МЕТОДЫ ПРИМЕНЕНИЯ VAD В СИСТЕМАХ РАСПОЗНАВАНИЯ КАЗАХСКОЙ РЕЧИ

М. Н. Калимолдаев, О. Ж. Мамырбаев*, Р. Р. Мусабаев, Б. Б. Тусупова*

Институт проблем информатики и управления Министерства образования и науки Республики Казахстан, 050010, Алма-Ата, Казахстан

* Казахский национальный технический университет им. К. И. Сатпаева, 005013, Алма-Ата, Казахстан

УДК 519.7

Рассмотрена возможность применения алгоритма "Voice activity detection" в системе распознавания казахской речи. Предложены математическая модель VAD и способы обнаружения речевых данных: пауз между фразами, словами, отдельными звуками. Алгоритм VAD приспособлен к распознаванию казахской речи с учетом ее основных свойств. Впервые проведено исследование обнаружения голосовой активности в казахской.

Ключевые слова: распознавание речи, обнаружение голосовой активности, речевой сигнал.

This article considers the algorithm "Voice activity detection" and the using VAD algorithm in the system of kazakh speech recognition. The paper presents a mathematical model VAD and methods for detecting voice data: pauses between sentences, words, individual sounds. VAD algorithm is adapted to the recognition of Kazakh speech counting the basic properties of Kazakh language. Voice activity detection researches in Kazakh speech are being conducted for the first time. The results of the spectral analysis are displayed on the picture.

Key words: speech recognition, voice activity detection, speech signal.

Введение. Исследования в области распознавания речи ведутся достаточно давно. Речь как природный источник информации обладает избыточностью, в ней содержится большое количество данных, не несущих смысловой нагрузки.

В настоящее время для увеличения объемов передаваемой информации применяются различные методы, например частотное и временное уплотнение сигналов. Для выполнения задачи распознавания речи в первую очередь необходимо определить моменты начала и окончания входного слова и пауз внутри него [1].

Постановка задачи. Определение моментов начала и окончания фразы при наличии шума является важной задачей распознавания речи. В частности, при автоматическом распознавании речи важно точно определить моменты начала и окончания слова [2].

Процедура обнаружения моментов начала и окончания фразы существенно уменьшает число арифметических операций, если обрабатывать только те сегменты, в которых имеется речевой сигнал. Вследствие этого скорость обработки будет увеличиваться. Наиболее распространенным способом сжатия речевых данных является удаление пауз между фразами, словами, отдельными звуками. Как показали многочисленные исследования, в речи может содержаться до 50 % пауз, а в диалоге их объем может достигать 70 %. Поэтому были созданы различные алгоритмы, которые устраняют избыточность речи, выделяя только значимые ее параметры [3].

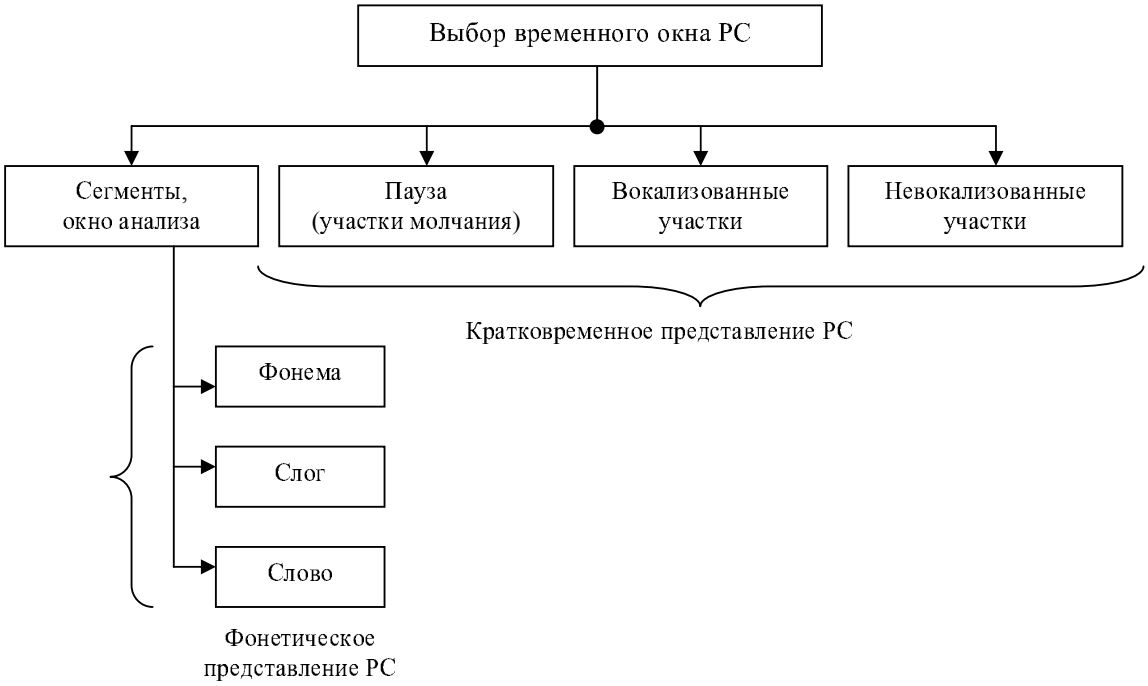

Рис. 1. Схема классификации кадров РС

Voice activity detector (VAD) — метод определения активности речи, технология сжатия речевого сигнала за счет поиска речи и пауз и их кодирования. В системах распознавания речи эффективность системы распознавания определяется в первую очередь эффективностью использования VAD [4].

Алгоритм VAD работает в процессе кодирования речевого сигнала перед распознаванием речи. Наличие пауз определяется на основе анализа и синтеза речевых данных, которые содержат отрезки сигнала. Предположим, речь содержит паузу, которую можно предсказать, и данный пакет содержит паузу, а не речь — наиболее сложный элемент алгоритма VAD. В наиболее простой реализации наличие паузы в наборе цифровых отсчетов определяется на основе сравнения суммарной энергии пакета речевых данных с некоторым пороговым значением, которое отделяет паузу от пакета с голосом. В этом случае порог необходимо подобрать таким образом, чтобы не допустить чрезмерно частое устранение ошибочных пауз, так как это может привести к ухудшению качества, потере важных данных и как следствие к снижению эффективности алгоритма VAD. Обычно для определения пауз применяется сложный алгоритм, учитывающий не только энергию пакета, но и энергию спектральных составляющих отрезка сигнала [5, 6].

Алгоритм разделения речевого сигнала на вокализованные и невокализованные участки и участки молчания. Звуки речи, в которых присутствует основной тон, называются вокализованными. При исследовании динамики изменения характеристик речевого сигнала (РС) важной задачей является выбор длительности временных кадров, на которые он разбивается. На рис. 1 представлена схема классификации кадров РС [7].

Длительность кадра РС должна быть достаточно малой, чтобы последовательность кадров более точно отражала кратковременную динамику изменения РС, и достаточно большой, чтобы последовательность кадров более точно отражала долговременную динамику РС.

Согласно условиям регистрации РС, указанным в таблице, длительность его кадра должна быть не меньше периода основного тона $T_{\text{ОТ}} = 1000/100 = 10$ мс. На рис. 2 приведен график речевого сигнала [8].

Речевой сигнал (f_g = 8000 Гц, $f_{от} \geq 100$ Гц)

Число отсчетов	Длительность кадра, мс	Свойства окна
32	32/8 = 4	Отражает кратковременную динамику РС и не отражает его периодический характер
64	64/8 = 8	Отражает кратковременную динамику РС и не полностью отражает его периодический характер
128	128/8 = 16	Не полностью отражает кратковременную и долговременную динамику РС, полностью отражает его периодический характер
256	256/8=32	Не отражает кратковременную динамику РС, отражает долговременную динамику РС, полностью отражает его периодический характер

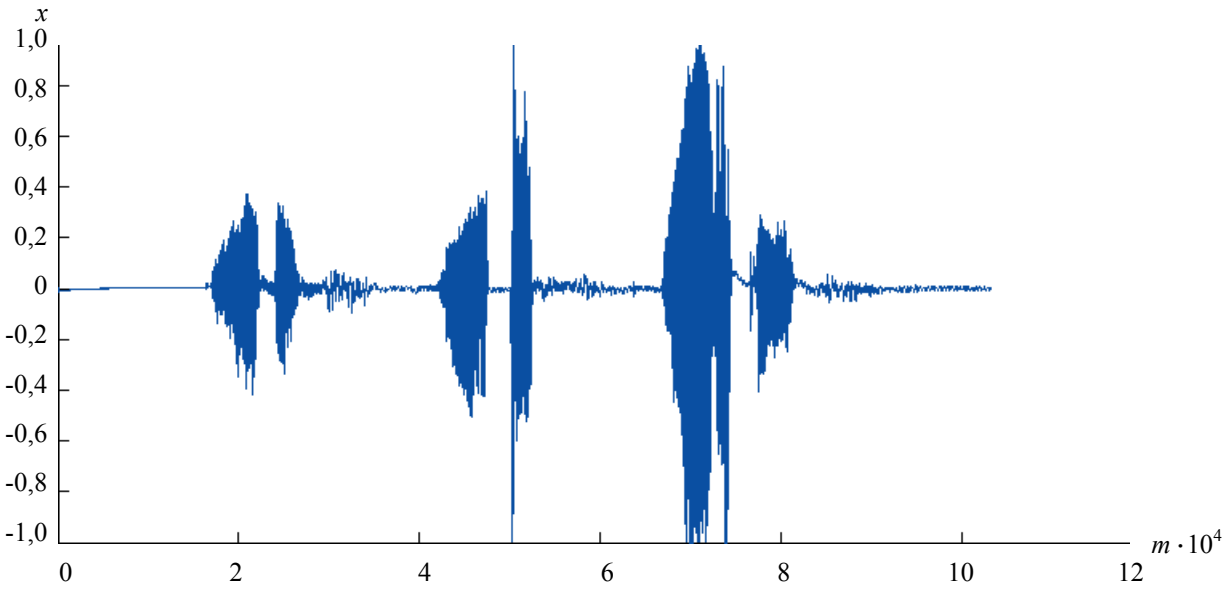

Рис. 2. График речевого сигнала

На рис. 3 представлена блок-схема алгоритма разделения речевого сигнала на вокализованные и невокализованные участки и участки молчания. Данный алгоритм основан на предположении, что речевой сигнал — это нестационарный процесс со значительными изменениями кратковременной энергии и числа пересечений нуля между смежными окнами [9].

Алгоритм включает 7 блоков.

Блок 1. Исходный речевой сигнал $x(m)$, $m = \overline{0, N-1}$.

Блок 2. Разделение РС на кадры длительностью 16 мс.

Блок 3. Вычисление значений кратковременной энергии E_n (или кратковременное значение модуля энергии) и числа пересечений нуля $Z_n n$-го кадра. Например, кратковременная энергия равна $E_n = \sum\limits_{m=n-N+1}^{n} x^2(m)$, или $E_n = \sum\limits_{m=-\infty}^{\infty} [x(m)w(n-m)]^2$, или $E_n = \sum\limits_{m=0}^{N-1} x^2(N-n+m)$, где n — номер кадра;

$$w(m) = \begin{cases} 1, & m = \overline{0, N-1}, \\ 0, & m \neq \overline{0, N-1} \end{cases} -$$

оконная функция кадра; $n = \overline{0, L}$; L — число кадров; $M = LN$ — число отсчетов речевого сигнала.

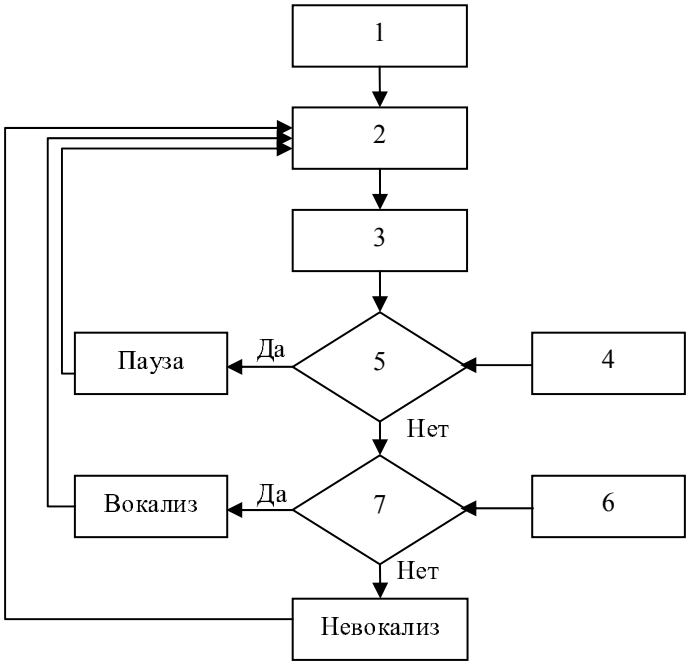

Рис. 3. Блок-схема алгоритма разделения речевого сигнала на вокализованные и невокализованные участки и участки молчания

Кратковременная функция среднего числа переходов через нуль, или нулевых пересечений, основана на сравнении знаков соседних отсчетов [10, 11]. Например,

$$z_n = \sum_{m=-\infty}^{\infty} |\mathrm{sgn}(x(m)) - \mathrm{sgn}(x(m-1))|\, w(n-m),$$

где

$$W(m) = \begin{cases} 1/2, & 0 \le m \le N-1, \\ 0, \end{cases} \qquad \mathrm{sgn}(X(m)) = \begin{cases} 1, & X(m) > 0, \\ -1, & X(m) < 0 \end{cases}$$ –

знаковая функция.

Блоки 4, 6. Установка пороговых значений $E_{\text{пор}}$ и $Z_{\text{пор}}$ для E_n и Z_n.

Блок 5. Проверка выполнения условия $E_n < E_{\text{пор}}$?: да — n-й кадр относится к участку молчания; нет — к блоку 7.

Блок 7. Проверка выполнения условия $Z_n < Z_{\text{пор}}$?: да — n-й кадр относится к вокализованному участку; нет — n-й кадр относится к невокализованному участку.

Недостатком данного алгоритма является высокая чувствительность E_n к большим значениям сигнала. Полученные данные представлены на рис. 4, 5.

Для уменьшения ошибок принятия решения относительно того, является ли участок вокализованным, предлагается использовать соотношение

$$R_{rms} = \frac{E_{rms}}{Z_n},$$

где $E_{rms} = \sqrt{\bar{x}^2(m)} = \sqrt{\frac{1}{N}\sum\limits_{m=1}^{N} x^2(m)}$ — квадратный корень среднего квадратов значений PC (rootmeansquare), или квадратичное среднее.

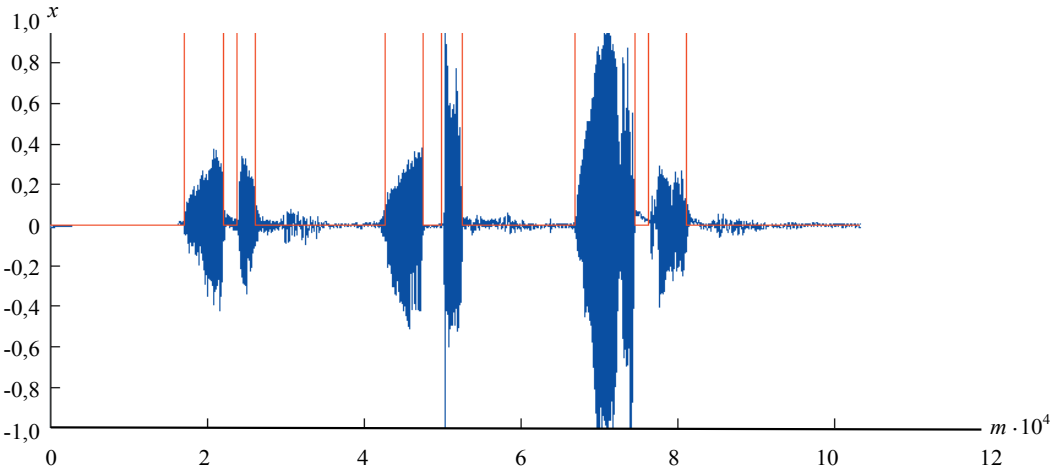

Рис. 4. График определения VAD в речевом сигнале

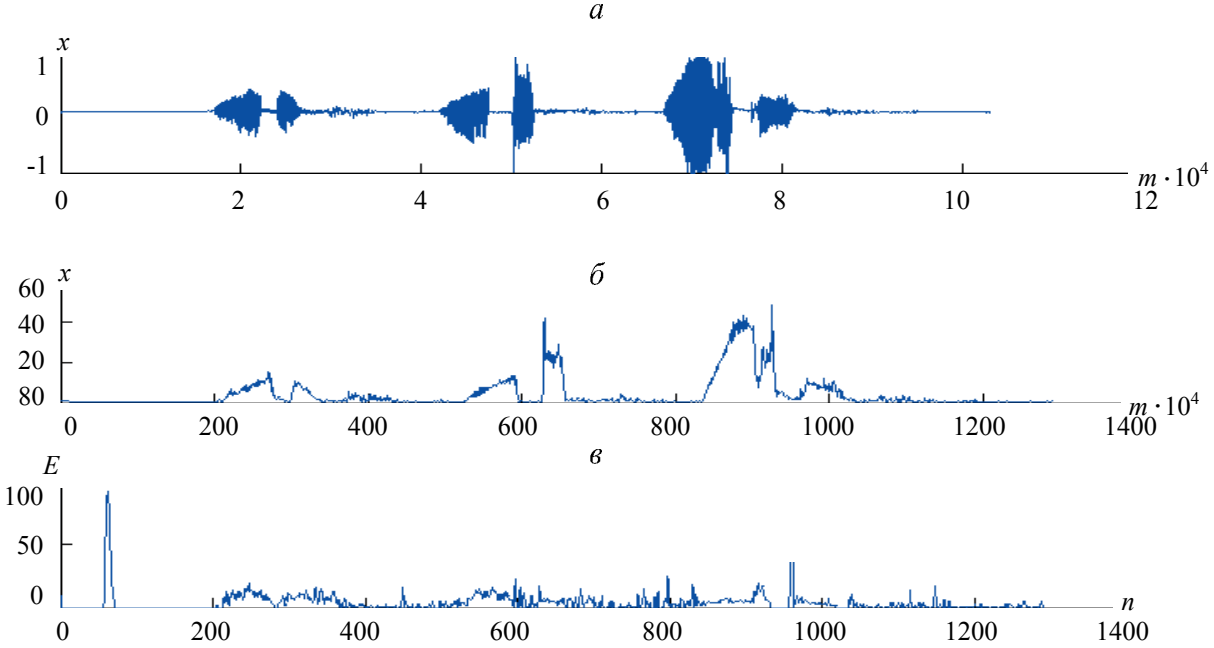

Рис. 5. График определения вокализованных (*а*), невокализованных (*б*) участков и энергии (*в*) в речевом сигнале

Вокализованная речь характеризуется большим значением E_{rms} и малым Z_n, а невокализованная речь характеризуется малым значением E_{rms} и большим Z_n, поэтому справедливо условие: R_{rms} является большим для вокализованного кадра и малым для невокализованного кадра. В данном случае требования к выбору порогового значения R_{rms} являются более простыми, что уменьшает возможность ошибочного принятия решения относительно того, является ли кадр вокализованным.

Выводы. Предложенный алгоритм используется для поиска конечной точки различных изолированных слов. В эксперименте получены графики для казахской речи. Алгоритм позволяет получить более точные результаты по сравнению с результатами поиска конечной точки РС вручную. На рис. 4 приведены примеры РС и обнаружения речевой активности. Для программирования использован язык MATLAB.

На рис. 5 показан процесс определения вокализованных и невокализованных участков, энергии РС. Общее число образцов, необходимых для представления речи разными дикторами, варьируется в зависимости от спектральных характеристик речи.

Алгоритм показывает хороший результат во многих кадрах сегментированной речи для классификации РС. Он эффективен для обнаружения конечных точек различных РС, позволяет снижать требования к объему памяти компьютера и время, затрачиваемое на вычисления. Алгоритм действует более эффективно, чем сегментация, выполняемая вручную.

Список литературы

1. Дорохин О. А., Старушко Д. Г. Сегментация речевого сигнала // Искусств. интеллект. 2000. № 3. С. 450–478.

2. Шелепов В. Ю., Ниценко А. В. Амплитудная сегментация речевого сигнала, использующая фильтрацию и известный фонетический состав // Искусств. интеллект. 2003. № 6. С. 120–123.

3. Lamel L. F., Rabiner L. R., Rosenberg A. E., Wilpon J. G. An improved endpoint detector for isolated word recognition // IEEE Trans. Acoust., Speech, Signal Process. 1981. V. 29, N 4. P. 23–31.

4. Rabiner L. Fundamentals of speech recognition / L. Rabiner, Juang Biing-Hwang. Englewood Cliffs: Prentice Hall, 1993.

5. Deller J. R. (Jr.). Discrete-time processing of speech signals / J. R. Deller (Jr.), J. H. L. Hansen, J. G. Proakis.John Wiley and Sons. IEEE Press.

6. Nilsson M., Ejnarsson M. Speech recognition using hidden Markov model // 2002. Degree of Master of Science in Electrical Engineering. Blekinge Institute of Technology. Karlskrona: Kazerntryckriet AB, 2002.

7. Aida-Zade K. R. Investigation of combined use of MFCC and LPC features in speech recognition systems // K. R. Aida–Zade, C. Ardil, S. S. Rustamov. World Acad. of Sci., Eng. and Technol. 2006.

8. Rabiner L. R., Sambur M. R. An algorithm for determining the endpoints of isolated utterances // Bell System Tech. J., 1995. 1975. P. 298–315.

9. Atal B., Rabiner L. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition // IEEE Trans. Acoust., Speech, Signal Process. V. 24. P. 201–212, 197.

10. Rabiner L. R. Digital processing of speech signals / L. R. Rabiner, R. W. Schafer. Englewood Cliffs: Prentice Hall, 1978. P. 666–667.

11. Рабинер Л. Р. Цифровая обработка речевых сигналов / Л. Р. Рабинер, Р. В. Шафер. М.: Радио и связь, 1981.

Калимолдаев Максат Нурадилович — д-р физ.-мат. наук, проф., директор Института проблем информатики и управления Министерства образования и науки Республики Казахстан; тел.: 8-727-272-3712;

Мамырбаев Оркен Жумажанович — докторант PhD Казахского национального технического университета им. К. И. Сатпаева; e-mail: morkenj@mail.ru;

Мусабаев Рустам Рафикович — канд. техн. наук, ст. науч. сотр. Института проблем информатики и управления Министерства образования и науки Республики Казахстан; тел.: 8-727-272-3712;

Тусупова Белла Борисовна — канд. техн. наук, доц. Казахского национального технического университета им. К. И. Сатпаева; тел.: 8-777-226-3362

Дата поступления — 20.12.12