УДК 004.934
doi:10.21685/2587-7704-2021-6-2-2

Open Access RESEARCH ARTICLE

Применение энергетического оператора Тигера в задаче сегментации речевых сигналов

Алан Казанферович Алимурадов
Пензенский государственный университет, Россия, г. Пенза, ул. Красная, 40
alansapfir@yandex.ru

Аннотация. Представлен способ сегментации речевых сигналов на основе анализа фрагментов речевых сигналов с помощью энергетического оператора Тигера и последующего анализа значений кратковременной энергии и количества пересечения через нулевую ось функции энергетической характеристики. Проведено исследование предложенного способа, в рамках которого выявлено, что за счет хорошей восприимчивости энергетического оператора Тигера к кратковременным изменениям амплитуды и частоты речевых сигналов предложенный способ обеспечивает повышение эффективности сегментации на 2,97 и 2,49 % для ошибок первого и второго рода соответственно.

Ключевые слова: обработка речи, сегментация речи, вокализованная, невокализованная речь и паузы, энергетический оператор Тигера

Для цитирования: Алимурадов А. К. Применение энергетического оператора Тигера в задаче сегментации речевых сигналов // Инжиниринг и технологии. 2021. Т. 6(2). С. 1–5. doi:10.21685/2587-7704-2021-6-2-2

Application of Teager energy operator for speech signal segmentation

Alan K. Alimuradov
Penza State University, 40 Krasnaya Street, Penza, Russia
alansapfir@yandex.ru

Abstract. The article presents a method for speech signal segmentation based on the analysis of speech signal fragments using the Teager energy operator and subsequent analysis of the values of short-term energy and zero-crossing rate of the energy characteristic function. The research has revealed that the proposed method provides an increase in the efficiency of segmentation by 2.97 % and 2.49 % for the first and second kind errors, respectively, due to good susceptibility of the Teager energy operator to short-term changes in the amplitude and frequency of speech signals.

Keywords: speech processing, speech segmentation, voiced and unvoiced speech, pauses, Teager energy operator

For citation: Alimuradov A.K. Application of Teager energy operator for speech signal segmentation. *Inzhiniring i tekhnologii = Engineering and Technology.* 2021;6(2):1–5. (In Russ.). doi:10.21685/2587-7704-2021-6-2-2

Задача сегментации речевых сигналов представляет собой точное обнаружение границ начала и окончания информативных участков: пауз, вокализованной и невокализованной речи. На сегодняшний день задача сегментации речевых сигналов решается разными способами во временной и частотной области [1]. К временным относятся способы на основе анализа количества пересечения сигнала через нулевую ось (ПСЧН), отклонения автокорреляционной функции (АКФ), кратковременной энергии (КрЭ), а также одномерного расстояния Махаланобиса. К частотным относятся способы на основе анализа мел-частотных кепстральных коэффициентов (МЧКК) и линейно-частотных кепстральных коэффициентов (ЛЧКК).

В статье представлен способ сегментации речевых сигналов на основе энергетического оператора Тигера (ЭОТ). Предлагаемый способ представляет собой модернизацию существующего спосо-

© Алимурадов А. К., 2021. Контент доступен по лицензии Creative Commons Attribution 4.0 License / This work is licensed under a Creative Commons Attribution 4.0 License.

ба сегментации на основе анализа значений КрЭ и количества ПСЧН. Статья является результатом научной работы [2], выполняемой в рамках проекта «Поисковые исследования паттернов речевых сигналов, релевантных естественно выраженным эмоциям человека, и разработка системы обнаружения и классификации психоэмоциональных состояний для эксплуатации в условиях повышенной ответственности». Заявка № 21-19-00668 на получение финансирования данного проекта находится на экспертизе в рамках конкурса 2021 г. «Проведение фундаментальных научных исследований и поисковых научных исследований отдельными научными группами» Российского научного фонда.

Способы сегментации речевых сигналов на основе анализа значений КрЭ и количества ПСЧН применяются ограниченно. Это связано с невозможностью выбора и обоснования корректных пороговых значений, соответствующих вокализованной, невокализованной речи и паузам.

Вычисление количества ПСЧН основано на сравнении знаков соседних дискретных отсчетов времени и определяется по следующей формуле:

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

где $x(n)$ – исследуемый сигнал; n – дискретный отсчет времени; s – номер фрагмента; N – количество дискретных отсчетов в исследуемом фрагменте; $sgn(x)$ – знаковая функция ($sgn(x) = 1$ при $x \geq 0$ и $sgn(x) = -1$ при $x \leq 0$).

Вычисление КрЭ представляет собой нахождение суммы квадратов амплитуд дискретных отсчетов сигнала для короткой последовательности (фрагмента) и определяется по следующей формуле:

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

Анализ количества ПСЧН построен на предположении, что количество пересечений функции сигнала с нулевой осью для пауз с фоновым шумом больше по сравнению с вокализованной и невокализованной речью. Аналогично построен анализ КрЭ: энергия вокализованной и невокализованной речи больше, чем энергия пауз с фоновым шумом. Однако данные предположения не всегда корректны. Не решен главный вопрос: насколько текущие значения КрЭ и количества ПСЧН должны быть больше, чем пороговые, для корректной сегментации речевых сигналов. Кроме того, известно, что пороговые значения могут варьироваться для каждого конкретного анализируемого речевого сигнала.

ЭОТ – это дифференциальный энергетический оператор второго порядка, позволяющий вычислять энергетические характеристики сигнала [3]. На сегодняшний день ЭОТ получил широкое практическое применение в задачах обработки речевых сигналов, в том числе для сегментации на информативные участки [4]. Для дискретных сигналов ЭОТ имеет следующий вид:

$$TEO\left(n\right) = x\left(n\right)^2 - x\left(n-1\right)\cdot x\left(n+1\right).$$

На рис. 1 структурно представлен способ сегментации речевых сигналов на основе энергетического анализа фрагментов речевого сигнала с помощью ЭОТ и последующего анализа значений КрЭ и количества ПСЧН.

Суть сегментации заключается в линейном разделении речевого сигнала на фрагменты (блок 1); вычислении энергетической характеристики речевого сигнала с помощью ЭОТ (блок 2); вычислении значений КрЭ и количества ПСЧН фрагментов энергетической характеристики (блок 3, 4); определении статуса «речь/пауза» фрагментов (блок 7) на основе вычисленных пороговых значений КрЭ и количества ПСЧН (блок 5, 6). Блоки 8 и 9 не относятся к способу и предназначены для постобработки ошибок сегментации, а также для сравнения результатов с сегментацией, осуществленной вручную.

Для корректной сегментации речевых сигналов в предложенном способе представлено решение проблемы выбора пороговых значений КрЭ и количества ПСЧН. Предлагается использовать начальную паузу в качестве исходных данных для формирования пороговых значений КрЭ и количества ПСЧН. Вычисляются математическое ожидание μ_E, μ_{ZCR} и дисперсия σ_E, σ_{ZCR} значений КрЭ и количества ПСЧН для фрагментов, соответствующих начальной паузе 200 мс (фоновому шуму):

$$\mu_{ZCR} = \frac{1}{S}\sum_{s=1}^{S} ZCR_s,$$

$$\mu_E = \frac{1}{S}\sum_{s=1}^{S} E_s ,$$

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

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

где ZCR_s, E_s – значения КрЭ и количества ПСЧН исследуемого фрагмента соответственно; S – количество фрагментов, соответствующих фоновому шуму.

Рис. 1. Структура способа сегментации речевых сигналов на основе энергетического анализа фрагментов речевого сигнала с помощью ЭОТ и последующего анализа значений КрЭ и количества ПСЧН

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

$$\frac{\left|ZCR_s - \mu_{ZCR}\right|}{\sigma_{ZCR}} \geq K\sigma_{ZCR} ,$$

$$\frac{\left|E_s - \mu_E\right|}{\sigma_E} \geq K\sigma_E ,$$

где выражения $\left|ZCR_s - \mu_{ZCR}\right|$, $\left|E_s - \mu_E\right|$ являются естественной мерой одномерного расстояния Махаланобиса от текущих значений КрЭ и количества ПСЧН к средним значениям, соответствующим фоновому шуму; K – коэффициент порога (K всегда больше 1).

Если разница между текущим и средним значениями количества ПСЧН больше или равна $K\sigma_{ZCR}$, то фрагмент соответствует паузе. И наоборот, если условие не выполняется, то фрагмент соответствует речи. Аналогично, если разница между текущим и средним значениями КрЭ больше или

равна $K\sigma_E$, то фрагмент соответствует речи. И наоборот, если условие не выполняется, то фрагмент соответствует паузе.

Для оценки предложенного способа сегментации сформирована база речевых сигналов. Эффективность сегментации речевых сигналов оценивалась посредством определения ошибок первого (α) и второго (β) рода. В рамках исследования предложенного способа оценивалось влияние коэффициента порога на эффективность сегментации речевых сигналов в сравнении с классическим способом сегментации на основе анализа значений КрЭ и количества ПСЧН. В табл. 1 представлены усредненные значения ошибок первого и второго рода для классического способа сегментации речевых сигналов и предложенного способа.

Таблица 1

Усредненные значения ошибок первого и второго рода для способа сегментации речевых сигналов на основе анализа КрЭ и количества ПСЧН, предложенного способа на основе энергетического анализа с помощью ЭОТ и последующего анализа значений КрЭ и количества ПСЧН

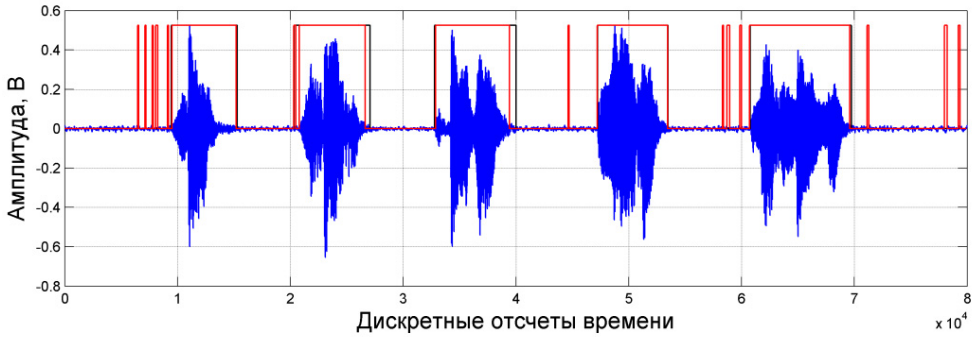
Значение коэффициента порога	Способ на основе анализа значений КрЭ и количества ПСЧН		Способ на основе энергетического анализа с помощью ЭОТ и последующего анализа значений КрЭ и количества ПСЧН	
	Ошибки первого и второго рода, %			
	α	β	α	β
1	1,37	36,06	0,92	47,25
2	1,60	15,10	0,92	22,03
3	5,03	4,44	0,92	13,68
4	6,64	1,78	0,92	10,48
5	7,09	1,24	0,92	7,46
6	8,24	0,89	1,37	5,51
7	11,67	0,89	1,37	3,91
8	12,59	0,89	1,83	2,66
9	14,42	0,89	1,83	2,66
10	16,02	0,89	2,06	1,95
11	17,39	0,89	2,29	1,95
12	18,08	0,89	2,52	1,95
13	18,76	0,89	2,52	1,95
14	18,99	0,89	2,52	1,95
15	19,45	0,89	2,52	1,95

Анализ полученных результатов в табл. 1 выявил, что наиболее оптимальные значения ошибок первого и второго рода достигаются предложенным способом – 2,06 и 1,95 % соответственно при значении коэффициента порога, равном 10. Для классического способа сегментации речевых сигналов оптимальные значения ошибок первого и второго рода достигаются при значении коэффициента, равном 3, – 5,03 и 4,44 %.

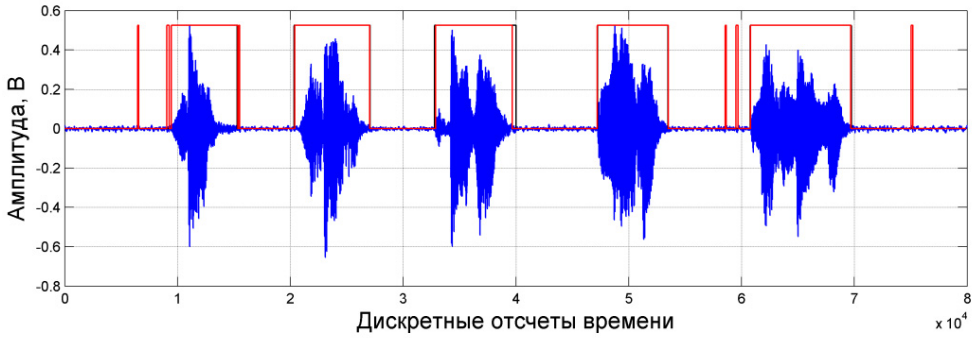
При сравнении оптимальных значений ошибок первого и второго рода предложенный способ обеспечивает повышение эффективности сегментации речевых сигналов на 2,97 и 2,49 % соответственно. Это обеспечивается за счет хорошей восприимчивости ЭОТ к кратковременным изменениям амплитуды и частоты речевых сигналов.

На рис. 2 представлен пример, иллюстрирующий результаты сегментации речевого сигнала длительностью 10 с, представляющего собой сочетание следующих 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.

Как видно из рис. 2, ошибки сегментации предложенного способа (рис. 2,***б***) в основном наблюдаются в пограничных областях между участками речи и пауз, так как в большинстве практических случаев параметры глухих сонорных и шумных (смычных, щелевых) звуков соответствуют параметрам паузы с фоновым шумом. Как правило, ошибочно сегментированные участки имеют длительность менее 20 мс. Следовательно, мелкие ошибки сегментации в пограничных областях практически не будут влиять на эффективность обработки речевых сигналов.



а)



б)

Рис. 2. Пример, иллюстрирующий результаты сегментации речевого сигнала (линией красного цвета обозначены достигнутые результаты сегментации, линией черного цвета – результат сегментации, осуществленной вручную): ***а*** – способ сегментации речевых сигналов на основе анализа значений КрЭ и количества ПСЧН; ***б*** – способ сегментации речевых сигналов на основе энергетического анализа с помощью ЭОТ и последующего анализа значений КрЭ и количества ПСЧН

Список литературы

1. Huang X., Acero A., Hon H.-W. Spoken Language Processing. Guide to Algorithms and System Development // Prentice Hall. New Jersey, 2001. 980 p.
2. Алимурадов А. К., Тычков А. Ю., Чураков П. П. Способ автоматизированной сегментации речевых сигналов для определения временных паттернов естественно выраженных психоэмоциональных состояний // Измерение. Мониторинг. Управление. Контроль. 2019. № 3 (29). С. 48–60.
3. Kaiser J. F. On a simple algorithm to calculate the ‘energy’ of a signal // International Conference on Acoustics, Speech, and Signal Processing (Albuquerque, NM, USA, April 3–6, 1990). Albuquerque, USA, 1990. Vol. 2. P. 381–384.
4. Жуйков В. Я., Харченко А. Н. Алгоритм классификации сегментов речевого сигнала // Электроника и связь. Тематический выпуск «Электроника и нанотехнологии». Ч. 1. 2009. № 2–3, С. 130–137.

References

1. Huang X., Acero A., Hon H.-W. Spoken Language Processing. Guide to Algorithms and System Development. *Prentice Hall*. New Jersey, 2001:980.
2. Alimuradov A.K., Tychkov A.Yu., Churakov P.P. A method for automated segmentation of speech signals to determine temporal patterns of naturally expressed psycho-emotional states. *Izmerenie. Monitoring. Upravlenie. Kontrol' = Measuring. Monitoring. Management. Control*. 2019;3(29):48–60. (In Russ.)
3. Kaiser J.F. On a simple algorithm to calculate the ‘energy’ of a signal. *International Conference on Acoustics, Speech, and Signal Processing (Albuquerque, NM, USA, April 3–6, 1990)*. Albuquerque, USA, 1990;2:381–384.
4. Zhuykov V.Ya., Kharchenko A.N. Algorithm for classification of speech signal segments. *Elektronika i svyaz'. Tematicheskiy vypusk «Elektronika i nanotekhnologii». Ch. 1. = Electronics and Communications. Thematic issue “Electronics and Nanotechnology”. Part 1.* 2009;2–3:130–137. (In Russ.)

Поступила в редакцию / Received 10.04.2021

Поступила после рецензирования и доработки / Revised 20.05.2021

Принята к публикации / Accepted 07.06.2021