.E., Colominas M.A., Schlotthauer G., Flandrin P. A complete Ensemble Empirical Mode decomposition with adaptive noise. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-11) (May 22-27, 2011)*. Prague, Czech Republic, 2011, pp. 4144-4147.

24. Colominasa M.A., Schlotthauera G., Torres M.E. Improved complete ensemble EMD: a suitable tool for biomedical signal processing. *Biomed. Signal Proces*, 2014, vol. 14, pp. 19-29.

25. Alimuradov A.K., Fokina E.A., Zhurina A.E. Issledovanie vliianiia dlitel'nosti analiziruemykh rechevykh signalov na chastotno-izbiratel'nye svoistva dekompozitsii na empiricheskie mody [Investigation of the influence of the duration of the analyzed speech signals on the frequency-selective properties of the decomposition into empirical modes]. *Novye informatsionnye tekhnologii i sistemy. Sbornik nauchnykh statei XVI Mezhdunarodnoi nauchno-tekhnicheskoi konferentsii (Penza, 27-29 November 2019)*. Penza: Penzenskii gosudarstvennyi universitet, 2019, pp. 201-205.

26. Alimuradov A.K., Churakov P.P., Tychkov A.Yu., Artemov I.I., Kuzmin A.V. Improvement of the Efficiency of Voice Control Based on the Complementary Ensemble Empirical Mode Decomposition. *International Siberian Conference on Control and Communications (SIBCON 2016) (May 12-14, 2016)*. Russia, Moscow, 2016, 6 p.

Сведения об авторах

Алимурадов Алан Казанферович (Пенза, Россия) – кандидат технических наук, доцент кафедры «Радиотехника и радиоэлектронные системы» Пензенского государственного университета (440026, Пенза, ул. Красная, 40, e-mail: alansapfir@yandex.ru).

About the author

Alan K. Alimuradov (Penza, Russian Federation) – Ph. D. in Technical Sciences, Associate Professor of the Department of Radio Engineering and Radioelectronic Systems Penza State University (440026, Penza, 40, Krasnaya str., e-mail: alansapfir@yandex.ru).

Получено 01.05.2021