Способ адаптивного измерения просодических характеристик речевых сигналов

А. К. Алимурадов[1], А. Ю. Тычков[2], П. П. Чураков[3]
Пензенский государственный университет
[1]alansapfir@yandex.ru, [2]tychkov-a@mail.ru, [3]churakov-pp@mail.ru

***Аннотация*. Предложен способ адаптивного измерения просодических характеристик речевых сигналов. Суть способа заключается в адаптивной обработке речевых сигналов с помощью улучшенной полной множественной декомпозиции на эмпирические моды с адаптивным шумом; измерении частоты основного тона и ее функционалов. Представлена блок-схема способа и краткое математическое описание. Проведено исследование с использованием сформированной верифицированной базы сигналов здоровых пациентов и пациентов с пограничными психическими расстройствами мужского и женского пола, в возрасте от 18 до 60 лет. Результаты исследований оценивались в сравнении с известными способами измерения частоты основного тона. В соответствии с результатами исследования, разработанный способ обеспечивает повышение точности определения пограничных психогенных расстройств: для ошибки первого рода в среднем точнее на 13,33 % и для ошибки второго рода на 4,35 %.**

Ключевые слова: речевой сигнал; частота основного тона; улучшенная полная множественная декомпозиция на эмпирические моды с адаптивным шумом; пограничные психические расстройства

I. Введение

На сегодняшний день мониторинг состояния психического здоровья человека является социально-значимой проблемой для каждого государства. Оценка психоэмоционального состояния особенно важна в отраслях человеческой деятельности, сопряженных с повышенным риском для жизни населения: пилоты, космонавты, военнослужащие, диспетчеры аэропортов, диспетчеры опасных производственных объектов (АЭС, ТЭС, объектов химической промышленности) и других операторов систем управления с высокой степенью ответственности.

Важность анализа речи с целью выявления нарушений работы нервной системы подробно рассмотрена в работе [1], в которой авторы подчеркивают, что вид и степень выраженности психических расстройств кодируются в информативные параметры речевых сигналов.

[1]Работа выполнена при финансовой поддержке Российского научного фонда, проект № 17-71-20029.
[2]Работа выполнена при финансовой поддержке Совета по грантам Президента РФ, проект СП-246.2018.5

Эффективность оценки состояния психического здоровья человека зависит от корректной обработки речевых сигналов, которая определяется точностью измерения его амплитудных, временных, частотных, энергетических и других характеристик. Основная причина низкой точности и больших погрешностей в измерениях связана с использованием неадаптивных методов обработки нестационарных речевых сигналов, характеристики которых быстро изменяются во времени.

В данной работе авторы решают две основные задачи: выбор и обоснование адаптивных методов обработки речевых сигналов[1]; поиск уникально новых признаков и скрытых паттернов пограничных психических расстройств[2]. Исследования являются продолжением ранее опубликованных трудов авторов [2, 3].

II. Материалы и методы

A. *Адаптивная обработка*

Исследования методов обработки речевых сигналов, выявили перспективность использования адаптивной технологии разложения нестационарных сигналов, возникающих в нелинейных системах – декомпозиции на эмпирические моды (ДЭМ) [4]. ДЭМ обеспечивает локальное разложение сигнала на быстрые и медленные колебательные функции. В результате разложения исходный сигнал представляется в виде суммы амплитудных и частотных модулированных функций, называемых эмпирическими модами (ЭМ). Аналитическое выражение ДЭМ выглядит следующим образом:

$$x(n) = \sum_{i=1}^{I} IMF_i(n) + r_i(n)$$

где $x(n)$ – исходный сигнал, $IMF_i(n)$ – ЭМ, $r_i(n)$ – конечный остаток, i=1, 2, …, I – номер ЭМ, n – дискретный отсчет времени ($0<n\leq N$, N – количество дискретных отсчетов в сигнале).

Среди всех разновидностей методов декомпозиции наиболее адаптивным к речевым сигналам является усовершенствованная полная множественная декомпозиция на эмпирические моды с адаптивным шумом (ПМДЭМАШ) [5].

Математическое описание метода улучшенной ПМДЭМАШ представлено ниже:

Этап 1. С помощью аппарата ДЭМ и выражая из формулы $\left\langle E_1(x_j(n))\right\rangle = \left\langle x_j(n)\right\rangle - \left\langle M(x_j(n))\right\rangle$ — локальные средние значения шумовых копий исходного сигнала ($x_j(n) = x(n) + \beta_0 E_1(w_j(n))$) определяется первый остаток:

$$r_1(n) = \left\langle M(x_j(n))\right\rangle$$

где $E_i(\cdot)$ – аппарат извлечения ЭМ методом ДЭМ (i – номер моды), $x_j(n) = x(n) + w_j(n)$ – шумовые копии исходного сигнала ($x(n)$ – исходный речевой сигнал, $w_j(n)$ – реализации белого шума с нулевой средней единичной дисперсией), $M(\cdot)$ – аппарат, создающий локальное среднее значение применяемого сигнала, $\beta_i = \varepsilon_i \mathrm{std}(r_i)$ – коэффициент, допускающий выбор различных значений отношения сигнал/шум.

Этап 2. На первом этапе для $i = 1$ вычисляется первая мода: $IMF_1(n) = x(n) - r_1(n)$.

Этап 3. Вычисляется второй остаток как усредненное локальное среднее значение шумовых копий первого остатка $r_1(n) + \beta_1 E_2(w_j(n))$ и определяется вторая мода:

$$\widetilde{IMF}_2(n) = r_1(n) - r_2(n) = r_1(n) - \left\langle M\left(r_1(n) + \beta_1 E_2(w_j(n))\right)\right\rangle$$

Этап 4. На последующих этапах для $i = 3, ..., I$ вычисляется i-й остаток

$$r_i(n) = \left\langle M\left(r_{i-1}(n) + \beta_{i-1} E_i(w_j(n))\right)\right\rangle$$

Этап 5. Вычисляется i-ая мода

$$\widetilde{IMF}_i(n) = r_{i-1}(n) - r_i(n)$$

Этап 6. Переход к этапу 4 для следующей моды i.

Константы β_i выбираются так, чтобы получить желаемое отношение сигнал/шум между добавленным шумом и остатком, к которому добавляется шум.

Применение улучшенной ПМДЭМАШ на этапе адаптивной обработки обеспечивает [4, 5]:

- адаптивное разложение, так как базисные функции, используемые при декомпозиции, извлекаются непосредственно из исходного речевого сигнала и позволяют учитывать только ему свойственные особенности (скрытые модуляции, области концентрации энергии и т.п.);
- минимальный уровень остаточного шума и отсутствие паразитных ЭМ, возникающих на ранних этапах декомпозиции вследствие перекрытия масштабно-энергетических пространств мод.

B. Измерение просодических характеристик

Для эффективного детектирования новых признаков и скрытых паттернов состояний с высоким и низким уровнем психоэмоционального возбуждения лучше всего подходят просодические характеристики речевых сигналов [1]. Просодические характеристики – особенности речи, не являющиеся фонематическими, характеризующие речевую мелодию, темпоральные и тембральные особенности речи, её ритм, словесные тоны, стыки, паузы и интонации (т.е. фонация основного тона на уровне фраз). Речь представляет собой нестационарный акустический сигнал сложной формы, состоящий из вокализованных и невокализованных участков, образующихся соответственно в результате периодических и непериодических колебаний голосовых связок. Периодические колебания голосовых связок называется основным тоном (ОТ). Частота колебаний связок является важным просодическим информативным параметром речи– частотой основного тона (ЧОТ).

Особенностью ЧОТ при пограничных психических расстройствах являются колебания голосовых связок характеризующие нерегулярность, которая проявляется в виде значительных изменений длительности периодов ОТ (на 30 – 40%) и в виде небольших флуктуации соседних периодов тона. Нерегулярности возникают из-за неполного смыкания голосовых связок в начале и в конце вокализованных участков.

III. Описание способа

На рис. 1 представлена упрощенная блок-схема способа адаптивного измерения просодических характеристик речевых сигналов. Структурно способ делится на два этапа: адаптивная обработка (блоки 2-4) и измерение просодических характеристик (блоки 5-7). Блоки 8 и 9 применяются для исследования предложенного способа.

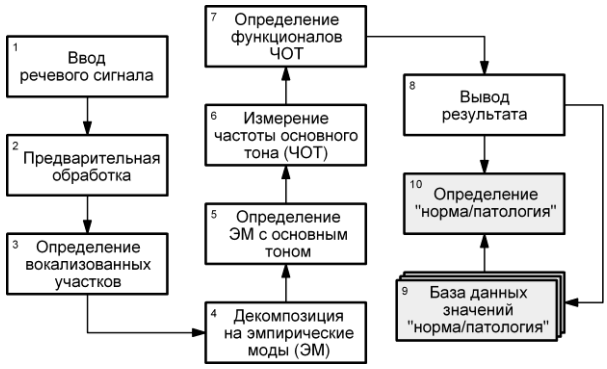

Рис. 1. Блок-схема способа адаптивного измерения просодических характеристик речевых сигналов

Сегментация речевого сигнала представляет собой обнаружение границ вокализованных и невокализованных участков (паузы и дыхание) в общем речевом потоке (рис. 2). Определение вокализованных участков осуществлялось на основе вычисления следующих параметров в скользящем окне и кластерного анализа: скорости пересечения сигнала через нулевое значение, автокорреляционной функции, энергии/мощности.

Декомпозиция методом улучшенной ПМДЭМАШ позволит разложить речевой сигнал на моды, отражающие информативные шумовые и сигнальные составляющие, свободные от тренда (оконечного остатка), на основе которых определяется ЭМ, содержащая ОТ и измеряется ЧОТ.

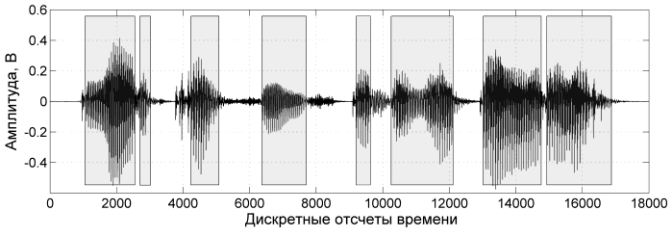

Рис. 2. Вокализованные участки речевого сигнала

Результат разложения вокализованного участка речевого сигнала с использованием улучшенной ПМДЭМАШ приведен на рис. 3.

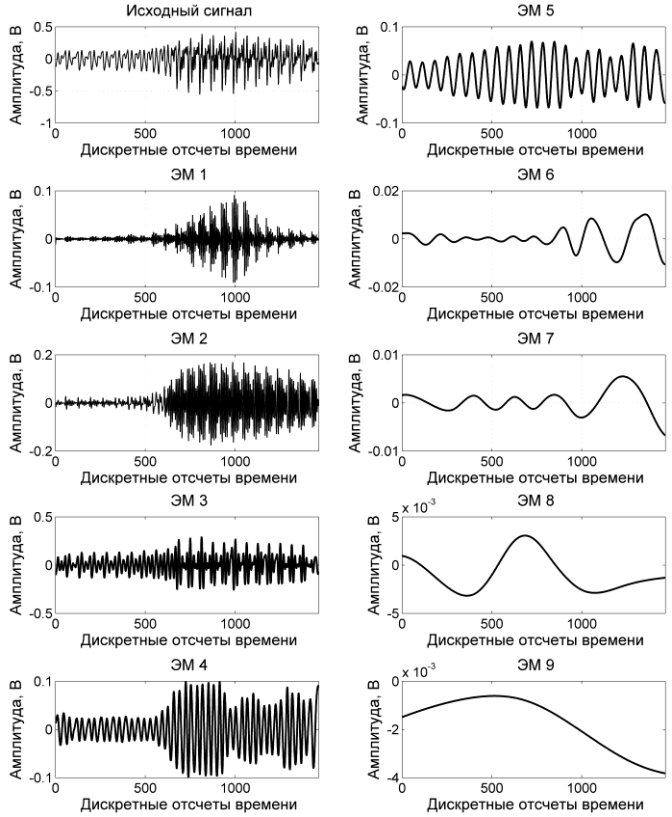

Рис. 3. Результат разложения вокализованного участка речевого сигнала методом улучшенной ПМДЭМАШ

Как видно из рис. 3 вокализованный участок речевого сигнала разложен на 9 ЭМ. Две первые моды содержат основной шум, присутствующий в исходном сигнале. Шестая мода и последующие являются низкочастотными и соответствуют присутствующему в сигнале тренду. Ценная информация, связанная со смыканием голосовых складок, появляется с третьей по пятую ЭМ.

Определение ЭМ, содержащей ОТ, заключается в последовательном вычислении разницы значений логарифмов энергии между текущей и последующей модами по модулю. Логарифмирование энергии применяется для сжатия амплитуды сигнала в большом динамическом диапазоне. В результате из последовательности полученных значений разницы, большему из них соответствует резкий спад энергии между информативной ЭМ, содержащей ОТ и трендовой.

Измерение ЧОТ осуществляется с использованием функция измерения мгновенной энергии сигнала – оператора Тигра, обладающего простотой, эффективностью и хорошей восприимчивостью к изменению речевого сигнала:

$$T(n) = (IMF_{i,PF}(n))^2 - IMF_{i,PF}(n-1) \times IMF_{i,PF}(n+1)$$

где $T(n)$ – функция оператора Тигра; $IMF_{i,PF}(n)$ – ЭМ, содержащая ОТ.

Для измерения частоты используются близкорасположенные максимумы, функции оператора Тигра между которыми определяется разница в дискретных отсчетах времени, вычисляется период ОТ в секундах и ЧОТ в герцах:

$$P_0 = \frac{T_{\max}(n+2) - T_{\max}(n)}{f_d}, \quad f_0 = \frac{1}{P_0}$$

где P_0 – ОТ, f_0 – ЧОТ; $T_{max}(n)$, $T_{max}(n+1)$ – максимумы функции оператора Тигра; f_d – частота дискретизации.

Для расширения информационного пространства о ЧОТ определяются следующие функционалы:

- среднее значение ЧОТ в Гц:

$$f_0 = \frac{1}{P}\sum_{p=1}^{P} f_{0,p}$$

где f_0 – ЧОТ, p=1, 2, …, P – номер периода ОТ;

- максимальное $max(f_0)$ и минимальное $mix(f_0)$ значения ЧОТ, в Гц;
- стандартное отклонение контура ЧОТ:

$$SD_{f0} = \frac{1}{P-1}\sum_{p=1}^{P}(f_{0,p} - f_{0,mean})^2$$

- диапазон фонационных частот:

$$PFR = 12 \times \frac{\log\left(\frac{\max(f_0)}{\min(f_0)}\right)}{\log 2}$$

- среднее абсолютное значение джиттера:

$$MAJ = \frac{1}{P-2}\sum_{p=P-1}^{1}\left|f_{0,p-1} - f_{0,p}\right|$$

- джиттер:

$$J = \frac{MAJ}{f_{o,mean}}$$

- среднее относительное возмущение ЧОТ, сглаженное за 3 периода ОТ:

$$RAP = \frac{\frac{1}{P-2}\sum_{p=2}^{P-2}\left|\left(f_{0,p+1} + f_{0,p} + f_{0,p-1}/3\right) - f_{0,p}\right|}{f_{0,mean}} \times 100$$

- коэффициент возмущения ЧОТ, сглаженный за 5 периодов ОТ:

$$PPQ = \frac{\frac{1}{P-4}\sum_{p=3}^{P-2}\left|\left(\sum_{k=p-2}^{p+2} f_{0,k}/5\right) - f_{0,i}\right|}{f_{0,mean}} \times 100$$

Обозначенные выше паттерны максимально полно отражают информацию о скрытых нарушениях работы органов речевого аппарата вследствие пограничных психических расстройств.

IV. Исследование способа

Для оценки эффективности разработанного способа при поддержке Областной клинической больницы им. К.Р. Евграфова (г. Пенза) и Пензенского государственного университета сформирована группа испытуемых и верифицированная база сигналов. В группу испытуемых отобрано 220 чел. мужского и женского пола, в возрасте от 18 до 60 лет, поступивших с явно выраженной симптоматикой пограничных психических расстройств. Для оценки эффективности способа, использовался параметр – ошибки первого и второго рода.

Все этапы обработки сигналов и анализа данных были выполнены в среде математического моделирования © Matlab (MathWorks).

Результаты исследования способа оценивались в сравнении со способами измерения ЧОТ, программная реализация которых имеется в открытом доступе: на основе автокорреляционной функции и её модификаций («*YIN*») [6], устойчивого метода отслеживания основного тона (*Robust Algorithm for Pitch Tracking*, *RAPT*) [7] и оценки основного тона пилообразной формы (*Sawtooth Waveform Inspired Pitch Estimation*, *SWIPE*) [8].

Результаты определения пограничных психических расстройств представлены в табл. 1.

ТАБЛИЦА I РЕЗУЛЬТАТЫ ОПРЕДЕЛЕНИЯ ПОГРАНИЧНОГО ПСИХИЧЕСКОГО РАССТРОЙСТВА

Прогнозируемый результат	**Результат определения**		**Ошибки первого и второго рода, %**	
	Патология	***Норма***		
Способ на основе устойчивого отслеживания основного тона (RAPT)				
Патология	184 чел.	36 чел.	1-ого	16,36
Норма	18 чел.	202 чел.	2-ого	8,19
Способ на основе автокорреляционной функции («YIN»)				
Патология	156 чел.	64 чел.	1-ого	29,1
Норма	24 чел.	196 чел.	2-ого	10,9
Способ на основе оценки основного тона пилообразной формы (SWIPE)				
Патология	178 чел.	42 чел.	1-ого	19,1
Норма	15 чел.	205 чел.	2-ого	6,81
Разработанный способ				
Патология	202 чел.	18 чел.	1-ого	8,19
Норма	9 чел.	211 чел.	2-ого	4,1

Как видно из результатов, процент ложных присваиваний статуса «норма» речевым сигналам, произнесенным пациентами с пограничными психическими расстройствами у *RAPT* (16,36 %), «*YIN*» (29,1 %) и *SWIPE* (19,1 %) слишком велик. Ни один из трех способов не обеспечил допустимый показатель 10 % – 22 человека из 220, который условно определили для себя авторы статьи в качестве удовлетворительного для диагностических систем обнаружения пограничных психических расстройств. То же самое можно сказать о ложных присваиваниях статуса «патология» речевым сигналам, произнесенным здоровыми пациентами: 8,19 %, 10,9 % и 6,81 % соответственно. Намного превосходящие значения ошибок 1-ого и 2-ого роды были достигнуты разработанным способом: 8,19 % и 4,1 % соответственно.

Исходя из полученных результатов и учитывая, что все четыре способа исследовались в равных условиях, можно сделать вывод – в условиях нерегулярности моторики органов речевого аппарата при пограничных психических расстройствах, возможности аналогов существенно ограничены. В первую очередь ограничение обусловлено стационарной моделью речевого сигнала, лежащей в их основе, которая подразумевает точное повторение периода ОТ. При изменениях периода, связанных с расстройствами нервной системы, точность измерения ЧОТ существенно снижается.

Разработанный способ может быть успешно использован в диагностических системах обнаружения пограничных психических расстройств и внедрен в клиническую практику врача-психиатра.

Список литературы

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

[2] Alimuradov A.K. Speech/pause detection algorithm based on the adaptive method of complementary decomposition and energy assessment of intrinsic mode functions / A.K. Alimuradov, A.Yu. Tychkov, A.V. Ageykin, P.P. Churakov, Yu.S Kvitka, A.P. Zaretskiy // 2017 XX IEEE International Conference on Soft Computing and Measurements (SCM), May 24-26, 2017, Russia, St. Petersburg, p. 610-613.

[3] Alimuradov A.K. Measurement of Speech Signal Patterns under Borderline Mental Disorders / A.K. Alimuradov, A.Yu. Tychkov, A.V. Kuzmin, P.P. Churakov, A.V. Ageykin, G.V. Vishnevskaya // Proceedings of the 21st Conference of Open Innovations Association FRUCT, 6-10 November, 2017, Finland, Helsinki, p. 26-33.

[4] 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 A. 1998. Vol. A454. p. 903-995.

[5] Colominasa M.A., Schlotthauera G., Torres M.E. Improved complete ensemble EMD: A suitable tool for biomedicalsignal processing // Biomedical Signal Processing and Control. 2014. Vol. 14. p. 19-29.

[6] Cheveigne A., Kawahara H. «YIN» a fundamental frequency estimator for speech and music // The Journal of the Acoustical Society of America. 2002. Vol. 111, № 4. p. 1917-1930.

[7] Talkin D. A Robust Algorithm for Pitch Tracking (RAPT) // Chapter 14 in Speech Coding & Synthesis / D. Talkin; ed. by W. B. Kleijn and K. K. Paliwan. New York, USA, Elsevier Science. 1995. p. 495-518.

[8] Camacho A., Harris J.G. A. Sawtooth waveform inspired pitch estimator for speech and music // The Journal of the Acoustical Society of America. 2008. Vol. 123, № 4. p. 1638-1652.