Информационные процессы, Том 15, № 3, 2015, стр. 314–323

МАТЕМАТИЧЕСКИЕ МОДЕЛИ, ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

Улучшение качества речи с использованием адаптивных спектральных оценок

Ю. Сандовал-Ибарра[*], В. Диаз-Рамирез[*], В.И. Кобер[**], В.Н. Карнаухов[**]

[*] *Национальный политехнический институт, Центр исследований и развития цифровых технологий, Тихуана, 22510, Мексика*

[**] *Институт проблем передачи информации, Российская академия наук, Москва, 127051, Россия*

Поступила в редколлегию 08.06.2015

Аннотация—Известные методы статистического оценивания речевых сигналов основаны на некоторых предположениях о статистических свойствах речевых и шумовых процессов. В реальных приложениях эти предположения не всегда верны из-за нестационарной природы реальной окружающей среды. В данной работе мы предлагаем синтезировать новые робастные функции спектральных оценок речевых сигналов с использованием порядковых статистик. Предлагаемые оценки хорошо адаптируются к нестационарным характеристикам речевых сигналов и фонового шума в реальных условиях. С помощью компьютерного моделирования мы показываем, что предлагаемые методы улучшения речи превосходят традиционные методы с точки зрения объективных критериев качества.

КЛЮЧЕВЫЕ СЛОВА: фильтрация речевых сигналов, улучшение качества речи, робастные оценки.

ВВЕДЕНИЕ

Современные методы улучшения качества речи основаны на оптимизации метрик (таких как качество и разборчивость речи [1]), которые хорошо описывают субъективное восприятие человеком восстановленных речевых сигналов. Телекоммуникационные системы являются примерами приложений для успешной работы которых требуются надежные алгоритмы, обеспечивающие высокую разборчивость речи с уменьшенным содержанием мешающего шума. Улучшение качества речи обычно формулируется как проблема оценки модуля спектра неискаженного речевого сигнала из наблюдаемого сигнала. Известно несколько успешных оценок речевых сигналов, среди которых можно выделить следующие оценки: максимального правдоподобия (ML) [1]–[3], минимальной среднеквадратической ошибки (MMSE) [4]–[6], логарифмической среднеквадратической ошибки (log-MMSE) [7] и максимальной апостериорной вероятности (MAP) [8]. Эти оценки основаны на нескольких допущениях относительно статистических свойств речевых сигналов и шума. Например, обычно предполагаются асимптотический характер статистических характеристик спектральных свойств речевых сигналов, а также известное распределение речевого сигнала. В реальных приложениях локальные функция плотности распределения речевых сигналов и шума могут со временем меняться. Следовательно, существующие оценки могут приводить к неудовлетворительным результатам при обработке реальных речевых сигналов на неоднородном фоне окружающей среды. Таким образом, задача разработки локально адаптивных робастных оценок для улучшения качества речи является актуальной.

В цифровой обработке сигналов широко применяются фильтры, основанные на вычислении порядковых статистик [9]–[12]. Эти фильтры являются робастными к шуму с “тяжелыми” хвостами в плотности функции распределения и хорошо сохраняют в процессе обработки мелкие

детали и резкие перепады сигнала. При улучшении качества речи эти свойства нелинейных фильтров являются полезными для подавления нежелательного шума с одновременным сохранением разборчивости речи. Недавно была предложена локально адаптивная нелинейная фильтрация для обработки речи [12], которая способна уменьшать аддитивный шум, сохранять разборчивость речи почти без искусственных артефактов, таких как "музыкальный" шум. Однако полученная оценка не учитывает метрик, описывающих субъективное восприятие человеком речевых сигналов [13],[14]. В данной работе мы предлагаем использовать порядковые статистики для улучшения существующих оценок обработки речи с учетом субъективного восприятия человеком речевых сигналов.

Пусть дискретная функция $f(n) = s(n) + d(n)$ – входной речевой фрагмент длиной N отсчетов, функция $s(n)$ – неискаженный сигнал речи, а $d(n)$ – аддитивный шум с нулевым средним значением. В частотной области наблюдаемый сигнал может быть представлен как

$$F_k e^{j\theta_k^f} = S_k e^{j\theta_k^s} + D_k e^{j\theta_k^d}, \tag{1}$$

где F_k, S_k и D_k – амплитуды дискретных Фурье спектров сигналов $f(n)$, $s(n)$ и $d(n)$, соответственно; θ_k^f, θ_k^s и θ_k^d – фазы спектров наблюдаемого, неискаженного сигналов и шума, соответственно. После получения оценки спектра неискаженного сигнала по наблюдаемому сигналу вычисление восстановленного спектра сигнала производится как $S_k e^{j\theta_k^s} \approx \hat{S}_k e^{j\theta_k^f}$.

Существующие методы оценивания спектра речи предполагают, что коэффициенты Фурье (реальные и мнимые части) речевых сигналов и шума имеют нулевые средние значения, сигналы являются независимыми гауссовыми случайными процессами и являются квазистационарными в интервале 20–40 миллисекунд. В реальных приложениях эти предположения могут быть не верны, например, когда речевой сигнал искажен негауссовым и нестационарным шумом. В данной работе мы предлагаем метод построения робастных оценок на основе вычисления порядковых статистик для улучшения качества речи. Предлагаемые оценки хорошо адаптируются к нестационарным особенностям зашумленных речевых сигналов. Они также способны улучшить качество речи, сохраняя разборчивость речи и не вводя в сигнал искусственных артефактов.

Статья организована следующим образом. В разделе 1 кратко описывается локально-адаптивная обработка речи с использованием порядковых статистик. В разделе 2 описывается предлагаемый алгоритм для улучшения качества речи. В разделе 3 представлены экспериментальные результаты, полученные с помощью предлагаемого подхода. Эти результаты сравниваются с результатами, полученными с помощью современных известных методов с точки зрения объективных показателей. Наконец, в последнем разделе приведены наши выводы.

1. УЛУЧШЕНИЕ КАЧЕСТВА РЕЧИ С ИСПОЛЬЗОВАНИЕМ ЛОКАЛЬНОЙ ОБРАБОТКИ СИГНАЛОВ

Синтез фильтров на основе порядковых статистик обычно осуществляется в два этапа: выделяются однородные окрестности (структурный подход) и строятся оценки неискажённого сигнала (оценивание) [10]–[11]. Локально-адаптивная обработка сигнала выполняется в скользящем окне. На первом шаге выделяются однородные окрестности в скользящем окне – желательные структуры сигнала в окне. Затем, на основе выделенных элементов окрестностей строится оценка неискаженного сигнала для центрального элемента окна с использованием выбранного критерия. На Рис. 1 показан пример речевого сигнала и построение локальных окрестностей на основе порядковых статистик.

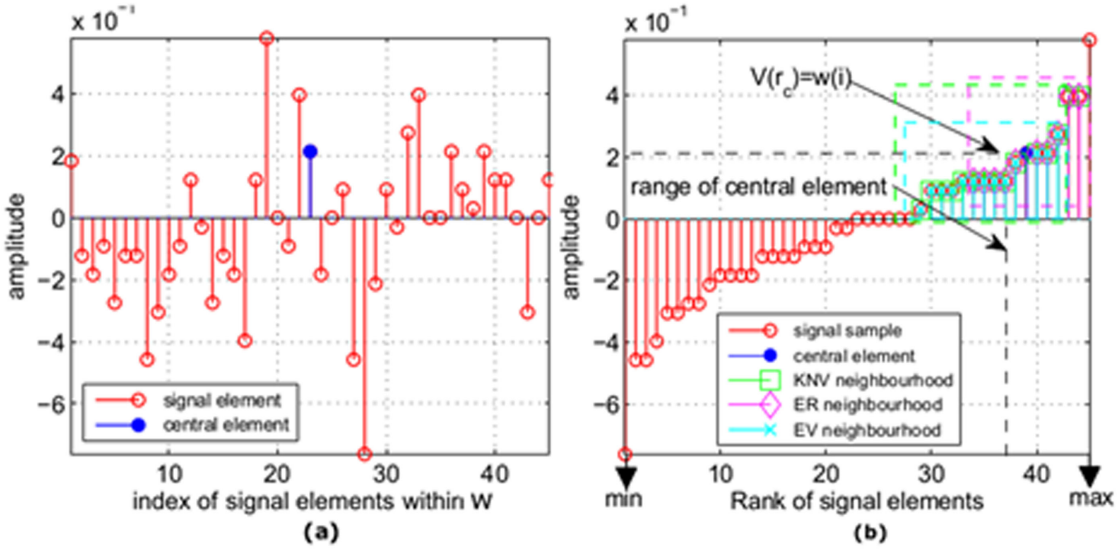

Рис. 1. Вычисление локально адаптивных окрестностей: (a) скользящее окно, (b) вариационный ряд и локальные окрестности.

Для обрабатываемого речевого сигнала вектор скользящего окна $\mathbf{w}$, состоящего из S элементов, можно представить следующим образом:

$$\mathbf{w} = \left[w\left(n - i + \frac{S+1}{2}\right) = f(n) : \ \ |n-i| \le \frac{S-1}{2} \right]^T, \tag{2}$$

где i является индексом центрального элемента в пределах текущего окна, а T обозначает транспонирование. Вариационный ряд $v(r)$ – упорядоченная последовательность элементов вектора $\mathbf{w}$, удовлетворяющая следующему условию: $v(1) \le v(2) \le \cdots \le v(S)$. Величины $v(r)$ и $r(v)$ являются r^{th}-вой порядковой статистикой и рангом величины v, соответственно [11]. Отметим, что порядковая статистика и ранг могут быть вычислены из локальной гистограммы сигнала $\{h(q);\ q = 0, \ldots, Q-1\}$ внутри скользящего окна: $r(v) = \sum_{q=0}^{v} h(q)$, где Q – количество уровней квантования сигнала.

Имеется несколько вариантов построения локальных окрестностей, основанных на порядковых статистиках [10]. Одной из самых популярных окрестностей для цифровой обработки сигналов является EV – соседняя область. Эта окрестность представляет из себя подмножество элементов вектора $\mathbf{w}$, значения которых отклоняются от величины центрального элемента $w(i)$ по меньшей мере на заданные величины $-\varepsilon_v$ и $+\varepsilon_v$ как указано ниже:

$$\mathbf{v} = \{v(n) = w(n) : \ \ w(i) - \varepsilon_v \le w(n) \le w(i) + \varepsilon_v\}, \tag{3}$$

где $\mathbf{v}$ является вектором $S_a \times 1 (S_a \le S)$, чьи элементы формируют подмножество элементов вектора $\mathbf{w}$. При улучшении качества речи при построении оценок используются методы статистического оценивания [3]. Так в работе [4] предложены оптимальные оценки амплитуды спектра неискажённого речевого сигнала с точки зрения среднеквадратической ошибки и логарифмической среднеквадратической ошибки. Эти оценки обычно дают хорошие результаты при подавлении стационарного шума в речевом сигнале. С другой стороны, использование этих оценок ухудшает субъективное качество речевых сигналов, так как восстановленный сигнал

содержит раздражающий “музыкальный” шум, который может вызывать быструю усталость слушателя.

Оценку минимальной среднеквадратической ошибки амплитуды спектра неискажённого сигнала речи [4] можно записать как

$$\hat{S}_k = \frac{\sqrt{\pi}}{2}\frac{\sqrt{\beta_k}}{\gamma_k}\exp\left(-\frac{\beta_k}{2}\right)\left[(1+\beta_k)I_0\left(\frac{\beta_k}{2}\right)+\beta_k I_1\left(\frac{\beta_k}{2}\right)\right]V_k, \tag{4}$$

где I_0 и I_1 представляют собой модифицированные функции Бесселя нулевого и первого порядков, соответственно, а β_k вычисляется как

$$\beta_k = \frac{\xi_k}{1+\xi_k}, \tag{5}$$

где ξ_k и λ_k являются априорным и апостериорным отношениями сигнал/шум (SNR), вычисляемыми по формулам:

$$\gamma_k = \frac{V_k^2}{1+\lambda_k^d} \tag{6}$$

и

$$\xi_k = \frac{\lambda_k^s}{\lambda_k^d}. \tag{7}$$

Здесь λ_k^s и λ_k^d – дисперсии неискаженного сигнала и шума, соответственно. Во временно́й области центральный элемент скользящего окна неискаженного сигнала можно вычислить следующим образом:

$$y(i) - ay_{k-1}(i) + (1-a)y_k(i), \tag{8}$$

где $a \in [0,1]$ – весовой коэффициент, а $y_k(i)$ получается из

$$y_k(i) = \mu_s + \frac{\lambda_k^s}{\lambda_k^s+\lambda_k^d}\left(s(i)-\mu_s\right), \tag{9}$$

где μ_s – среднее значение восстановленного сигнала, а $s(i)$ – центральный элемент скользящего окна после применения оценки минимальной среднеквадратической ошибки.

В данной работе мы предлагаем метод обработки речевых сигналов, модифицируя существующие оценки таким образом, чтобы улучшить качество речи, сохраняя разборчивость речи и не вводя в сигнал искусственных звуков.

2. ПРЕДЛАГАЕМЫЙ АЛГОРИТМ

В этом разделе мы описываем предлагаемый алгоритм для улучшения качества речи при помощи локально-адаптивной обработки сигналов. Блок-схема алгоритма представлена на Рис. 2, а его шаги рассмотрены детально ниже по тексту.

ШАГ 1: Считываем начальный входной сегмент $\mathbf{n}_0$ с S элементами при отсутствии речевого сигнала.

ШАГ 2: Считываем входной речевой сегмент $f(n)$ с N элементами и устанавливаем счетчик $i = 1$.

ШАГ 3: Создаем вектор окна $\mathbf{w}$, вокруг i-го зашумленного элемента, используя выражение (2).

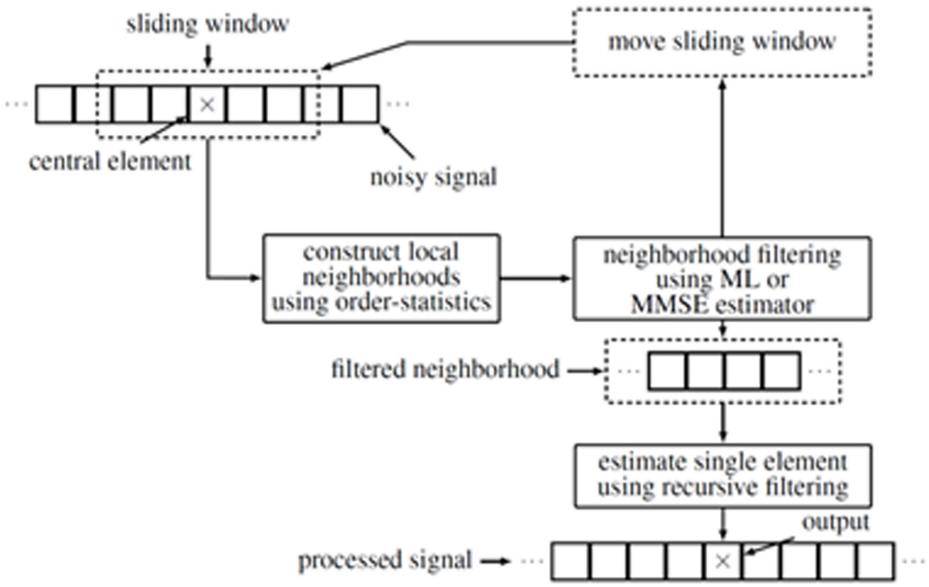

Рис. 2. Блок-схема предлагаемой ранговой фильтрации.

ШАГ 4: Вычисляем величину ε_v следующим образом [11]:

$$\varepsilon_v = \alpha_1 \sigma_f \left[1 - \frac{1}{1 - \Omega^{-a_2}}\right], \tag{10}$$

где Ω – локальное отношение сигнал/шум (SNR), которое вычисляется по формуле $\Omega = \frac{\mathbf{w}^T\mathbf{w}}{\mathbf{n}_0^T\mathbf{n_0}}$, σ_f – среднеквадратичное отклонение шума. Параметры $a_1 \geq 1$ и $a_2 \in (0, 1]$ учитывают априорную информацию о диапазоне речевого сигнала и флуктуации шума.

ШАГ 5: Строим EV – окрестность $\mathbf{v}$ вектора $\mathbf{w}$ с помощью выражений (3) и (10).

ШАГ 6: Применяем оценку минимальной среднеквадратической ошибки, используя выражение (4).

ШАГ 7: Вычисляем оценку выходного сигнала , используя выражения (8) и (9). Устанавливаем $i = i + 1$. Если $i \leq N_i$, то переходим к **ШАГУ 3**, иначе переходим к **ШАГУ 2**.
Результат работы алгоритма – восстановленный сигнал с использованием оптимальной среднеквадратичной оценки неискаженного сигнала и локально-адаптивной обработки на основе вычисления порядковых статистик.

3. ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

В этом разделе мы представляем экспериментальные результаты, полученные с помощью предлагаемого метода. Были проведены многочисленные эксперименты по проверке качества работы предлагаемого подхода. Полученные результаты были сравнены с результатами, полученными с помощью существующих известных алгоритмов улучшения качества речи. Мы протестировали классические алгоритмы максимального правдоподобия, обозначаемый ML [1]–[3], и минимальной среднеквадратической ошибки, обозначаемый MMSE [4], а также предлагаемые алгоритмы, основанные на порядковых статистиках и оценках среднеквадратичной ошибки и максимального правдоподобия, которые обозначаются как rank-MMSE и rank-ML, соответственно.

Все алгоритмы были протестированы на речевых сигналах, искаженных двумя типами шума: белый нормальный шум и автомобильный шум. Отношение сигнал/шум SNR принимало значения 20 дБ, 15 дБ и 10 дБ. Рассматриваемые алгоритмы были протестированы с использованием базы данных института инженеров по электротехнике и электронике [13]. Эта база данных содержит 600 речевых предложений, произносимых дикторами-мужчинами и дикторами-женщинами. Предложения в базе данных фонетически сбалансированы и имеют относительно низкую предсказуемость словарного контекста. Предложения были записаны с частотой выборки 8 кГц. Для оценки качества работы алгоритмов использовались следующие метрики.

– Качество речи характеризуется оценкой качества непрерывной речи (PESQ) [14];
– Разборчивость речи описывается кратковременной объективной разборчивостью (STOI) [15];
– Шумоподавление характеризуется отношением полезный сигнал/шумовая помеха (SIR) [16];
– Введение артефактов в речевой сигнал описывается отношением полезный сигнал/артефакты (SAR) [16].

Протестируем качество работы алгоритмов речевого сигнала, искаженного аддитивным шумом. Параметры для предложенных алгоритмов: $S = 121$, $a_1 = 2.0$, $a_2 = 0.45$, $a = 0.8$. На рисунке Рис. 3 показан пример улучшения качества речи, искаженной аддитивным шумом SNR=15 дБ, с помощью предлагаемых алгоритмов. Дополнительно на Рис. 3 показаны

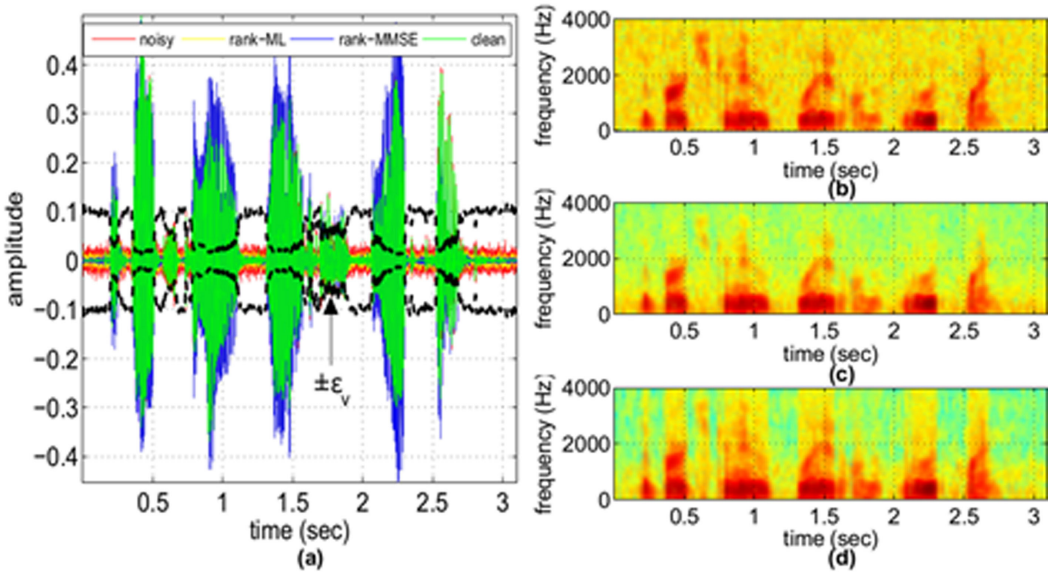

Рис. 3. Пример улучшения качества речи, искаженной аддитивным гауссовым шумом SNR=15 дБ, с помощью предлагаемых алгоритмов: (a) речевые сигналы, (b) спектрограмма зашумленного сигнала, (c) спектрограмма обработанного сигнала алгоритмом rank-ML и (d) спектрограмма обработанного сигнала алгоритмом rank-MMSE.

спектрограммы зашумленного и обработанных сигналов. Можно отметить, как ε_v адаптируется к локальным изменениям зашумленного сигнала: когда локальное отношение сигнал/шум низкое, то ε_v принимает большие значения. Это означает, что в этом случае предлагаемые алгоритмы выполняют более агрессивную фильтрацию, чем при больших отношениях сигнал/шум. На Рис. 4 показаны результаты обработки 600 искаженных речевых сигналов с помощью тестируемых алгоритмов с 95% доверительной вероятностью с точки зрения качества

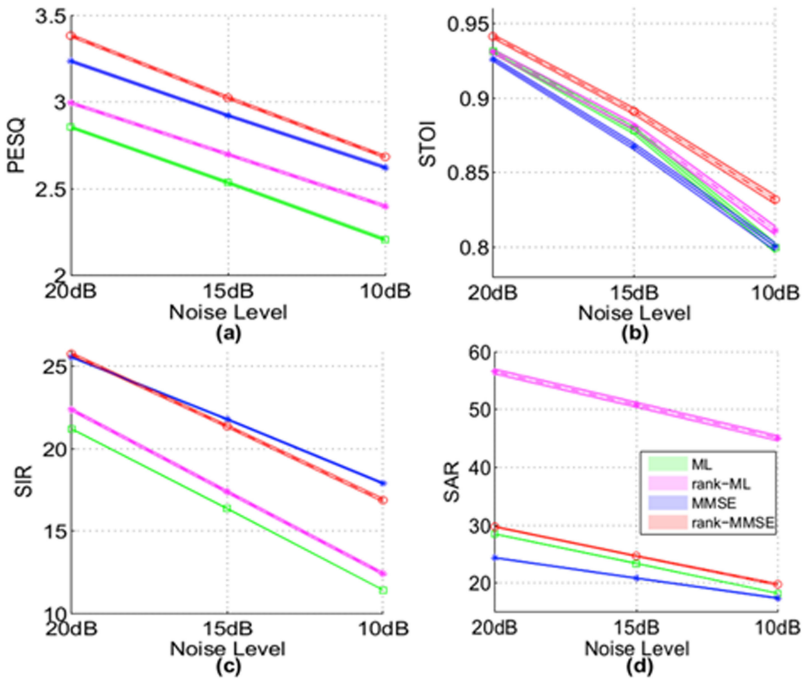

Рис. 4. Результаты обработки речи, искаженной аддитивным гауссовым шумом SNR=15 дБ, с помощью тестируемых алгоритмов с 95% доверительной вероятностью с точки зрения: (a) качества непрерывной речи (PESQ), (b) разборчивости (STOI), (c) шумоподавления (SIR) и (d) введения артефактов (SAR).

речи, разборчивости, шумоподавления и введения артефактов. Отметим, что предложенный алгоритм rank-ML дает лучшие результаты с точки зрения качества непрерывной речи и шумоподавления, чем классический алгоритм ML для всех рассматриваемых значений отношений сигнал/шум. Также можно видеть, что предлагаемый алгоритм rank-ML превосходит все тестируемые алгоритмы по критерию введения артефактов. Предложенный алгоритм rank-MMSE превосходит свой классический аналог по качеству непрерывной речи и по введению артефактов для всех рассматриваемых значений отношения сигнал/шум. Более того, алгоритм rank-MMSE является лучшим среди тестируемых алгоритмов с точки зрения разборчивости для всех значений отношения сигнал/шум. Классический алгоритм MMSE хорошо подавляет шум за счет введения существенного “музыкального” шума (наихудшие значения критерия – введение артефактов).

Теперь оценим качество работы алгоритмов по улучшению речи в окружающей среде с автомобильным шумом. Для этого эксперимента параметры предлагаемых алгоритмов следующие: $S = 65$, $a_1 = 2.0$, $a_2 = 0.5$, $a = 0.8$. На рисунке Рис. 5 показан пример улучшения качества речи, искаженной автомобильным шумом SNR=15 дБ, а также показаны спектрограммы зашумленного и обработанных сигналов. На Рис. 6 показаны результаты обработки 600 искаженных речевых сигналов с помощью тестируемых алгоритмов с 95% доверительной вероятностью с точки зрения качества речи, разборчивости, шумоподавления и введения артефактов. Отметим, что предложенный алгоритм rank-ML дает существенное улучшение речи с точки зрения качества непрерывной речи и шумоподавления, чем классический алгоритм ML для всех рассматриваемых значений отношений сигнал/шум. Отметим, что способность шумоподавления алгоритма rank-ML аналогична способности шумоподавления классического алгоритма MMSE. Обычный алгоритм ML является наихудшим с точки зрения качества непре-

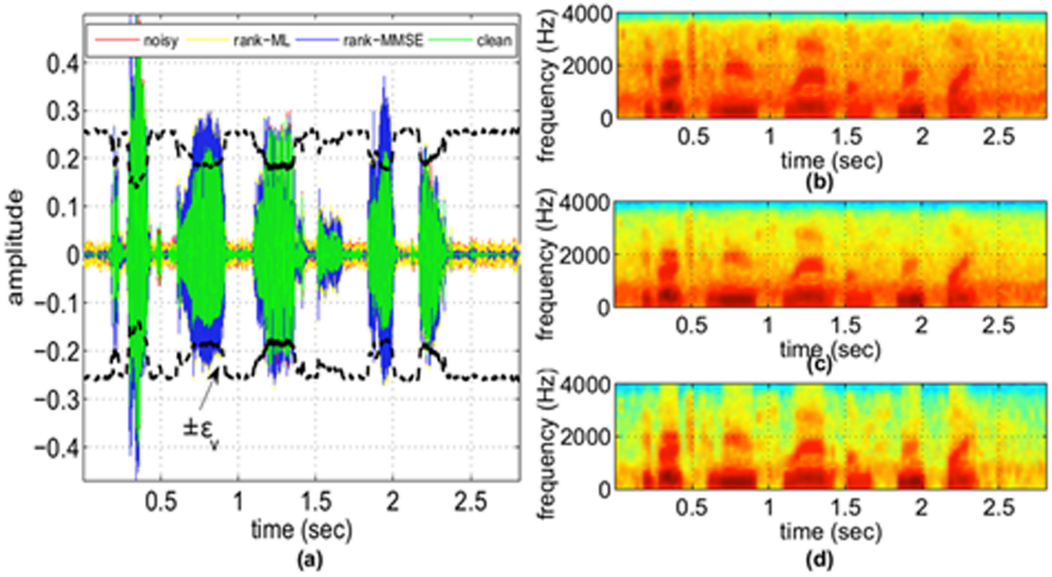

Рис. 5. Пример улучшения качества речи, искаженной автомобильным шумом SNR=15дБ, с помощью предлагаемых алгоритмов: (a) речевые сигналы, (b) спектрограмма зашумленного сигнала, (c) спектрограмма обработанного сигнала алгоритмом rank-ML и (d) спектрограмма обработанного сигнала алгоритмом rank-MMSE.

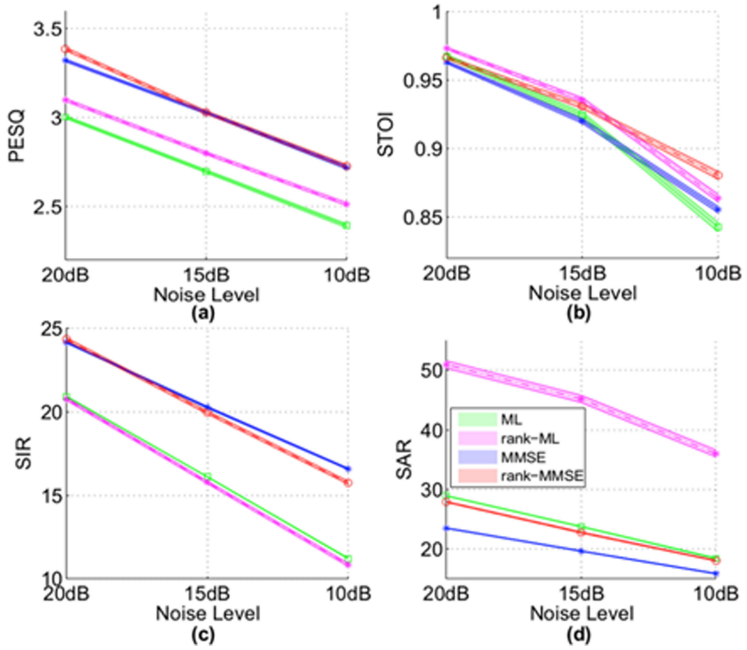

Рис. 6. Результаты обработки речи, искаженной автомобильным шумом SNR=15 дБ, с помощью тестируемых алгоритмов с 95% доверительной вероятностью с точки зрения: (a) качества непрерывной речи (PESQ), (b) разборчивости (STOI), (c) шумоподавления (SIR) и (d) введения артефактов (SAR).

рывной речи и шумоподавления среди всех тестируемых алгоритмов. Предлагаемый алгоритм rank-MMSE превосходит все алгоритмы по качеству непрерывной речи и шумоподавлению для

отношения сигнал/шум 15 дБ и 20 дБ. Обычный алгоритм MMSE дает слегка лучшие характеристик и шумоподавления при сильном зашумлении речевого сигнала 10 дБ. Однако, подобно предыдущим результатам с аддитивным шумом, алгоритм MMSE имеет наихудший результат среди всех протестированных алгоритмов по критерию введения артефактов.

ЗАКЛЮЧЕНИЕ

В данной работе были предложены новые алгоритмы улучшения речевых сигналов на основе известных оценок амплитуды спектра неискаженных сигналов и порядковых статистик. Предлагаемые оценки хорошо адаптируются к нестационарным характеристикам речевых сигналов и фонового шума в реальных условиях. С помощью компьютерного моделирования было показано, что предлагаемые алгоритмы улучшения речи превосходят классические методы с точки зрения объективных критериев качества.

СПИСОК ЛИТЕРАТУРЫ

1. Loizou P. Speech Enhancement: Theory and Practice, Second Edition. *Taylor & Francis*, 2013.
2. McAulay R., Malpass M. Speech enhancement using a softdecision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Processing*, 1980, vol. 28, No. 2, pp. 137-145.
3. Kim N.S. and Chang J.-H. Spectral enhancement based on global soft decision. *Signal Processing Letters, IEEE*, 2000, vol. 7, No. 5, pp. 108–110.
4. Ephraim Y. and Malah D. Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, 1984, vol. 32, No. 6 pp. 1109–1121.
5. Ding G.-H., Huang T. and Xu B. Suppression of additive noise using a power spectral density mmse estimator. *Signal Processing Letters, IEEE*, 2005, vol. 11, No. 6, pp. 585–588.
6. Martin R. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *Speech and Audio Processing, IEEE Transactions on*, 2005, vol. 13, No. 5, pp. 845–856.
7. Ephraim Y., Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, 1985, vol. 33, No. 2, pp. 443–445.
8. Lotter T. and Vary P. Speech enhancement by map spectral amplitude estimation using a super-Gaussian speech model. *EURASIP Journal on Applied Signal Processing*, 2005, pp. 1110–1126.
9. Yaroslavsky L. and Eden M. Fundamentals of Digital Optics. Birkhaeuser Boston, 1996.
10. Kober V., Mozerov M. and Alvarez-Borrego J. Nonlinear filters with spatially connected neighborhoods. *Opt. Eng.*, 2001, vol. 40, No. 6, pp. 971–983.
11. Huber P.J., Pop P. C., Ronchetti E.M. Robust Statistics, Second Ed., Wiley, 2009.
12. Diaz-Ramirez V.M. and Kober V. Robust speech processing using local adaptive nonlinear filtering. *IET Signal Processing*, 2013, vol. 7, No. 5, pp. 345–359.
13. IEEE Subcommittee (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE Trans. Audio and Electroacoustics*, 1969, AU-17(3), pp. 225–246.
14. ITU, Perceptual evaluation of speech quality (PESQ) An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *ITU-T Recommendation P.862*, 2001.
15. Taal C.H., Hendriks R.C., Heusdens R. and Jensen J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2011, vol. 19, No. 7, pp. 2125–2136.
16. Vincent E., Gribonval R. and F‘evotte C. Performance measurement in blind audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions*, 2006, vol. 14, No. 4, pp. 1462–1469.

Speech enhancement with adaptive spectral estimators

Sandoval-Ibarra Y., Diaz-Ramirez V., Kober V. and Karnaukhov V.

Common statistical estimators for speech enhancement rely on several assumptions about statistical properties of speech and noise processes. In real applications these assumptions may not be always satisfied due the effects of a nonstationary environment. We propose new robust spectral estimators for speech enhancement by incorporation of calculation of rank-order statistics to existing speech enhancement estimators. The proposed estimators are better adapted to nonstationary characteristics of speech signals and noise processes in real environments. By means of computer simulations we show that the proposed estimators yield a better performance in terms of objective metrics than that of known estimators.

KEYWORDS: speech filtering, speech enhancement, robust estimators.