Алгоритм измерения акустических паттернов речевых сигналов естественно выраженных психоэмоциональных состояний

А. К. Алимурадов[1], А. В. Агейкин[2]
Пензенский государственный университет
[1]alansapfir@yandex.ru, [2]keokushinkai@yandex.ru

***Аннотация*. Разработан алгоритм измерения акустических паттернов речевых сигналов, отражающих естественно выраженные психоэмоциональные состояния. Суть алгоритма заключается в сегментации речи на информативные участки с помощью адаптивной декомпозиции и энергетического анализа эмпирических мод; измерении акустических паттернов временных интервалов вокализованных, невокализованных участков и участков пауз. Проведено исследование с использованием сформированной базы речевых сигналов 100 испытуемых, переживаемых естественные положительные и отрицательные эмоции. Результаты исследований оценивались в сравнении с известным способом сегментации на основе детектора голосовой активности, с последующим измерением акустических паттернов. В соответствии с результатами разработанный алгоритм точнее определяет естественно выраженные психоэмоциональные положительные и отрицательные состояния: ошибка 1-го рода 12 % и 11 %, ошибка 2-го рода 5 % и 4%.**

Ключевые слова: обработка речевых сигналов; сегментация; адаптивная декомпозиция; акустические паттерны; естественно выраженные психоэмоциональные состояния

I. Введение

Речь представляет собой один самых сложных приобретаемых навыков человека, чрезвычайно чувствительный к нарушениям работы нервной системы. На протяжении долгих лет оценка нестабильности моторики речевого аппарата при естественно выраженных психоэмоциональных состояниях, ограничивалась перцептивными тестами или лабораторным анализом. На сегодняшний день эта задача успешно решается методами на основе анализа речевых сигналов [1]. Вид и степень выраженности психоэмоциональных состояний кодируются в паттерны речевых сигналов. Акустические паттерны в большей степени характеризуют не сами психоэмоциональные состояния, а их поведение в течение времени. Идея использования акустических паттернов для оценки психоэмоциональных состояний основывается на том факте, что люди используют акустические вариации (интонация, темп и др.), чтобы подчеркнуть значимость отдельных элементов речи [1].

В данной работе авторами представлен алгоритм измерения акустических паттернов речевых сигналов, соответствующих естественно выраженным психоэмоциональным состояниям. В статье авторы решают две основные задачи: выбор и обоснование способа обработки, адаптивного к нарушениям моторики речевого аппарата[1]; поиск релевантного набора паттернов естественно выраженных психоэмоциональных состояний[2]. Исследования являются продолжением ранее опубликованных трудов авторов [2, 3].

Первый этап работы алгоритма (рис. 1) заключается в сегментации речи на информативные участки с помощью адаптивной декомпозиции и энергетического анализа эмпирических мод (ЭМ). Второй этап заключается в измерении акустических паттернов временных интервалов вокализованных, невокализованных участков и участков пауз, отражающих нарушения моторики речевого аппарата, вызванных психоэмоциональным расстройством.

II. Материалы и методы

A. Адаптивная декомпозиция

Сегментация слитной речи на информативные участки представляет собой деление сигнала на равные кратковременные фрагменты для последующей адаптивной декомпозиции; энергетический анализ ЭМ на основе функционала слухового аппарата человека; формирование порога сегментации вокализованных, невокализованных участков и участков пауз на основе физиологического аспекта воспроизведения речи.

Важным условием адаптивной декомпозиции является возможность формирования адаптивного базиса, функционально зависимого от внутренней структуры исходного сигнала. Такой подход реализуется в 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 – ЭМ.

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

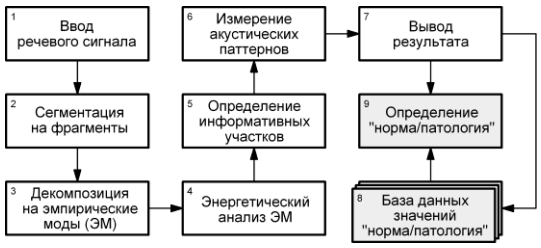


Рис. 1. Упрощенная блок-схема алгоритма измерения акустических паттернов речевых сигналов естественно выраженных психоэмоциональных состояний

Особенность ДЭМ заключается в том, что базисные функции, используемые для декомпозиции, извлекаются непосредственно из структуры исходного сигнала и позволяют учитывать только ему свойственные особенности (скрытые модуляции, области концентрации энергии и т.п.).

Анализ разновидностей методов ДЭМ [5], выявил наиболее адаптивный к речевым сигналам метод улучшенной полной множественной декомпозиции на эмпирические моды с адаптивным шумом (ПМДЭМАШ). Применение улучшенной ПМДЭМАШ обеспечит решение известных проблем ДЭМ [5]: смешивание несоизмеримых по амплитудному и частотному масштабам ЭМ; наличие минимального остаточного шума в ЭМ; отсутствие паразитных ЭМ, возникающих на ранних этапах декомпозиции вследствие перекрытия масштабно-энергетических пространств мод.

Аналитическое выражение декомпозиции 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).

B. Акустические паттерны

Обзор информативных параметров речевых сигналов [6] выявил следующие релевантные акустические паттерны естественно выраженных психоэмоциональных состояний:

- скорость распределения временных интервалов речи (*rate of speech timing*, *RST*), включая вокализованные, невокализованные участки и участки пауз;
- ускорение распределения временных интервалов речи (*acceleration of speech timing*, *AST*), включая вокализованные, невокализованные участки и участки пауз;
- продолжительность интервалов пауз (*duration of pause intervals*, *DPI*);
- энтропия распределения временных интервалов речи (*entropy of speech timing*, *EST*).

III. Описание алгоритма

Речь представляет собой процесс, спектр которого остается относительно неизменным в течение короткого периода времени. Это позволяет разделить речевой сигнал на равные фрагменты по 10 мс, в пределах которых можно считать сигнал условно стационарным. После сегментации сигнал представляет собой набор фрагментов, и дальнейшая работа алгоритма осуществляется с каждым фрагментом в отдельности.

Декомпозиция осуществлялась методом улучшенной ПМДЭМАШ. В результате декомпозиции каждый фрагмент представлен набором ЭМ.

Изменение амплитуды речевого сигнала во времени характеризуется важным информативным параметром – амплитудным распределением, которое описывается с помощью функции кратковременной энергии сигнала. В соответствии с функционалом слухового аппарата, человек воспринимает речь нелинейно, определяя разницу между энергиями элементов речи. Приближая работу алгоритма к функционалу слухового аппарата, для сжатия амплитуды сигнала в большом динамическом диапазоне применяют логарифмирование энергии:

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

где $LE_{s,i}$ – логарифм энергии ЭМ фрагмента речевого сигнала, s – номер фрагмента.

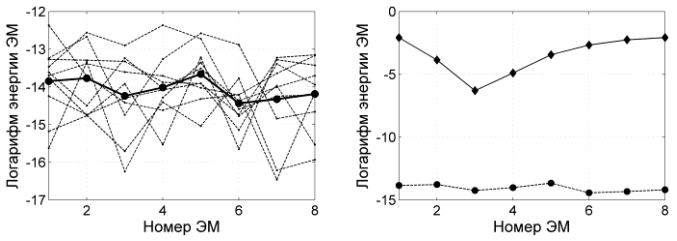
Определение информативных участков заключается в обнаружении точных границ вокализованных, невокализованных участков и участков пауз. Сегментация всех участков одновременно неэффективна. Для повышения точности сегментации необходимо последовательное разделение речевого сигнала на вокализованные, невокализованные участки и участки пауз.

В разработанном алгоритме сегментация осуществляется на основе энергетического анализа ЭМ фрагментов речевого сигнала в скользящем окне 10 мс. В соответствии с физиологическим аспектом человек перед воспроизведением речи делает кратковременную паузу – обычно от 200 до 500 мс, не содержащую речь и соответствующую тишине с фоновым шумом. Используя усредненные значения логарифмов энергии ЭМ фрагментов данного участка, можно определить пороговые значения логарифмов энергии разделения на полезный сигнал (вокализованные и невокализованные участки) и паузы. Определение пороговых значений логарифмов энергии ЭМ осуществляется по формуле:

$$LE_{thres.,i}(n) = \frac{1}{S}\sum_{s=1}^{S} LE_{IMFs,i},$$

где $LE_{thres.,i}$ – пороговое значение логарифмов энергии ЭМ, $LE_{IMFs,i}$ – логарифм энергии ЭМ.

На рис. 2а представлена графическая интерпретация определения пороговых значений.



а. Определение пороговых значений

б. Пороговая обработка при разделении речевого сигнала на полезный сигнал и паузы

Рис. 2. Энергетический анализ ЭМ

Штриховыми линиями отмечены значения логарифмов энергии ЭМ фрагментов участка, не содержащего речь и соответствующего тишине с фоновым шумом, утолщенной сплошной линией отмечены усредненные пороговые значения логарифмов энергии ЭМ. После определения пороговых значений, осуществляется первый этап разделения речевого сигнала на полезный сигнал и паузы. На рис. 2б представлена графическая интерпретация пороговой обработки при разделении речи на полезный сигнал и паузы. Штриховой линией отмечены пороговые значения логарифмов энергии ЭМ фрагментов, не содержащие речь, а сплошной линией – значения логарифмов энергии ЭМ фрагментов полезного сигнала.

Следующий этап сегментации заключается в разделении вокализованных и невокализованных участков. Сегментация осуществлялась также на основе энергетического анализа ЭМ в скользящем окне 10 мс и дополнительного вычисления следующих параметров для каждого фрагмента: скорость пересечения сигнала через нулевое значение (*Zero-Crossing Rate, ZCR*); автокорреляционная функция (*autocorrelation function*, *ACR*); энергия/мощность фрагмента (*PWR*). Значения *PWR*, *ACR* и *ZCR* используются для повышения точности сегментации и определяются с использованием следующих выражений:

$$PWR = \frac{1}{N}\sum_{n=1}^{N} x^2(n) * h(n),$$

$$R_x(k) = \frac{1}{N * \sigma_x^2}\sum_{n=1}^{N}\left(x(n)-\mu_x\right)*\left((n+k)-\mu_x\right),$$

$$ACR = \frac{1}{N-1}\sum_{n=1}^{N}\left(R_x(k)-\overline{R_x}\right),$$

$$ZCR = \frac{1}{N-1}\sum_{n=1}^{N-1}\left|sign\left(R_x(n+1)\right)-sign\left(R_x(n)\right)\right|,$$

$$sign\left(R_x(n)\right) = \begin{cases} 1, & x(n) \ge 0 \\ -1, & x(n) < 0 \end{cases},$$

где $x(n)$ представляет собой речевой сигнал в окне длинной N отсчетов, $h(n)$ – окно Хемминга, R_x представляет функцию автокорреляции, σ_x – стандартное отклонение сигнала и μ_x – среднее значение сигнала.

На рис. 3 представлен реализованный в разработанном алгоритме процесс сегментации на информативные участки.

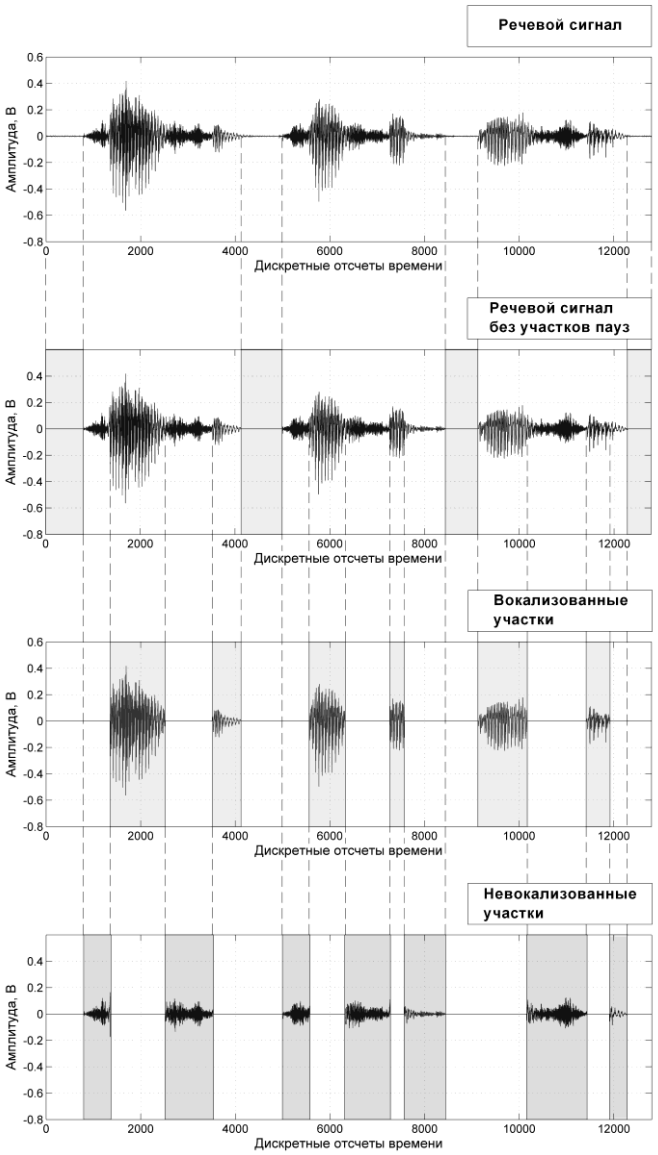


Рис. 3. Сегментация на информативные участки

Измерение акустических паттернов заключается в определении значений *RST*, *AST*, *DPI* и *EST*.

RST обеспечивает более точную оценку ухудшения скорости речи, чем простое измерение длительности пауз, поскольку данный паттерн учитывает не только паузы, но и вокализованные, невокализованные участки. Вокализованные участки предоставляют дополнительную информацию об ухудшении фонации, тогда как невокализованные участки предоставляют информацию о нечеткой артикуляции. *RST* приблизительно в комплексе равна скорости речи, поскольку ухудшение скорости речи

связано с недостатками во всех элементах речевых сигналов. Каждый вокализованный, невокализованный участок и участки пауз описываются временем возникновения, определяемым как среднее значения времени между началом участка и его окончанием. Общее число участков подсчитывалось для каждого момента в течение измерения.

AST определяет степень ускорения времени. Каждый анализируемый фрагмент речевого сигнала делится на две части с перекрытием 25%, что обеспечивает плавный переход между частями. Значение *AST* рассчитывается как разница между значениями *RST* обеих частей, разделенное на общую продолжительность фрагмента речевого сигнала.

DPI оценивает способность диктора начать воспроизведение речи. Сложное нарушение речи может вызвать трудности при воспроизведении, которые порождают удлинение пауз. *DPI* вычислялся как среднее время всех участков пауз.

EST описывает упорядоченность или предсказуемость речи, включая вокализованные, невокализованные участки и участки пауз. Соответственно, уменьшение энтропии равнозначно нарушенной речи. Вычисляется количество всех интервалов речи, в том числе количество вокализованных участков *kv*, количество невокализованных участков *ku*, число участков паузы *kp* и общее количество участков *kt*. *EST* была определена следующим образом:

$$EST = -\frac{kv}{kt} * \log_2\left(\frac{kv}{kt}\right) - \frac{ku}{kt} * \log_2\left(\frac{ku}{kt}\right) - \frac{kp}{kt} * \log_2\left(\frac{kp}{kt}\right).$$

IV. Исследование алгоритма

Для оценки эффективности разработанного алгоритма при поддержке Областной клинической больницы им. К.Р. Евграфова (г. Пенза) и медицинского института Пензенского государственного университета сформирована группа испытуемых и зарегистрирована верифицированная база речевых сигналов. В группу испытуемых отобрано 100 чел. мужского и женского пола, в возрасте от 18 до 60 лет, переживаемых естественные положительные и отрицательные эмоции. 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 алгоритма оценивались в сравнении с известным способом сегментации на основе детектора голосовой активности (*voice activity detection*, *VAD*) с последующим измерением акустических паттернов *RST*, *AST*, *DPI* и *EST*.

В таблице представлены результаты определения психоэмоциональных положительных и отрицательных состояний.

ТАБЛИЦА I Результаты определения психоэмоциональных состояний

Прогнозируемый результат	Результат определения		Ошибки первого и второго рода, %	
	Патология	*Норма*		
Способ на основе детектора голосовой активности (VAD)				
Положительное психоэмоциональное состояние				
Патология	63 чел.	27 чел.	1-ого	27
Норма	12 чел.	88 чел.	2-ого	12
Отрицательное психоэмоциональное состояние				
Патология	81 чел.	19 чел.	1-ого	19
Норма	7 чел.	93 чел.	2-ого	7
Разработанный алгоритм				
Положительное психоэмоциональное состояние				
Патология	88 чел.	12 чел.	1-ого	12
Норма	5 чел.	95 чел.	2-ого	5
Отрицательное психоэмоциональное состояние				
Патология	89 чел.	11 чел.	1-ого	11
Норма	4 чел.	96 чел.	2-ого	4

Из таблицы видно, что процент ложных присваиваний статуса «норма» речевым сигналам, произнесенным пациентами, находящимися в состоянии психоэмоционального возбуждения у *VAD* слишком велик: 27% и 19% соответственно. То же самое можно сказать о ложных присваиваниях статуса «патология» речевым сигналам, произнесенным пациентами в нейтральном состоянии: 12% и 7% соответственно. Лучшие значения ошибок 1-ого и 2-ого родов были достигнуты разработанным алгоритмом: всего лишь 12% и 11% для положительных эмоций, 5% и 4% для отрицательных эмоций. Исходя из результатов, можно сделать вывод, что разработанный алгоритм точнее измеряет акустические паттерны *RST*, *AST*, *DPI* и *EST* психоэмоциональных состояний. Это достигается за счет корректной сегментации информативных участков речевых сигналов методом улучшенной ПМДЭМАШ и энергетического анализа ЭМ.

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, pp. 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, pp. 26-33.

[4] 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. Vol. A454. pp. 903-995.

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

[6] Hlavnička J. 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). pp. 13.