А.К. Алимурадов, А.Ю. Тычков, П.П. Чураков

НОВЫЙ ПОДХОД К СЕГМЕНТАЦИИ РЕЧЕВЫХ СИГНАЛОВ НА ОСНОВЕ ДЕКОМПОЗИЦИИ НА ЭМПИРИЧЕСКИЕ МОДЫ ДЛЯ ОЦЕНКИ ПСИХОЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ ЧЕЛОВЕКА

(Пензенский государственный университет)

Контроль психоэмоционального состояния очень важен в тех отраслях человеческой деятельности, которые сопряжены с повышенным риском возникновения техногенных и биогенных аварий [1]. Особенно это актуально для операторов систем управления с высокой степенью ответственности: пилотов, космонавтов, диспетчеров аэропортов и опасных производственных объектов (АЭС, ТЭС, объектов химической и ядерной промышленности).

В настоящее время исследования в области обнаружения и оценки психоэмоционального расстройства человека по речи активно поддерживаются грантами крупных компаний и международных технологических фондов: Industry Cooperation with BMW AG, проект «Sentiment Analyses», 01.05.2018 - 31.05.2021; EU H2020 Marie Sklodowska-Curie Innovative Training Networks European Training Networks (MSCA-ITN-ETN: ENG), проект «Training network on Automatic Processing of Pathological Speech» (#766287), 01.11.2017 - 31.10.2021.

Наибольший интерес в данных исследованиях представляют новые подходы и технологии обработки речевых сигналов, однако в силу коммерческой тайны данная информация не распространяется. По этой причине модернизация существующих и разработка новых технологий обработки речи остаются в центре внимания исследователей при решении задач по обнаружению и оценке психоэмоционального расстройства человека.

В данной статье представлена новая технология сегментации речевых сигналов на информативные участки, основанная на методе декомпозиции на эмпирические моды (ДЭМ) [2]. Работа является продолжением ранее опубликованной статьи [3] и выполнена при финансовой поддержке Совета по грантам Президента РФ, проект № СП-246.2018.5, 2018-2020 гг.

В зависимости от степени участия голосовых связок, речь человека делится на вокализованную, невокализованную и паузы. В соответствии с физиологическим аспектом формирования речи человек перед произношением делает начальную кратковременную паузу от 200 до 500 мс, соответствующую тишине. Речевой аппарат чрезвычайно чувствителен к нарушениям работы нервной системы. Продолжительность, скорость, ускорение и энтропия распределения временных интервалов вокализованных, невокализованных участков и участков пауз характеризуют работу речевого аппарата.

Сегментация на информавтиные участки представляется собой процесс определение точных границ вокализованных, невокализованных участков и участков пауз в слитной речи. Корректное определение границ информативных участков речи не только повышает эффективность обнаружения и оценки

психоэмоционального расстройства человека, но и уменьшает количество вычислительных операций.

В настоящее время существует много различных подходов к сегментации речевых сигналов. Среди наиболее известных можно выделить следующие способы: основанные на анализе кратковременной энергии (*Short-time Energy*, *STE*) и скорости пересечения сигнала через нулевое значение (*Zero-crossing Rate*, *ZCR*); основанные на анализе статистических свойств фонового шума и одномерного расстояния Махаланобиса.

Метод ДЭМ был разработан Норденом Хуангом в 1998 году и предназначался для разложения нестационарных сигналов, возникающих в нелинейных системах [2]. Принцип ДЭМ состоит в разложении сигнала в сумму функций с ограниченной полосой, называемых эмпирическими модами (ЭМ):

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

где $x(n)$ - исходный сигнал; $IMF_i(n)$ - ЭМ; $r_I(n)$ - конечный остаток, $i = 1, 2, \ldots, I$ - номер ЭМ, n - дискретный отсчет времени.

При разложении, модель сигнала не задаётся заранее, ЭМ вычисляются в ходе процедуры отсеивания с учетом локальных особенностей (таких как экстремумы и нули сигнала) и внутренней структуры каждого конкретного сигнала. Результаты подробных исследований технологий декомпозиций, выявили перспективность использования улучшенной полной множественной декомпозиции на эмпирические моды с адаптивным шумом (ПМДЭМАШ) [4], базисом которой является классическая ДЭМ.

Особенностью улучшенной ПМДЭМАШ является добавление к исходному сигналу контролируемого шума для создания новых экстремумов:

$$x_j(n) = x(n) + w(n),$$
$$x_j(n) = \sum_{i=1}^{I} IMF_{ji}(n) + r_{jI}(n),$$
$$IMF_i(n) = \sum_{j=1}^{J} \frac{IMF_{ji}(n)}{J},$$
$$r_I(n) = \sum_{j=1}^{J} \frac{r_{jI}(n)}{J},$$

где $j = 1, 2, \ldots, J$ - количество реализаций белого шума; $x_j(n)$ - шумовые копии речевого сигнала; $w_j(n)$ - реализации белого шума.

Способ сегментации на основе новой технологии состоит из двух этапов, в рамках которых этапа осуществляется: линейное деление речевого сигнала на кратковременные фрагменты длительностью 30 мс; декомпозиция фрагментов на ЭМ; определение параметров ЭМ; формирование пороговых значений параметров ЭМ фрагментов, соответсвующих начальной паузе; определение вокализованных, невокализованных участков и участков пауз.

В соответствии с новым подходом каждый кратковременный фрагмент речевого сигнала представляется набором ЭМ, полученных методом улучшенной ПМДЭМАШ. Особенностью новой технологии сегментации является то, что соотнесение анализируемого фрагмента сигнала к вокализованной, невокализованной речи или к паузе осуществляется, исследуя свойства каждой ЭМ фрагмента в отдельности. Учитывая, что каждая ЭМ

обладает определенными параметрами, сравнительный анализ мод по отдельности значительно повышает эффективность определения границ вокализованных, невокализованных участков и участков пауз при нестабильной работе речевого аппарата.

К исследуемым параметрам ЭМ фрагментов речевого сигнала относятся:

– логарифм энергии:

$$LE_{s,i} = log_2\left(\sum_{n=1}^{N}\left(IMF_{s,i}(n)\right)^2\right),$$

где $LE_{s,i}$ - логарифм энергии ЭМ фрагмента речевого сигнала; s - номер фрагмента;

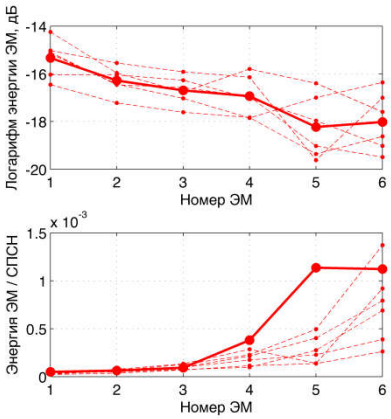
– отношение кратковременной энергии ЭМ к скорости пересечения сигнала ЭМ через нулевое значение (СПСН):

$$Z_{s,i} = \frac{\sum_{n=1}^{N}\left(IMF_{s,i}(n)\right)^2}{ZCR_{s,i}},$$

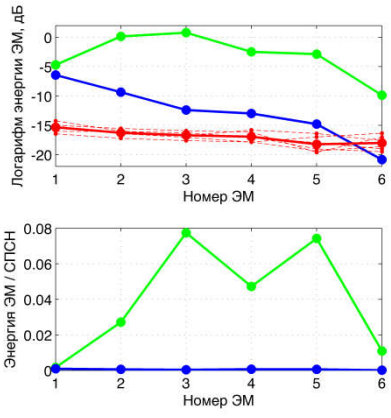
$$ZCR_{s,i} = 0{,}5\sum_{n=1}^{N-1}\left|sgn\left(IMF_{s,i}\left((n-1)N+n+1\right)\right) - sgn\left(IMF_{s,i}\left((n-1)N+n\right)\right)\right|,$$

где $ZCR_{s,i}$ - скорость пересечения сигнала ЭМ через нулевое значение; sgn - знаковая функция ($sgn(x) = 1$, если $x \geq 0$ и -1 при $x < 0$).

Применяя усреднее значений параметров LE и Z ЭМ для фрагментов начальной паузы (200 мс), можно определить пороговые значения $LE_{tresh.}$ и $Z_{tresh.}$. На рисунке 1а представлена графическая интерпретация формирование пороговых значений по шести первым модам. Пунктирными линиями красного цвета отмечены значения параметров ЭМ фрагментов начальной паузы, утолщенной сплошной линией красного цвета отмечены усредненные пороговые значения параметров ЭМ.



а. Формирование пороговых значений



б. Пороговая обработка

Рисунок 1 - Анализ параметров ЭМ фрагментов речевого сигнала

На рисунке 1б представлена графическая интерпретация пороговой обработки. Утолщенной сплошной линией зеленого цвета отмечены значения параметров ЭМ вокализованного фрагмента речи. Утолщенной линией синего цвета - для невокализованного фрагмента.

Для исследования новой технологии сегментации сформирована группа исследуемых – 220 человек с признаками психоэмоциональных расстройств.

Зарегистрирована база данных речевых сигналов. Эффективность сегментации оценивалась посредством коэффициента действительного обнаружения (*Detection Rate*, *DR*) - безразмерной величины, равной отношению правильно определенных фрагментов к общему числу фрагментов. В качестве эталона использовалась сегментация речевых сигналов в ручном режиме в специализированном аудио редакторе. В таблице 1 представлены усредненные результаты сегментации речевых сигналов на вокализованную, невокализованную речь и паузы в сравнении с другими способами.

Таблица 1 - Результаты сегментации речевых сигналов

Информативные участки	***DR*, %**		
	Способ на основе анализа *STE+ZCR*	Способ на основе анализа расстояния Малаханобиса	Способ на основе новой технологии
Вокализованная речь	88,6	46,4	78,5
Невокализованная речь	78,5	90,2	58,6
Паузы	97,5	77,3	91,4

В соответствии с полученными значениями коэффициента действительного обнаружения можно сделать вывод, что способ на основе новой технологии обеспечивает наилучшие результаты сегментации речевых сигналов, зарегистрированных с признаками психоэмоциональных расстройств. Данные результаты достигаются за счет исследования свойств каждой ЭМ анализируемого фрагмента, позволяющего точно определять границы информативных участков речи при нестабильной работе речевого аппарата. Таким образом, предлагаемая технология сегментации, может успешно тестироваться на этапах предварительной обработки в системах обнаружения и оценки психоэмоционального расстройства человека.

Литература

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

2. Huang, N. E. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis / N. E. Huang, Sh. Zheng, R. L. Steven // Proc. R. Soc. Lond. - 1998. - A 454. - P. 903 - 995.

3. Алимурадов А. К. 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.

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