УДК 621.391.037.372

АЛГОРИТМ ОЦЕНКИ ОТНОШЕНИЯ СИГНАЛ/ШУМ РЕЧЕВЫХ СИГНАЛОВ

М.Б. Столбов

Предложен алгоритм оценки интегрального значения отношения сигнал/шум и его значений в частотных полосах для определения качества фонограмм в системе верификации дикторов. Особенность разработанного алгоритма состоит в робастности по отношению к большой вариативности условий записи и качества фонограмм, а также возможности выполнения оценки в режиме реального времени, т.е. в темпе поступления речевого сигнала. В основу алгоритма положены новые способы оценки спектра шума и детектирования речи. Эксперименты показали достаточную для практических применений достоверность оценок отношения сигнал/шум в диапазоне от 6 до 26 дБ на записях длительностью от 10 с и более.

Ключевые слова: отношение сигнал/шум, детектирование речевых кадров, оценка спектра шума.

Введение

Предварительная оценка качества речевого материала является важной в задаче идентификации личностей по голосу (идентификации диктора). Одним из основных показателей, влияющих на качество идентификации диктора, является отношение уровня исходного речевого сигнала к уровню присутствующего в фонограмме шума. Данное отношение может меняться на протяжении фонограммы вследствие вариаций уровня фонового шума и параметров речи диктора. В качестве интегральной меры качества фонограммы целесообразно использовать среднее по фонограмме отношение сигнал/шум (ОСШ).

Средняя величина ОСШ может использоваться для оценки тестовых и обучающих данных, идентификации канала записи фонограммы, выбора рабочего диапазона частот и т.д. Для решения данных задач могут быть использованы оценки ОСШ в частотных полосах или интегральное по частотам значение ОСШ. Мерой качества речевого сигнала в фонограмме является средняя по всем речевым фрагментам фонограммы оценка ОСШ – так называемое сегментное ОСШ.

Закрепленные в стандартах алгоритмы оценки ОСШ предполагают, что известны как шум, так и полезный сигнал [1, 2]. Однако эти алгоритмы не могут быть использованы в ситуации, когда имеется единственная фонограмма с зашумленным речевым сигналом. В этом случае необходимо решить задачу слепой оценки ОСШ. Предложено несколько групп методов слепой оценки ОСШ (см., например, список литературы в работе [3]):

– распознавание участков речевого сигнала и шума с применением детектора речи, по которым вычисляются оценки спектра шума и речи;
– оценка гистограммы амплитуд огибающих спектра, из которой отдельно определяются распределения речи и шума;

– оценка текущих спектров шума (например, на основе отслеживания минимумов огибающих спектра сигнала);
– оценка статистических параметров распределений спектральных амплитуд (например, статистик высокого порядка).

Последние два направления получили к настоящему моменту наибольшее распространение. Однако задача слепой оценки ОСШ по-прежнему далека от своего решения [4]. Непосредственное использование алгоритмов сторонних разработчиков представляется затруднительным по ряду причин:
– разнообразие требований к алгоритмам в зависимости от области их практического применения;
– использование в алгоритмах параметров, зависящих от типа сигнала (частота дискретизации и пр.), которые в конечном итоге подбираются эмпирически;
– отсутствие общей методики сравнения различных алгоритмов (базы данных, критерии оценки и пр.).

Перечисленные обстоятельства обусловили разработку собственного алгоритма оценки ОСШ, удовлетворяющего следующим требованиям:
– робастность оценки ОСШ для шумов различных типов (в том числе нестационарных);
– работоспособность в интервале значений ОСШ (динамическом диапазоне) фонограмм от 6 до 24 дБ;
– возможность оценки ОСШ на коротких фонограммах (на общей длительности речи не менее 10 с);
– возможность вычисления оценки ОСШ в режиме реального времени;
– устойчивость работы алгоритма к различным типам помех;
– инвариантность к нормировке и частоте дискретизации речевого сигнала;
– низкие вычислительные затраты.

Цель работы – описание практической реализации алгоритма слепой оценки ОСШ, пригодного для работы в широком диапазоне условий (по типам шумов, каналам записи, изменчивости акустической обстановки и пр.).

Алгоритм оценки ОСШ

В основу алгоритма оценки ОСШ положен метод, основанный на вычислении оценок текущего спектра шума. Пусть обрабатываемый сигнал $x(i)$ представляет собой сумму речевого сигнала $s(i)$ и шума $n(i)$:

$$x(i) = s(i) + n(i).$$

При этом принимается, что речевой сигнал и шум статистически независимы. Во многих практических случаях данное условие выполняется. Тогда теоретические кратковременные спектры мощности можно записать как

$$Px(k, m) = Ps(k, m) + Pn(k, m),$$

где k и m – индексы частоты и кадра соответственно; $Px(k, m)$, $Ps(k, m)$, $Pn(k, m)$ – спектры мощности зашумленного сигнала, речи и шума соответственно. Тогда ОСШ на кадре данных в частотных полосах запишется как:

$$SNR(k, m) = Ps(k, m)/Pn(k, m) = (Px(k, m) - Pn(k, m))/Pn(k, m) = Px(k, m)/Pn(k, m) - 1.$$

Интегральное по частоте значение ОСШ на кадре данных выражается следующей формулой:

$$SNR(m) = Ps(m)/Pn(m),$$

где $Ps(m)$, $Pn(m)$ – мощности речи и шума на кадре m.

Интегральной характеристикой качества фонограммы в целом является среднее по всему файлу отношение сигнал/шум. Более представительной характеристикой качества речевого сигнала является сегментное ОСШ, вычисляемое как среднее значение покадровых оценок ОСШ на речевых сегментах сигнала:

$$SSNR(k) \text{ [дБ]} = 10 \log_{10} \langle SNR(k, m) \rangle_{SP},$$

$$SSNR \text{ [дБ]} = 10 \log_{10} \langle SNR(m) \rangle_{SP},$$

где $\langle \rangle_{SP}$ обозначает усреднение по кадрам речи.

На практике теоретические значения спектров мощности сигнала $Px(k, m)$ и шума $Pn(k, m)$ неизвестны, поэтому используются их оценки. Оценка ОСШ в частотных полосах выражается следующим образом:

$$SNR(k, m) = \max\{ \delta, |Y(k, m)|^2/| \breve{N}(k,m) |^2 - 1\},$$

где $\breve{N}(k,m)^2$ – оценки спектральной плотности мощности шума (СПМ) на кадре с индексом m; δ – минимальное значение оценки ОСШ; $|Y(k, m)|^2$ – сглаженная по времени оценка СПМ сигнала на кадре с индексом m:

$$|Y(k, m)| = \alpha\, |Y(k, m-1)| + (1 - \alpha)\, |X(k, m)|,$$

где $X(k, m)$ – кратковременный спектр сигнала $x(i)$; α – коэффициент забывания (от 0,75 до 0,8).

Таким образом, центральными элементами оценки ОСШ являются алгоритм оценки текущего спектра шума и детектор речевого сигнала. Рассмотрим вкратце предложенные нами алгоритмы.

В основе базового алгоритма оценки спектра шума лежит итеративный алгоритм оценки амплитудного спектра, построенный на идее управляемого порога, вычисляемого по отношению спектральных амплитуд зашумленного сигнала и шума [5]. Недостаток базового алгоритма состоит в том, что при появлении резких всплесков энергии сигнала он останавливает подстройку спектра шума. На практике такие ситуации (например, увеличение коэффициента усиления записывающего устройства и др.) могут привести к полной остановке обновления оценки спектра шума.

Вместо управляемого порога предлагается применить управляемые коэффициенты сглаживания. В этом случае оценка амплитуды спектра шума осуществляется рекурсивно, кадр за кадром, без поиска пауз речи, с использованием экспоненциального сглаживания с коэффициентами $\beta(k, m)$, управляемыми в каждой спектральной полосе индивидуально:

$$\breve{N}(k,m) = \beta(k, m)\, \breve{N}(k,m-1) + (1 - \beta(k, m))\, |Y(k, m)|.$$

Коэффициенты сглаживания меняются в зависимости от отношений $\breve{N}(k,m-1)/|Y(k, m)|$. Управление коэффициентами учитывает тот факт, что основная часть спектральных амплитуд $|Y(k, m)|$ на шуме распределена в интервале $0{,}5N(k, m)$–$2N(k, m)$. При выходе амплитуд $|Y(k, m)|$ за верхнюю границу этого интервала величина коэффициента $\beta(k, m)$ уменьшается, а в случае выхода за нижнюю границу увеличивается. Алгоритм является эффективным в вычислительном отношении и позволяет получать оценки ОСШ в режиме реального времени.

Предложенный алгоритм оценки спектра шума продемонстрировал работоспособность на шумах различных типов (стационарных и нестационарных) и в условиях различных видов помех. В качестве критерия точности алгоритма оценки ОСШ использована интегральная по частоте относительная ошибка [6],

$$\mathrm{LogErr}(m) = < 20\log_{10} |N(k, m)|/ \breve{N}(k,m) >,$$

где < > обозначает усреднение по кадрам сигнала.

На рис. 1 приведен пример зависимости относительной ошибки оценки ОСШ от времени для предложенного алгоритма для речи, зашумленной реальным шумом кондиционера с отношениями сигнал/шум 34 дБ (кривая 1), 16 дБ (кривая 2), 4 дБ (кривая 3), и шума без речи (кривая 4).

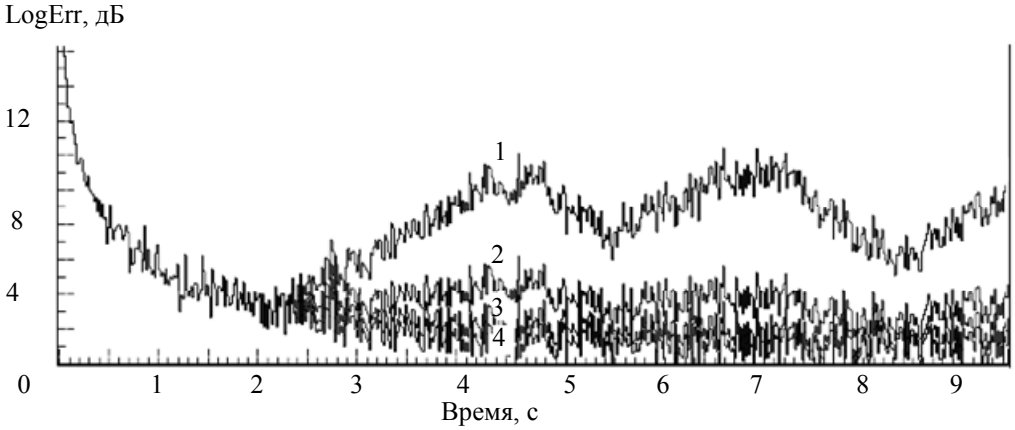

Рис. 1. Зависимость логарифмических ошибок от времени для речевого сигнала с разными значениями ОСШ

Из рис. 1 видно, что на начальном участке настройки оценки спектра шума ошибка оценки уменьшается. На участке, содержащем речь (начиная со 2-й секунды), ошибки оценки возрастают по мере увеличения величины ОСШ сигнала.

Сопоставление уровней ошибки оценки ОСШ для предложенного алгоритма и других известных алгоритмов [6] показывает, что предложенный алгоритм дает меньшие ошибки, вплоть до ОСШ, равного +28 дБ. Аналогичные результаты были получены при исследовании зависимости ошибки от используемой полосы частот.

Оценка интегрального по частоте значения ОСШ на кадре рассчитывается как отношение мощностей речевого сигнала и шума в полосе частот 300–3300 Гц:

$$SNR(m) = \max\left\{\delta, \frac{\sum_{k=Kb}^{k=Ke} |Y(k,m)|^2}{\sum_{k=Kb}^{k=Ke} \breve{N}(k,m)^2} - 1\right\},$$

где Kb, Ke – значения индексов частоты, соответствующие 300 Гц и 3300 Гц; δ – минимальное значение оценки ОСШ.

Вторым важным элементом алгоритма оценки сегментного ОСШ является детектирование кадров с речью. Для детектирования речевых кадров была применена интегральная мера

$$T(m) = \frac{1}{(Ke - Kb)} \sum_{k=Kb}^{k=Ke} \frac{|Y(k,m)|^2}{\bar{N}(k,m)^2},$$

где Kb, Ke – значения индексов частоты, соответствующие 300 Гц и 3300 Гц.

В случаях, когда значение $T(m)$ на кадре больше 1,3, кадр классифицируется как речевой. Отличие предложенного детектора от традиционных энергетических детекторов заключается в его устойчивости к наиболее распространенным тональным помехам. Действительно, в случае большого значения отдельной тональной компоненты спектра отношение $Y(k, m)/\bar{N}(k,m)$ будет близким к единице и не внесет значительного вклада в величину $T(m)$. Величины всех порогов и сглаживающих констант были определены экспериментальным путем.

Эксперименты

Целью экспериментов было исследование диапазона работоспособности реализованного алгоритма оценки ОСШ. Алгоритм исследовался на фонограммах с известными значениями ОСШ.

В качестве иллюстрации на рис. 2 показаны зависимости оценок ОСШ сигнала в частотных полосах для зашумленной речи при ОСШ, равном 4 дБ, 10 дБ, 16 дБ, 22 дБ, 28 дБ.

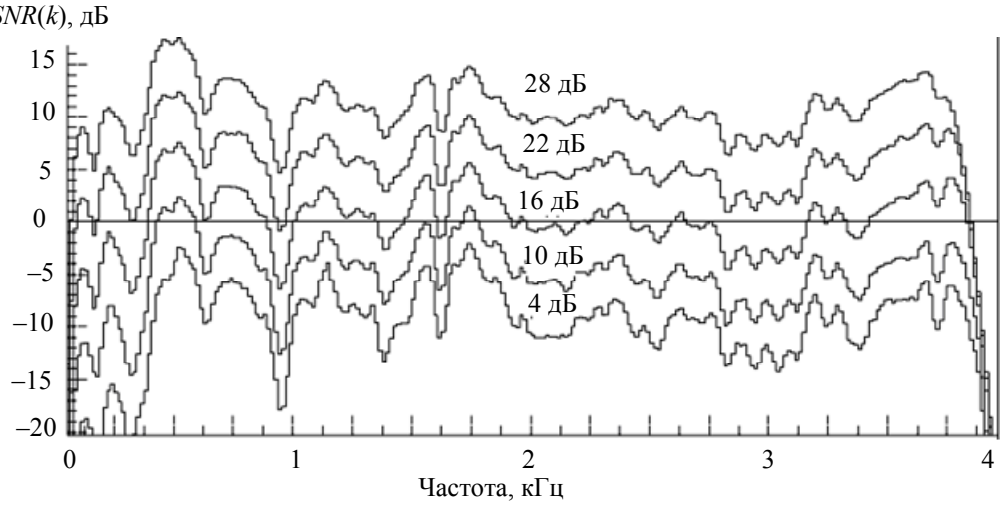

Рис. 2. Оценки $SSNR(k)$ [дБ] для зашумленной речи при разных значениях ОСШ сигнала

Графики, приведенные на рис. 2, демонстрируют изменение оценки сегментного ОСШ, соответствующего величине ОСШ сигнала.

Проведенные эксперименты выявили следующие характеристики алгоритма:

- по мере уменьшения величины ОСШ количество распознанных речевых кадров уменьшается, поскольку в качестве речевых сегментов распознаются лишь кадры с относительно большими локальными значениями ОСШ. Вследствие этого для сигналов с малыми значениями ОСШ алгоритм показал завышенные оценки *SSNR*.
- для сигналов с большими значениями ОСШ (больше 30 дБ) алгоритм показал заниженные оценки *SSNR*, что связано с тем, что алгоритм оценки спектра шума дает завышенные оценки спектра шума на участках речевого сигнала с большими величинами ОСШ.

Для расширения рабочего диапазона алгоритма оценки значения *SSNR* корректировались путем следующего нелинейного преобразования:

$SSNR_c$ [дБ] = 1,4 ($SSNR$ [дБ] – 1).

Пример скорректированной оценки интегрального значения ОСШ для различных типов шумов и значений ОСШ приведен в таблице.

Экспериментально определенные значения времени настройки алгоритма показали, что оценки ОСШ приближались к устойчивым значениям на участке речи длительностью 3 с и становились достоверными на участке речи длительностью около 10 с. Время настройки алгоритма в значительной мере определяется параметром скорости адаптации в алгоритме оценки спектра шума. В данном варианте постоянная времени адаптации была задана равной 1 с. Поскольку для некоторых применений (например, текстозависимой верификации диктора) предъявляется требование к длительности сигнала не более 5 с, задача сокращения времени настройки остается актуальной.

Эксперименты с реальными зашумленными фонограммами показали хорошее соответствие оценок ОСШ ожидаемым значениям.

	$SSNR_c$							
	–2 дБ	4 дБ	10 дБ	16 дБ	22 дБ	28 дБ	34 дБ	40 дБ
Шум 1	–1,7	2,9	8,9	15,6	22,5	29,4	36,6	38,7
Шум 2	–1,35	3,11	9,11	15,9	22,9	30,2	37,3	38,6
Шум 3	–1,16	3,52	9,86	16,5	23,5	30,5	37,6	39,0
Шум 4	0,03	5,26	11,87	18,8	26,0	33,1	38,4	39,7
Среднее по шумам	–1,05	3,95	9,93	16,7	23,7	30,1	37,5	39,0

Таблица. Оценки $SSNR_c$ для различных уровней и типов шумов и уровней ОСШ, дБ

Применение алгоритма в системе верификации диктора

Пример графического экрана для процедуры оценки ОСШ представлен на рис. 3.

Работа алгоритма в системе происходит в два этапа. Сначала на кадрах речи вычисляется матрица ОСШ, $SNR(k, m)$. По ней в качестве выходных параметров вычисляются средние по всей фонограмме оценки ОСШ в частотных полосах и интегральное значение ОСШ, по которым принимается решение о качестве фонограммы. Затем, в случае пригодности фонограммы, производится более детальный анализ, и отбрасываются «плохие» кадры с малым значением ОСШ.

Основным результатом работы является практическое внедрение разработанного алгоритма в систему идентификации диктора. К настоящему моменту алгоритм прошел практическую проверку на больших объемах данных.

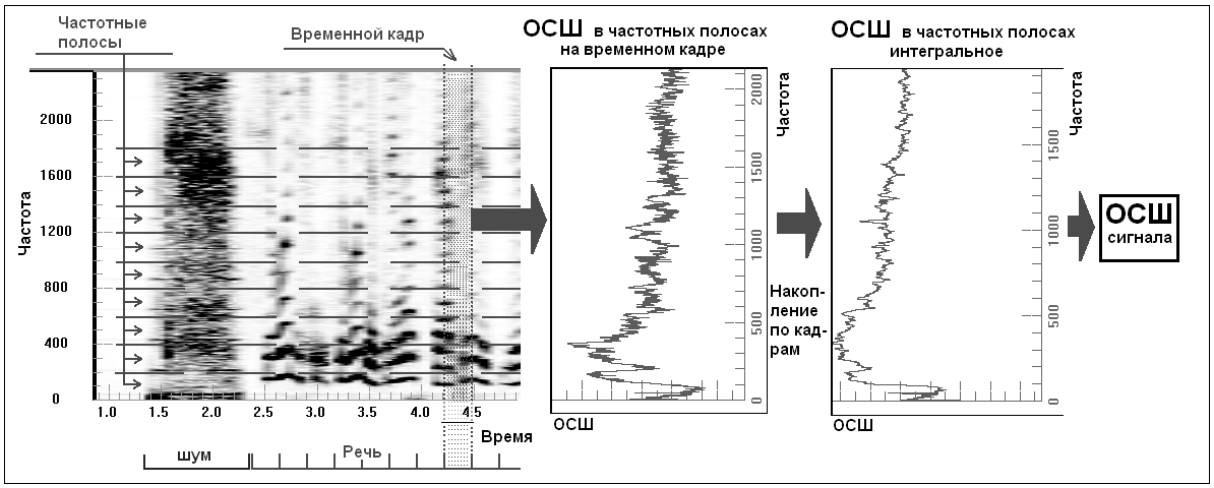

Рис. 3. Иллюстрация процедуры оценки ОСШ в частотных полосах

Заключение

Предложен алгоритм автоматической оценки интегральных и спектральных (в частотных полосах) значений ОСШ на фонограммах с зашумленной речью, использующий оценки текущих значений амплитудного спектра шума.

Основными компонентами реализованной схемы оценки ОСШ являются рекурсивный алгоритм оценки амплитудного спектра шума и детектор речи. Алгоритм оценки амплитудного спектра шума не требует наличия пауз речи в сигнале и устойчив к различным помехам. Детектор речи устойчив к присутствию в сигнале мощных тональных помех.

Предложенный алгоритм продемонстрировал свою работоспособность как на тестовых, так и на реальных записях речи. В настоящее время алгоритм используется в ряде продуктов ООО «ЦРТ».

Разработанный алгоритм оценки ОСШ удовлетворяет предъявленным требованиям, в частности: достоверная оценка ОСШ в интервале от 6 до 24 дБ на фонограммах, содержащих речь длительностью от 10 с.

Основными задачами дальнейшей работы являются обеспечение достоверности оценок при значениях ОСШ менее +6 дБ и на длительностях речи не менее 10 с.

Литература

1. ITU-T Rec. P. 56. Objective measurement of active speech level. – 1993. – Approved in Dec. 2011. – Printed in Switzerland, Geneva, 2012. – 17 p.
2. ITU-T G. 160. Objective measures for the characterization of the basic functioning of noise reduction algorithms. – 2008. – Approved in Nov. 2009. – Printed in Switzerland, Geneva, 2010. – 14 p.

3. Kim C., Stern R.M. Robust Signal-to-Noise Ratio Estimation Based on Waveform Amplitude Distribution Analysis // Proc. INTERSPEECH-2008. – Brisbane, Australia, 2008. – P. 2598–2601.
4. Nemer E., Goubran R., Mahmoud S. SNR Estimation of Speech signals Using Subbands and Fourth-Order Statistics // IEEE Signal Processing Letters. – 1999. – V. 6. – № 7. – P. 171–174.
5. Hirsch H.G., Ehrlicher C. Noise estimation techniques for robust speech recognition // Proc. ICASSP. – Detroit, Michigan, 1995. – V. 1. – P. 153–156.
6. Hergolz C., Jeub M., Nelke C., Beaugeant C., Vary P. Evaluation of Single- and Dual-channel Noise Power Spectral Density Estimation Algorithms for Mobil Phones // Proc. Konferenz Elektronische Sprachsignalverarbeitung (EESV). – Aachen, Germany, 2011. – P. 1–10.

Столбов Михаил Борисович – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, ООО «ЦРТ-инновации», кандидат технических наук, ст. научный сотрудник, доцент, stolbov@speechpro.com