УДК 621.391.037.372

С. В. Алейник, К. К. Симончик

АЛГОРИТМЫ ВЫДЕЛЕНИЯ ТИПОВЫХ ПОМЕХ И ИСКАЖЕНИЙ В РЕЧЕВЫХ СИГНАЛАХ

Исследованы способы выделения типовых аддитивных помех в системах обработки речевых сигналов. Проведена экспериментальная оценка влияния того или иного детектора помех на эффективность системы верификации диктора. Предложены усовершенствованные алгоритмы выделения помех.

Ключевые слова: *шум, акустические помехи, импульсные помехи, обработка речевых сигналов.*

Введение. Акустические речевые сигналы зачастую искажены аддитивными помехами, значительно снижающими эффективность систем верификации диктора. В общем случае данные аддитивные помехи могут быть разделены на две большие группы: стационарные, присутствующие на всем протяжении сигнала (например, широко известный белый и розовый шум), и нестационарные кратковременные, присутствующие на отдельных участках сигнала.

При наличии помех второй группы входные сигналы редко бывают полностью искажены. Незначительно искаженные участки сигнала чередуются с участками, сильно искаженными импульсными помехами различных типов: клиппированием, кратковременными электрическими наводками, перегрузками и т.п. Именно эти нестационарные помехи и искажения оказывают наибольшее отрицательное влияние. Соответственно используя детекторы, способные на этапе предобработки с высокой вероятностью обнаруживать подобного рода помехи и искажения (с целью их дальнейшего подавления или исключения из анализа), можно существенно улучшить качество систем обработки речи. Основными типовыми помехами и искажениями, рассматриваемыми в настоящей статье, являются щелчки, перегрузки, короткие тональные сигналы, клиппирование.

Следует также отметить, что важными дополнительными требованиями к таким детекторам являются высокая скорость и низкая ресурсоемкость, т.е. типовые требования, предъявляемые к устройствам предобработки.

Щелчки. Несмотря на кажущуюся простоту, обнаружение щелчков представляет собой определенные трудности, поскольку короткие импульсы, воспринимаемые человеком на слух как „щелчки“, могут в общем случае существенно различаться как во временном, так и в частотном представлении (рис. 1, *1* — короткий „классический“ высокочастотный щелчок; *2* — низкочастотный щелчок; *3* — щелчок с короткими осцилляциями; *4* — „длинный“ щелчок с шумовым или осциллирующим заполнением).

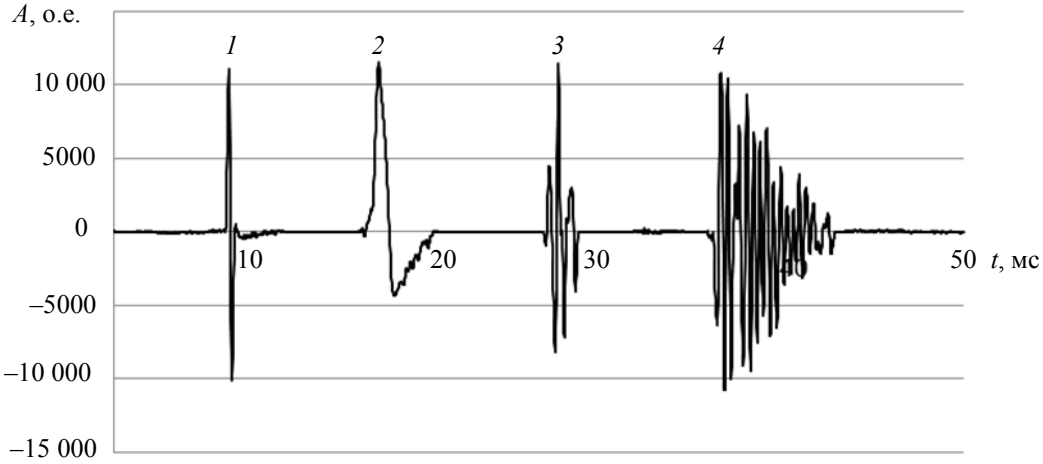

Рис. 1

Например, короткий высокочастотный щелчок хорошо обнаруживается следующим способом. Анализируемый сигнал $x(i)$, где i — дискретный временной индекс, вначале пропускается через высокочастотный (ВЧ) фильтр с частотой среза порядка 2—4 кГц. Затем вычисляется первая разность $d(i) = y(i) - y(i - 1)$, где $y(i)$ — сигнал на выходе фильтра, далее ее абсолютная величина сравнивается с пороговым значением. К сожалению, данный способ не работает на низкочастотных (НЧ) щелчках (кривая *2*), так как, во-первых, основная часть их энергии сосредоточена в низкочастотной области и „срезается“ ВЧ-фильтром, а во-вторых, значение $d(i)$ щелчков данного вида и речевых сигналов различается несущественно.

Результаты исследований различных алгоритмов, основанных на методах линейного предсказания и авторегрессионных моделях [1, 2] показали их высокую вычислительную сложность, поэтому авторы разработали более простой алгоритм обнаружения щелчков различных типов (рис. 2, сплошная кривая — участок анализируемого сигнала со щелчком, пунктир — выходная величина алгоритма (умноженная на 1000 с целью отображения на одном графике с сигналом); t_0—t_3 — временные метки границ окна анализа).

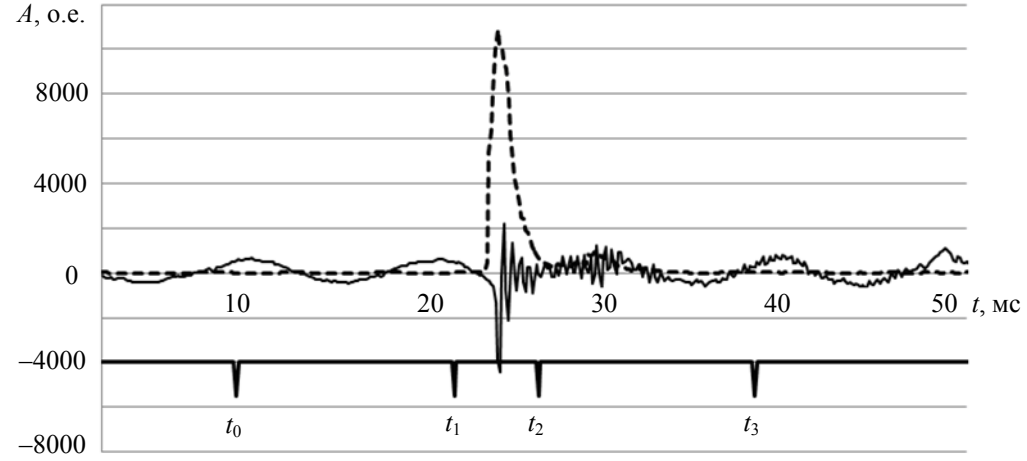

Рис. 2

Разработанный алгоритм включает следующие шаги.

1. Выбирается длина окна анализа (t_0, t_3) таким образом, чтобы выполнялось условие

t_3–$t_0 = KL_c$, где L_c — предполагаемая длительность щелчка и K — масштабный коэффициент, изменяющийся в диапазоне от 10 до 100.

2. Окно разбивается на три части (см. рис. 2), причем длина центральной части выбирается соизмеримой с предполагаемой длиной щелчка, и t_1—$t_0 = t_3$—t_2.

3. Выходная величина V_c, сравниваемая в дальнейшем с пороговым значением, рассчитывается как:

$$V_c(t_{\text{center}}) = \frac{2(t_1 - t_0)}{t_2 - t_1} \frac{\sum_{t=t_1}^{t_2} x^2(t)}{\sum_{t=t_0}^{t_1} x^2(t) + \sum_{t=t_2}^{t_3} x^2(t)}, \quad (1)$$

где $x(t)$ — анализируемый сигнал; $t_{\text{center}} = 0{,}5(t_0+t_3)$ — центр интервала $[t_3, t_0]$.

Нетрудно понять, что V_c в (1) есть отношение мощностей сигнала на различных участках, нормированное таким образом, что в случае стационарного сигнала (например, белого шума) $V_c = 1$. Для речевых сигналов полученные значения V_c колебались от нуля до нескольких единиц. Величина $V_c > 8$ сигнализирует о наличии щелчка (строго говоря, конкретное пороговое значение зависит от выбранной допустимой вероятности ложной тревоги и размеров окна анализа и определяется экспериментально).

Очевидно, что длина интервала t_2–t_1 в идеальном случае должна соответствовать длительности щелчка, подлежащего обнаружению, что в реальных условиях труднодостижимо. В проведенных экспериментах установлено, что если это значение находится в пределах нескольких длин щелчка, то результаты детектора также вполне приемлемы. В противном случае, при значительной априорной неопределенности в длительности предполагаемых щелчков, приходится осуществлять перебор.

Путем моделирования были получены следующие временные параметры детектора: интервал t_2–t_1 5 мс; t_1–t_0 и t_3–t_2 — 60 мс. При таких значениях получены хорошие результаты по детектированию типовых щелчков на реальных речевых сигналах.

Следует заметить, что при обнаружении коротких высокочастотных щелчков бывает полезна предварительная фильтрация ВЧ-фильтром с частотой среза 2—4 кГц.

Перегрузки. Перегрузкой называются короткие (1—2 отсчета) скачки сигнала, импульсы или серии подобных импульсов большой амплитуды, вызванные изменением знака сигнала при так называемом „целочисленном переполнении“. Причины перегрузок кроются в следующем. На практике наиболее широко используемый тип квантования при переводе аудиосигналов в цифровую форму — 16-битовое квантование. При таком типе квантования каждый отсчет сигнала представляет собой целое двухбайтовое число в формате “signed short int” (стандарт ANSI), т.е. амплитуда отсчета изменяется от –32 768 до 32 767. В то же время обработка сигнала может выполняться, например, в форматах “long”, “float” или “double”. При этом если число, получившееся после обработки, выходит за пределы интервала [–32 768, 32 767], то при его простом преобразовании к типу “signed short int” (при записи, например, на диск в WAV-формате) произойдет „переброс знака“, и число, например 32 768, преобразуется в –32 768, число –32 769 — в 32 767 и т.д.

Общие выражения для результата могут быть записаны как:

$$\text{if } (x > 32\,767) \text{ then } y = (x \bmod 32\,767) - 32\,768,$$

$$\text{if } (x < -32\,768) \text{ then } y = -(|x| \bmod 32\,768) + 32\,768,$$

где x — число до преобразования, y — результат преобразования, mod — операция вычисления по модулю.

На слух одиночная перегрузка воспринимается как высокочастотный щелчок, а серия подобных щелчков — как резкий громкий треск, существенно ухудшающий как разборчивость речевого сигнала, так и показатели систем обработки речи.

На рис. 3 приведен типичный пример перегрузки, возникшей при преобразовании величины в формате double (время перегрузки 6,68 мс, значение x = 56 981) в двухбайтовый формат signed short int.

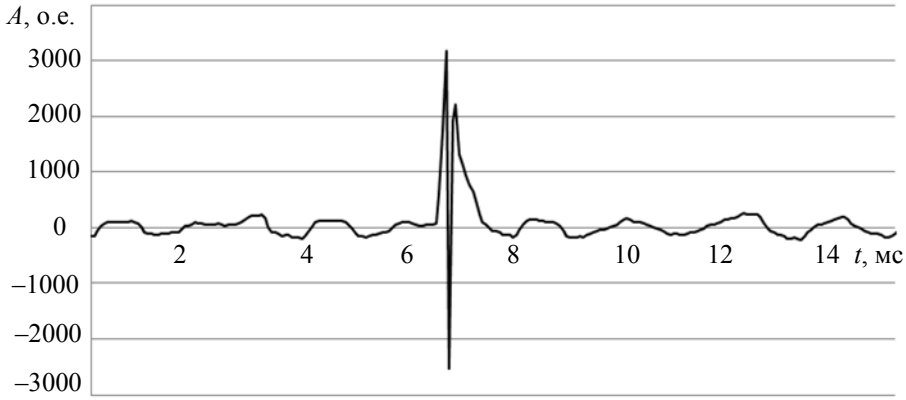

Рис. 3

Одиночная перегрузка (в отличие от серии) с успехом может быть обнаружена с помощью детектора ВЧ-щелчков. Однако, используя первую разность (которая была ранее описана как неэффективная при обнаружении НЧ-щелчков второго типа), возможно создать алгоритм, обнаруживающий как одиночные, так и множественные перегрузки. Дело в том, что „переброс“ знака вызывает сильные резкие скачки амплитуды за один отсчет, часто соизмеримые с динамическим диапазоном сигнала. В данном случае коэффициент вычисляется следующим образом:

$$d(i) = \frac{|x(i) - x(i-1)|}{A_{\max} - A_{\min}}, \qquad (2)$$

где $A_{\max}$ и $A_{\min}$ — максимальное и минимальное значения амплитуды сигнала, вычисленные по всей выборке. Теоретически $0 \le d(i) \le 1$, однако на чистой речи, без перегрузок, величина $d(i)$, как правило, значительно меньше единицы.

Наши эксперименты по определению плотности распределения коэффициента $d(i)$ на большом наборе речевых сигналов показали, что при пороге $T_d = 0,7$ и принятии решения о наличии перегрузки по условию $d(i) > T_d$ вероятность ошибки первого рода (вероятность принять речь за перегрузку) равна приблизительно 10^{-8} на один отсчет сигнала, что дает хорошие результаты даже на длинных сигналах.

Алгоритм детектирования перегрузок представлен ниже.

1. Выбирается величина порога T_d, например, 0,7.

2. По всей выборке сигнала вычисляются его максимальное $A_{\max}$ и минимальное $A_{\min}$ значения.

3. Для каждого отсчета сигнала $x(i)$, $i = 1, N-1$ (здесь N — полная длина сигнала) по формуле (2) вычисляется коэффициент $d(i)$.

Производится сравнение $d(i)$ с выбранным ранее порогом, и в случае $d(i) > T_d$ принимается решение о наличии перегрузки.

Короткие тональные сигналы — это широко известные сигналы телефонного вызова, представляющие собой обычно одну или две гармоники длиной около одной секунды. Отличительной особенностью таких сигналов является высокий уровень и стабильность частоты составляющих гармоник. Соответственно в подавляющем большинстве алгоритмов обнаружения тонов используется анализ спектров мощности (или модулей спектров мощности)

сигналов [3, 4]. Отметим, что тональные сигналы без примеси постороннего шума или в сумме с шумом малой мощности могут быть также с успехом обнаружены детектором клиппированных сигналов, базирующемся на анализе гистограммы [5].

Нами были исследованы два алгоритма обнаружения коротких тонов: на основе подсчета локальных максимумов в спектре и детектор оценки постоянства амплитуды спектральных максимумов. Детектор на основе подсчета локальных максимумов использует тот факт, что при наличии в сигнале тональной компоненты большой амплитуды спектр мощности такого сигнала имеет ярко выраженный узкий пик.

Алгоритм детектирования следующий.

1. Выбирается величина M — длина сегмента сигнала для вычисления спектра мощности.

2. Для каждого сегмента сигнала длиной M вычисляется модуль мгновенного спектра мощности $S(m)$, где $m = 0, M/2$ — дискретная частота.

3. Для всех $m = 0, M/2$ находится спектральный максимум $S_{\max}$.

4. Вычисляется пороговый уровень $T_s = T_{s0} S_{\max}$.

5. Для всех $m = 0, M/2$ подсчитывается целевая величина K_s — количество спектральных отсчетов, превышающих уровень T_s, т.е.: $K_s = \sum_{m=0}^{M/2} k_s$, где

$$k_s = \begin{cases} 1 \;\; if \;\; S(m) \ge T_s, \\ 0 \;\; if \;\; S(m) < T_s. \end{cases}$$

6. Производится сравнение: если $K_s \le 3$, то принимается решение о наличии тональной составляющей в исследуемом фрагменте сигнала.

В алгоритме оценки постоянства амплитуды спектральных максимумов используется тот факт, что на соседних сегментах сигнала амплитуда тональной составляющей изменяется незначительно. В данном алгоритме сравниваются максимумы модулей спектров мощности двух соседних сегментов сигнала $S^j_{\max}$ и $S^{j+1}_{\max}$ (где j — индекс сегмента) и вычисляется их относительная разность:

$$D_s = \frac{\left| S^{j+1}_{\max} - S^j_{\max} \right|}{S^j_{\max}}.$$

Сравнение величины D_s с заранее выбранным порогом T_d дает искомый результат: если $D_s < T_d$, то принимается решение о наличии тональной составляющей в *j*-м фрагменте сигнала.

Пороговые величины T_{s0} и T_d были определены нами в ходе моделирования: $T_{s0} = 0,01$ и $T_d = 0,001$.

Клиппирование — искажение формы сигнала, происходящее при перегрузке усилителя и при выходе выходного напряжения усилителя из его динамического диапазона. На осциллограмме клиппирование обычно выглядит как ограничение сигнала по амплитуде.

На слух клиппирование воспринимается как появление излишней звонкости, „металлического“ звучания и может существенно снижать качество обработки речи.

Алгоритм детектирования клиппирования на основе анализа гистограммы сигнала приведен в работе [5].

Экспериментальная оценка эффективности разработанных алгоритмов. Эффективность предложенных алгоритмов была оценена в ходе экспериментов на примере системы верификации диктора на основе *i*-векторов, описанной в работе [6].

Для тестирования алгоритмов выделения типовых помех использовались записи телефонных разговоров в стандартном GSM-канале: 610 фонограмм различной длительности. Тестовые фонограммы поступали на вход блока предобработки, содержащего параллельно соединенные детекторы: участки фонограмм, на которых срабатывал хотя бы один из включенных детекторов, исключались из дальнейшего анализа. Показателем качества системы был выбран равновероятный уровень ошибок первого и второго рода (Equal Error Rate, EER), широко применяемый для оценки эффективности биометрических систем. Результаты экспериментов представлены в таблице.

Алгоритм детектирования				EER, %
щелчков	перегрузок	клиппирования	тональных помех	
–	–	–	–	13,6
–	–	–	+	10,4
–	–	+	–	10,4
–	+	–	–	10,91
+	–	–	–	10,85

Из таблицы видно, что при отсутствии детекторов (первая строка) качество системы наихудшее (высокий EER). Включение какого-либо детектора приводит к уменьшению EER, т.е. к повышению качества верификации. Следует отметить одинаковое улучшение при работе детекторов клиппирования и тональных помех. По мнению авторов, данный эффект был вызван тем, что, во-первых, в тестовых фонограммах клиппирование практически отсутствовало (в отличие от тональных сигналов телефонных вызовов). И, во-вторых, как уже отмечалось ранее, детектор клиппирования с успехом обнаруживает тональные сигналы, состоящие из одной гармоники.

Заключение. В статье рассмотрены алгоритмы обнаружения типовых помех, наиболее часто встречающихся при обработке речевых сигналов. Указаны характеристики данных алгоритмов, полученные путем моделирования на реальных записях речи. С помощью экспериментального исследования показано, что обнаружение и исключение из анализа речевых сигналов участков с помехами или искажениями способно повысить качество систем верификации диктора.

Работа проводилась при финансовой поддержке Министерства образования и науки Российской Федерации.

СПИСОК ЛИТЕРАТУРЫ

1. *Esquef P. A. A., Karjalainen M., Välimäki V.* Detection of clicks in audio signals using warped linear prediction // Proc. of the 14th Intern. Conf. on Digital Signal Processing. Greece, 2002. Vol. 2. P. 1085—1088.

2. *Esquef P. A. A., Biscainho L. W. P., Diniz P. S. R., Freeland F. P.* A double-threshold-based approach to impulsive noise detection in audio signals // Proc. EUSIPCO. Finland, 2000. Vol. 4. P. 2041—2044.

3. *So H. C., Chan Y. T., Ma Q., Ching P. C.* Comparison of Various Periodograms for Sinusoid Detection and Frequency Estimation // IEEE Trans. on Aerospace and Electronic Systems. 1999. Vol. 35. P. 945—952.

4. *Grigorakis A.* Application of Detection Theory to the Measurement of the Minimum Detectable Signal for a Sinusoid in Gaussian Noise Displayed on a Lofargram. Research Report, Aeronautical and Maritime Research Laboratory, Melbourne, Australia, 1997.

5. *Алейник С. В., Матвеев Ю. Н., Раев А. Н.* Метод оценки уровня клиппирования речевого сигнала // Науч.-техн. вестн. информационных технологий, механики и оптики. 2012. № 3 (79). С. 79—83.

6. *Белых И. Н., Капустин А. В., Козлов А. В., Лоханова А. И., Матвеев Ю. Н., Пеховский Т. С., Симончик К. К., Шулипа А. К.* Система идентификации дикторов по голосу для конкурса NIST SRE 2010 // Информатика и ее применения. 2012. Т. 6, № 1. С. 91—98.

Сведения об авторах

Сергей Владимирович Алейник — ООО „ЦРТ-инновации“, Санкт-Петербург; научный сотрудник; E-mail: aleinik@speechpro.com

Константин Константинович Симончик — канд. техн. наук; ООО „ЦРТ“, отдел верификации и идентификации диктора, Санкт-Петербург; руководитель отдела; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; доцент; E-mail: simonchik@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.