УДК 621.395

МЕТОДЫ И АЛГОРИТМЫ ОБНАРУЖЕНИЯ ПАУЗ В РЕЧИ

Волченков В.А., *старший преподаватель кафедры телекоммуникаций и основ радиотехники ФГБОУ ВО «Рязанского государственного радиотехнического университета им. В.Ф. Уткина», e-mail: volchenkov.rzn@yandex.ru.*

VOICE ACTIVITY DETECTION METHODS AND ALGORITHMS

Volchenkov V.A.

Problems of accuracy increase in voice activity detection (VAD) are considered. General information about some VAD methods is given. A new method of voice activity detection is offered. Performance comparison of VADs is given.

Key words: voice activity detection, problems of accuracy, performance comparison.

Ключевые слова: обнаружение пауз, детектор активности речи, методы и алгоритмы, сравнительный анализ.

Рассмотрены вопросы увеличения точности обнаружения пауз. Приведена общая информация о некоторых методах детектирования активности речи. Представлен новый способ обнаружения пауз в речи. Приведено сравнение их производительности.

Введение

Реализация существующих алгоритмов обнаружения пауз базируется на предположении, что речь является нестационарным сигналом, форма ее спектра обычно изменяется через короткие отрезки времени – 10-30 мс. Также считают, что фоновый шум обычно стационарен на более длинном отрезке времени, немного изменяясь со временем, а уровень речевого сигнала обычно выше уровня фонового шума [1]. Речь обычно делят на отрезки длительностью 16-32 мс, и анализируют уровень энергии сигнала на каждом интервале, а также количество переходов сигнала через ноль. В том случае, когда временной интервал определяется обнаружителем как пауза, перед окончательным принятием решения, что сигнал отсутствует, системе необходимо последовательно продетектировать ещё несколько фреймов (в системе GSM 5-6). Таким образом, существующие на сегодняшний день способы определения активности речи позволяют выявить паузы, длительность которых значительно превышает 40 мс. Обнаружение коротких пауз и установление более точных границ для длинных пауз становятся важными задачами, решению которых и посвящена данная работа.

В настоящей работе предложен детектор активности речи, обеспечивающий существенное повышение вероятности правильного разделения речевых сигналов на периоды активной речи и паузы.

Сравнение эффективности алгоритмов VAD

Проведем сравнение эффективности предлагаемого способа с алгоритмом VAD кодера G.729B, а также со способом детектирования пауз на основе отношения правдоподобия (VAD LR – Likelihood-Ratio-Based VAD).

Алгоритм VAD кодера G.729B

Кодер G.729B делит речь на интервалы по 10 мс и вырабатывает решение о наличии или отсутствии речи для каждого фрейма, оценивая при этом четыре параметра [2, 3, 4]:

– разность энергий всего диапазона – $\Delta E_f = \overline{E_f} - E_f$,

– разность энергий диапазона НЧ – $\Delta E_l = \overline{E_l} - E_l$,

– искажение спектра – $\Delta LSF = \sum_{i=o}^{9} (\overline{LSF_i} - LSF_i)^2$,

– разность частоты переходов через ноль – $\Delta ZC = \overline{ZC} - ZC$, где E_f – энергия всего диапазона, E_l – энергия диапазона НЧ, LSF_i – i-я частота спектра сигнала и ZC – частота переходов через ноль входного сигнала, $\overline{E_f}$, $\overline{E_l}$, $\overline{LSF_i}$, $\overline{ZC}$ – параметры, характеризующие шум и обновляемые посредством анализа фонового шума.

Блок схема алгоритма VAD кодера G.729B представлена на рис. 1 [4]. Входные параметры для анализа VAD могут быть получены из входного сигнала или из промежуточных значений речевого кодера. Затем рассчитываются параметры разницы между параметрами входного сигнала и шума ΔE_f, ΔE_l, ΔLSF и ΔZC. Решение о наличии речи получают путем анализа интервалов речевого сигнала по четырем параметрам, которые поступают на схему анализа предыдущих решений. Блок обновления параметров шума основан на схеме авторегрессии первого порядка. Они обновляются, если разница энергии всего диапазона меньше заданного фиксированного порога.

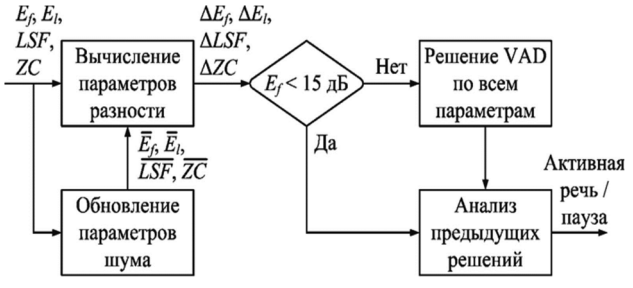

Рис. 1. Алгоритм VAD кодера G.729B

Алгоритм VAD на основе отношения правдоподобия

В [5] был предложен метод, который в отличие от традиционных методов VAD базируется на статистической модели [2], и, как описано, позволяет значительно повысить точность обнаружения [5]. Причиной высокой производительности является учет критериев подавления шума, предложенных Ефремом и Малом [6] для критериев принятия решения о речевой активности.

Решение о речевой активности может быть рассмотрено, как проверка гипотез: H_0 и H_1, которые означают отсутствие и присутствие речи, соответственно. Предполагая, что каждая спектральная составляющая речи и шума имеет комплексное гауссово распределение [6], в котором шум является аддитивным и не коррелирован с речью, функции плотности условного распределения вероятностей (conditional probability density functions – PDF) спектральной составляющей шума Y_k, данных $H_{0,k}$ и $H_{1,k}$ запишем следующим образом:

$$p(Y_k \mid H_{0,k}) = \frac{1}{\pi\lambda_{N,k}} \exp\left\{-\frac{|Y_k|^2}{\lambda_{N,k}}\right\}, \quad (1)$$

$$p(Y_k \mid H_{1,k}) = \frac{1}{\pi(\lambda_{N,k}+\lambda_{X,k})} \exp\left\{-\frac{|Y_k|^2}{\lambda_{N,k}+\lambda_{X,k}}\right\}, \quad (2)$$

где k означает индекс элемента спектральной выборки, а $\lambda_{N,k}$ и $\lambda_{X,k}$ обозначают дисперсию спектров шума и речи соответственно.

Отношение правдоподобия (likelihood ratio – LR) k – го элемента спектральной выборки Λ_k определяется из упомянутых выше двух функций плотности условного распределения вероятностей [5]:

$$\Lambda_k = \frac{p(Y_k \mid H_{1,k})}{p(Y_k \mid H_{0,k})} = \frac{1}{1+\xi_k} \exp\left\{\frac{(1+\gamma_k)\xi_k}{1+\xi_k}\right\}, \quad (3)$$

где γ_k и ξ_k являются апостериорным и априорным ОСШ, определяемые как $\gamma_k = |Y_k|^2 / \lambda_{N,k} - 1$ и $\xi_k = \lambda_{X,k} / \lambda_{N,k}$. Заметим, что определение апостериорного ОСШ немного отличается от оригинального, $\gamma_k = |Y_k|^2 / \lambda_{N,k}$ [7]. Предполагается, что дисперсия шума известна в результате адаптации шума. Тем не менее, дисперсия речи является неизвестной, таким образом, априорное ОСШ n-го фрейма $\xi_k^{(n)}$ оценивают, используя метод прямого принятия решения (decision-directed (DD) method) [6]:

$$\hat{\xi}_k^{(n)} = \alpha \frac{\left|\hat{X}_k^{(n-1)}\right|^2}{\lambda_{N,k}^{(n-1)}} + (1-\alpha) MAX\left\{\gamma_k^{(n)}, 0\right\}, \quad (4)$$

где α – взвешивающий элемент, например, 0,98, и амплитуда спектра незашумленной речи $|\hat{X}_k|$ оценивается с использованием минимальной среднеквадратической ошибки оценочной функции амплитуды логарифмического спектра [7]. Решение о наличии речевой активности представляется средним геометрическим значений Λ_k для всех спектральных выборок:

$$\Lambda = \exp\left\{\frac{1}{K}\sum_{k=1}^{K} \log \Lambda_k\right\}, \quad (5)$$

где K обозначает количество спектральных выборок.

Апостериорное ОСШ γ_k сильно колеблется от фрейма к фрейму из-за большой флуктуации амплитуды спектра на кратковременном интервале $|Y_k|$. С другой стороны, априорное ОСШ $\hat{\xi}_k$ меняется медленно вследствие сглаживающего эффекта. Если значение α увеличивается, $\hat{\xi}_k$ становится более сглаженной. Изменения γ_k и $\hat{\xi}_k$ уравновешивают друг друга при вычислении Λ_k и, следовательно, в результате увеличивается производительность VAD. Потому оценка DD для априорного ОСШ полезна не только для избегания явления музыкального шума при усилении речи [8], но также для уменьшения количества ошибок в обнаружении речевой активности.

Предлагаемый способ детектирования активности речи

В настоящей работе предложен детектор активности речи, базирующийся на использовании вспомогательного сигнала специальной частоты [10].

Структурная схема детектора активности речи изображена на рис. 2.

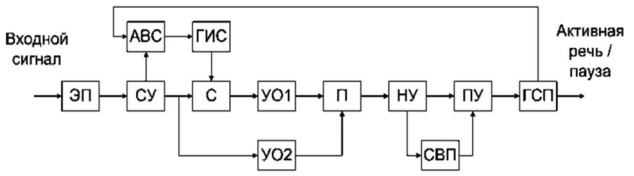

Рис. 2. Структурная схема детектора активности речи

В состав устройства входят: электроакустический преобразователь (ЭП), селективный усилитель (СУ), анализатор входного сигнала (АВС), генератор измерительного сигнала (ГИС), сумматор (С), амплитудные усилители-ограничители (УО) 1 и 2, перемножитель (П), накопитель-усреднитель (НУ), схема вычисления порога (СВП) и пороговое устройство (ПУ) и генератор сигнала паузы (ГСП).

Речевой сигнал с выхода электроакустического преобразователя усиливается селективным усилителем и подается на анализатор входного сигнала и вход сумматора. На второй вход сумматора подается сигнал с выхода генератора измерительного сигнала. Значение необходимой мощности измерительного сигнала вычисляется в анализаторе входного сигнала путем анализа первых 100 мс, т.к. на этом интервале речь обычно отсутствует, и пересчитывается заново при принятии системой решений о наличии паузы в течение 500 мс подряд. Суммарный сигнал с выхода сумматора поступает на вход усилителя-ограничителя 1, где происходит усиление сигнала, а затем ограничение по амплитуде. Аналогичная операция проводится над сигналом, поступающим с выхода селективного усилителя на вход усилителя-ограничителя 2. Сигнал с выхода усилителя-ограничителя 1 подается на первый вход перемножителя. На второй вход перемножителя подается сигнал с выхода усилителя-ограничителя 2. Сигнал с выхода перемножителя поступает на вход накопителя-усреднителя, где происходит выделение сигнала, по амплитуде которого принимают решение о наличии периода активного речевого сигнала или паузы в пороговом устройстве. Значение порога вычисляется в схеме вычисления порога пу-

тем анализа интервала длительностью 50 мс после изменения мощности вспомогательного сигнала. Сигнал с выхода порогового устройства поступает на вход генератора сигнала паузы, который при наличии паузы генерирует сигнал паузы длительностью 10 мс.

Сравнение эффективности предлагаемого способа с методами обнаружения пауз

Тестовым сигналом была речь длительностью 108 секунд, состоящая из фраз, надиктованных разными дикторами, смешанная с транспортным шумом с ОСШ: 5, 10, 15, 20 и 25 дБ. Интервалы активной речи и паузы были отмечены вручную. Пропорции между неактивными и активными участками речи были 0,46 и 0,54, соответственно. Результаты оценки ошибок определения речи и пауз для приведенных выше методов представлены в табл. 1-3.

Таблица 1. Оценка ошибок определения речи и пауз алгоритмом VAD кодера G.729B

ОСШ, дБ	Ошибка определения речи, %	Ошибка определения пауз, %
5	19,16	1,71
10	12,58	3,97
15	5,34	5,89
20	4,57	7,71
25	2,62	10,27

Таблица 2. Оценка ошибок определения речи и пауз способом детектирования пауз на основе отношения правдоподобия

ОСШ, дБ	Ошибка определения речи, %	Ошибка определения пауз, %
5	9,73	20,73
10	3,13	18,52
15	1,67	15,47
20	0,89	10,29
25	0,42	9,58

Таблица 3. Оценка ошибок определения речи и пауз разрабатываемым способом детектирования пауз

ОСШ, дБ	Ошибка определения речи, %	Ошибка определения пауз, %
5	21,47	2,41
10	9,71	2,53
15	5,01	3,09
20	2,03	4,14
25	0,29	6,86

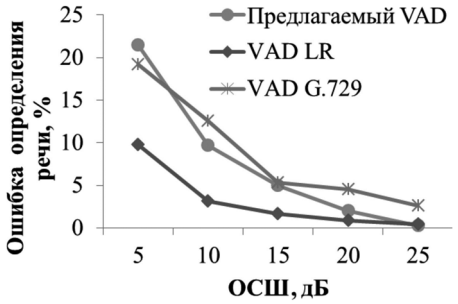

Рис. 3. Сравнение производительности методов обнаружения пауз в речи при воздействии на речевой сигнал транспортного шума. Ошибка определения речи по отношению к ОСШ

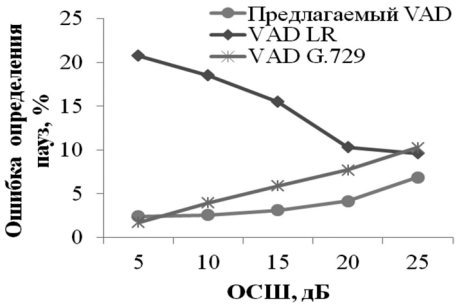

Рис. 4. Сравнение производительности методов обнаружения пауз в речи при воздействии на речевой сигнал транспортного шума. Ошибка определения пауз по отношению к ОСШ

Заключение

Предложенный способ показал себя лучше алгоритма VAD кодера G.729B почти для всех значений ОСШ. Только при ОСШ равном 5 дБ алгоритм VAD кодера G.729B справился лучше. Способ детектирования пауз на основе отношения правдоподобия показал себя лучше остальных приведенных способов с точки зрения параметра «ошибка определения речи», но с точки зрения параметра «ошибка определения пауз» оказался значительно хуже остальных. Самую низкую ошибку определения пауз почти для всех значений ОСШ показал способ детектирования пауз, предложенный в данной статье. В дальнейшей работе предполагается улучшить разрабатываемый алгоритм для уменьшения ошибки определения речи.

Литература

1. Шелухин О.И. Цифровая обработка и передача речи / О.И. Шелухин, В.Г. Лукьянцев; Под ред. О.И. Шелухина. – М.: Радио и связь, 2000. – 456 с.: ил.

2. Kondoz A.M. Digital Speech. Coding for Low Bit Rate Communication Systems. – John Wiley & Sons, Ltd. 2004. – 442 p.

3. ITU-T (1996) A silence compression scheme for G.729 optimised for terminals conforming to ITU-T V.70, ITU-T Rec. G.729 Annex B.

4. ITU-T (1996) Coding of speech at 8 kbit/s using conjugate-structure algebraiccode excited linear prediction (CS-ACELP), ITU-T Rec. G.729.

5. Sohn J., Kim N.S., and Sung W. (1999) 'A statistical model-based voice activity detection', in IEEE Signal Processing Letters, 6(1):1–3.

6. Y. Ephraim and D. Malah (1984) 'Speech enhan-cement using a minimum mean square error short-time spectral amplitude estimator', in IEEE Trans. on Acoust., Speech and Signal Processing, 32(6):1109-20.

7. Y. Ephraim and D. Malah (1985) 'Speech enhancement using a minimum mean square error log-spectral amplitude estimator', in IEEE Trans. on Acoust., Speech and Signal Processing, 33(2):443-5.

8. O. Capp´e (1994) 'Elimination of musical noise phenomenon with the Ephraim and Malah noise suppression', in IEEE Trans. Speech and Audio Processing, 2(2):345-9.

9. Пат. 2436173 Российская Федерация, МПК G10L 15/00, G10L 11/02, Способ обнаружения пауз в речевых сигналах и устройство его реализующее / Витязев В.В., Розов В.И., Волченков В.А.; заявитель и патентообладатель Рязанский государственный радиотехнический университет. – № 2010124342/08, заяв. 15.06.10; опубл. 10.12.11, Бюл. 34.