К вопросу формирования описаний для распознавания речевых команд

Жигулёвцев Ю.Н., *кандидат технических наук, с.н.с., доцент*

Рассмотрены методы построения описаний для распознавания речи в условиях шумов с применением алгоритмов авторегрессионного анализа и сингулярных разложений, приведены результаты экспериментального исследования качества распознавания при различных уровнях отношения сигнал-шум и использовании различных алгоритмов шумопонижения.

• распознавание речи • авторегрессионный анализ • сингулярное разложение • шумопонижение.

The methods of descriptions creation for speech recognition in noise conditions with application of the autoregressive analysis and singular value decomposition algorithms are considered, speech recognition quality experimental results are given at various levels of the signal-to-noise ratio and various noise reduction algorithms use.

• speech recognition • autoregressive analysis • singular value decomposition • noise reduction.

Введение

При решении задач речевых технологий важную роль играет первичный анализ речевых сигналов. Целью такого анализа является получение описания речевого сигнала в форме последовательности векторов параметров, характеризующих изменяющиеся во времени свойства акустического речевого сигнала в обобщённом (сигнальном) виде, содержащем максимум полезной информации о речевом сообщении и максимально инвариантных к акустической обстановке, для реальных приложений, характеризующихся порой значительным уровнем шумов. Большинство используемых параметрических представлений тем или иным образом связано со спектральными либо корреляционными характеристиками речевых сигналов, и критериями выбора могут служить описательная мощность и сложность реализации метода анализа. Кроме этого, важным преимуществом может являться «обратимость» метода анализа, позволяющая провести синтез («ресинтез») речевого сигнала по его параметрическому описанию, что особенно удобно при построении систем речевого диалогового управления. Названным критериям в значительной степени удовлетворяет метод авторегрессионного анализа, в речевых исследованиях получивший название «линейное предсказание речи».

Теоретические предпосылки метода вытекают из регрессионного анализа [35, 38, 41] и оптимальной фильтрации [23, 25, 40]. Детальная историческая справка по развитию методов спектрального оценивания, в том числе авторегрессионных, приведена в работе Э.А. Робинсона [36]. Два источника метода обусловили два параллельных направления его развития. Одно из них базируется на блочной, а другое на последовательной (рекуррентной) обработке отсчётов речевого сигнала. Результаты развития этих на-

правлений по состоянию на конец 70-х — начало 80-х годов обобщены в монографиях Маркела и Грэя [26] и Ю.Н. Прохорова [34] соответственно. В книге [27 Марпл] с общих позиций спектрального оценивания приведены подробные сведения об обоих подходах, однако без анализа их применений для обработки речи.

В настоящее время линейная авторегрессионная модель речеобразования является наиболее распространённой формой представления математической модели речевого тракта для анализа и синтеза речевых сигналов в задачах речевых технологий. Это обусловлено адекватностью этой модели представлению акустической модели речевого тракта в виде отрезков труб, обеспечивающей приемлемое для большого числа практических задач качество. Кроме того, за более чем 50-летний период развития подхода достаточно детально исследованы теоретические и прикладные аспекты метода, найдены эффективные алгоритмические решения, вычислительные процедуры и аппаратурные реализации.

С одной стороны, авторегрессионный анализ даёт метод оценки параметров сигнала, которые могут применяться для решения задач распознавания и синтеза речевых сигналов, рассматриваемых как реализации процессов, порождаемых в результате прохождения потока воздуха через речевой тракт. Эти параметры позволяют не только сформировать описание речевого сигнала в виде последовательности векторов признаков, но и получить оценку спектра сигнала. С другой стороны, коэффициенты этого уравнения могут быть интерпретированы как параметры модели речевого тракта, порождающей речевой сигнал. В этом случае просматривается связь подхода с методами идентификации динамических систем [39, 46], которые в свою очередь дают возможность оценивать как структуры, так и параметры идентифицируемых моделей, а также состояния моделируемых систем. При этом спектральные оценки интерпретируются как передаточные функции речевого тракта, давая возможность оценивания частоты и ширины полосы формант как полюсов этой функции.

Авторегрессионная модель строится в предположении, что текущий отсчёт сигнала может быть представлен линейной комбинацией взвешенных предыдущих отсчётов с некоторой погрешностью, величину которой необходимо минимизировать. Для минимизации обычно используется метод наименьших квадратов, для реализации которого производится суммирование квадратов погрешностей на конечном временном интервале, и частные производные весовых коэффициентов приравниваются нулю, в результате чего получают систему линейных алгебраических уравнений, решение которой даёт оценки коэффициентов линейного предсказания. В зависимости от определения интервала анализа различают автокорреляционный и ковариационный методы линейного предсказания, приводящие к несколько отличающимся по свойствам системам уравнений, для каждой из которых были найдены эффективные вычислительные алгоритмы. В результате реализации процедур могут быть получены различные оценки параметров модели авторегрессии — как собственно коэффициенты линейного предсказания, так и коэффициенты отражения или частной корреляции, между которыми существует взаимно однозначное соответствие. Анализ проводится, как было отмечено, на конечных, обычно частично перекрывающихся интервалах времени длительностью от нескольких миллисекунд для ковариационного метода до 20–30 мс. для автокорреляционного.

Авторегрессионным методам свойственны принципиальные ограничения, в первую очередь чувствительность к помехам, а также недостаточная для высококачественного синтеза речи описательная способность чисто полюсной

акустической модели речевого тракта, в действительности являющейся нелинейной нестационарной динамической системой с распределёнными параметрами [43]. Даже при сосредоточенных параметрах учёт нулей передаточной функции мог бы дать существенное приращение точности модели. Поэтому в последние годы появилось достаточно много работ, направленных на улучшение характеристик авторегрессионных и других параметрических методов оценивания параметров и моделей речевых сигналов.

Анализ существующих и перспективных подходов к улучшению свойств авторегрессионных моделей

Целью настоящей работы является изыскание возможностей преодоления недостатков и ограничений классических авторегрессионных методов. Следует отметить, что не существует единственного метода, обеспечивающего устранение всех недостатков классического авторегрессионного подхода. В современных исследованиях, как правило, применяется комплекс приёмов и решений, в совокупности обеспечивающий улучшение оценок параметров и моделей.

Речевой сигнал как нестационарный случайный процесс может быть представлен как временной ряд с изменяющимися во времени вероятностными характеристиками. Несмотря на непрерывный характер реальных речевых сигналов, представляемых в системах обработки речи в форме аналогового сигнала на выходе микрофона, в современных условиях при подавляющем преобладании цифровых методов обработки информации речевой сигнал рассматривается как дискретная последовательность квантованных по уровню цифровых отсчётов. Для описания свойств таких процессов применяются типовые параметрические модели, устанавливающие статистические связи между членами временного ряда. Таких моделей существует три: это модель авторегрессионного процесса (AP, Autoregressive — AR), процесса скользящего среднего (СС, MA — Moving Average) и комбинированная модель процесса авторегрессии — скользящего среднего (АРСС, ARMA). Последняя модель является наиболее общей и, соответственно, наиболее сложной в анализе и реализации.

Критерием для выбора одной из этих моделей может служить характер передаточных функций динамических систем, порождающих соответствующие процессы. АР-модель имеет передаточную функцию, содержащую только полюса, в то время как СС-модель представлена только нулями передаточной функции. АРСС-модель имеет дробно-рациональную передаточную функцию с нулями и полюсами и, таким образом, в большей степени соответствует модели речевого тракта, имеющей как резонансы, так и антирезонансы. Однако лишь авторегрессионная модель является линейной, что существенно облегчает её реализацию, поэтому именно она до сих пор в основном применяется для анализа речи. При этом увеличение порядка модели позволяет получить приемлемую точность аппроксимации спектральных характеристик. Правда, увеличение размерности противоречит вышеупомянутым требованиям минимизации объёма описаний и может быть преодолено только частично в процессе вторичного анализа речи.

Общая модель порождения многих детерминированных и стохастических процессов с дискретным временем может быть описана разностным уравнением [27]:

$$x(n) = -\sum_{k=1}^{p} a(k)\, x(n-k) + \sum_{k=0}^{q} b(k)\, u(n-k). \qquad (1)$$

Здесь $u(n)$ — входная последовательность, а $x(n)$ — выходная последовательность физически реализуемого, каузального фильтра. В случае речевых сигналов входной последовательностью является выход голосового источника, в большинстве случаев недоступный для наблюдения. Поэтому о нём принимается некоторое допущение, чаще всего в качестве такового принимается белый шум. Уравнение (1) представляет АРСС-модель случайного процесса, которую можно представить как выход цифрового фильтра с дробно-

рациональной передаточной функцией. Числитель этой передаточной функции определяет фильтр с бесконечно-импульсной характеристикой (БИХ), а знаменатель — фильтр с конечно-импульсной характеристикой (КИХ). Соответственно первый определяет модель процесса скользящего среднего, а второй — авторегрессионного процесса.

В частных случаях, когда коэффициенты $a(k)$ либо $b(k)$ равны нулю за исключением $a(0)$ =1, b(0) =1, получаем АР — модель:

$$x(n) = -\sum_{k=1}^{p} a(k)\,x(n-k) + u(n), \qquad (2)$$

либо СС — модель:

$$x(n) = \sum_{k=1}^{q} b(k)\,u(n-k) + u(n). \qquad (3)$$

Параметры этих трёх процессов взаимосвязаны, соотношения между ними рассмотрены, например, в [27]. Существенно то, что лишь АР-параметры линейно связаны с вероятностными характеристиками (например, автокорреляционной последовательностью) процессов, при этом СС и АРСС — параметры могут быть аппроксимированы набором АР-параметров высокого порядка. Отсюда вытекают методы оценивания параметров перечисленных процессов. Существенная проблема при этом заключается в том, что истинные вероятностные характеристики, во всяком случае, для речевых сигналов, неизвестны и к тому же нестационарны, поэтому их приходится заменять выборочными оценками, от их свойств и методов их получения существенно зависят результаты моделирования.

Возможности совершенствования алгоритмов авторегрессионного анализа

Известно, что погрешность предсказания имеет наибольшие значения на интервалах размыкания голосовых складок. Поэтому для улучшения оценок авторегрессионных параметров целесообразно исключать эти интервалы из анализа с применением локального выделителя основного тона [42, 45], либо разделять импульсные отклики голосового источника и речевого тракта, например, с помощью гомоморфной обработки [5,18].

Эффективным подходом для повышения качества формируемых описаний является учёт психоакустических особенностей восприятия речи человеком. Нелинейное преобразование масштаба частот по шкале мелов или барков применяется как в спектральном анализе («сжатое» или «неравномерное» дискретное преобразование Фурье, WDFT — Warped DFT или NDFT — Nonuniform DFT) [11, 29], так и в авторегрессионном и других методах анализа для решения большинства задач речевых технологий. Сюда относятся в первую очередь методы перцептивного линейного предсказания (Perceptual Linear Predictive analysis — PLP) [8] и его последующие модификации (RelAtive SpecTrAl — RASTA PLP) [9, 10]. Достаточно полный анализ и обобщение упомянутых и других подходов к реализации методов линейного предсказания приведены в [16]. В работах [9, 10] было предложено преобразовывать коэффициенты линейного предсказания в кепстральные коэффициенты, что стало в настоящее время практически стандартом, называемым MFCC — Mel Frequncy Cepstral Coefficients. Мел-кепстральные параметры, кроме вектора собственно коэффициентов кепстра, включают обычно первую, а иногда и вторую производные этого вектора, что увели-

чивает размерность вектора параметров втрое. По результатам ряда сравнительных исследований такое описание обеспечивает наиболее высокое качество распознавания. Здесь следует заметить, что, по нашему мнению, векторы производных следует использовать в отдельных, параллельных основному, процессах обработки параметров. Это должно позволить выявлять более полный набор акустических событий в потоке речи, включая кроме статических картин (например, спектральных портретов фонем) также и динамические признаки [32].

Существует два основных алгоритма оценивания MFCC — классический на основе преобразованного по шкале мелов спектра Фурье с последующим логрифмированием и обратным косинусным преобразованием, и на основе преобразования коэффициентов линейного предсказания в коэффициенты мел-частотного кепстра. Реализация этих методов осуществляется также двумя способами — блочным и рекуррентным (в зарубежной литературе называемый адаптивным) [21]. В упомянутой публикации определён метод обобщённого мел-кепстрального анализа, являющийся унифицированной реализацией кепстрального анализа и линейного предсказания, соотношение между которыми задаётся двумя параметрами α и γ, определяющими соответственно степень деформации частотной шкалы и характер спектральной оценки.

Сравнение блочных и рекуррентных подходов

В любом методе анализа сигналов поведение анализируемого процесса рассматривается во временном окне, длительность которого должна обеспечивать получение состоятельных в вероятностном смысле характеристик. В соответствии с этим число отсчётов сигнала в окне анализа должно быть, по крайней мере, на порядок больше числа оцениваемых параметров. Многие стандартные алгоритмы кодирования речевых сигналов используют окна длительностью 10…20 мс, что при частоте дискретизации 8 кГц даёт число отсчётов 80…160, при этом число параметров составляет 10. Например, минимизация погрешности линейного предсказания предполагает накопление некоторой последовательности отсчётов речевого сигнала, позволяющей получить выборочную оценку корреляционных связей между отсчётами в форме автокорреляционной функции либо ковариационной матрицы. На основе (или в результате получения) этих оценок формируется система уравнений, решение которых даёт оценки параметров авторегрессионной модели.

С другой стороны, темп получения информации о параметрах должен задаваться с учётом скорости изменения параметров сигнала. Исходя из оценок интервалов стационарности речевых сигналов, полученных на основе различных предпосылок (теория сигналов и случайных процессов, психоакустика и нейрофизиология и т.п.), интервал дискретизации вектора параметров должен составлять от одной до нескольких десятков миллисекунд. Поэтому перемещение окна анализа вдоль оси времени осуществляется с перекрытием на величину шага дискретизации вектора параметров. Это вызывает определённые проблемы и ограничения, если обработка сигнала в интервале анализа должна осуществляться с применением взвешивающего окна.

Надо учитывать, что взвешивание приводит к некоторому сужению интервала анализа за счёт спада на краях окна. Следовательно, получаемые таким образом оценки параметров будут охватывать относительно меньше отсчётов сигнала, чем при использовании прямоугольного окна, т.е. возникает необходимость достижения компромисса между разрешением анализа по времени и достоверностью получаемых оценок. Результат взвешивания прямоугольным окном существенно зависит от значений сигнала на краях окна, и при наличии выбросов на границе значение оценки достаточно сильно флуктуирует. Это особенно заметно, когда ширина окна близка к кратному значению основной периодичности в сигнале. При этом периодичность оценки определяется также и высшими гармониками основной частоты. Та же оценка, полученная с помощью весового окна, флуктуирует даже сильнее, но её периодичность определяется в большей степени основной частотой сигнала. В спектральной области отмеченные особенности объясняют-

ся более высоким уровнем боковых лепестков у частотной характеристики прямоугольного окна. Отметим, что рекуррентные алгоритмы предполагают использование, как правило, прямоугольного окна, взвешивание другими окнами приводит к необходимости дополнительной фильтрации выходных данных, несколько снижая вычислительную эффективность рекуррентных методов [32, 33].

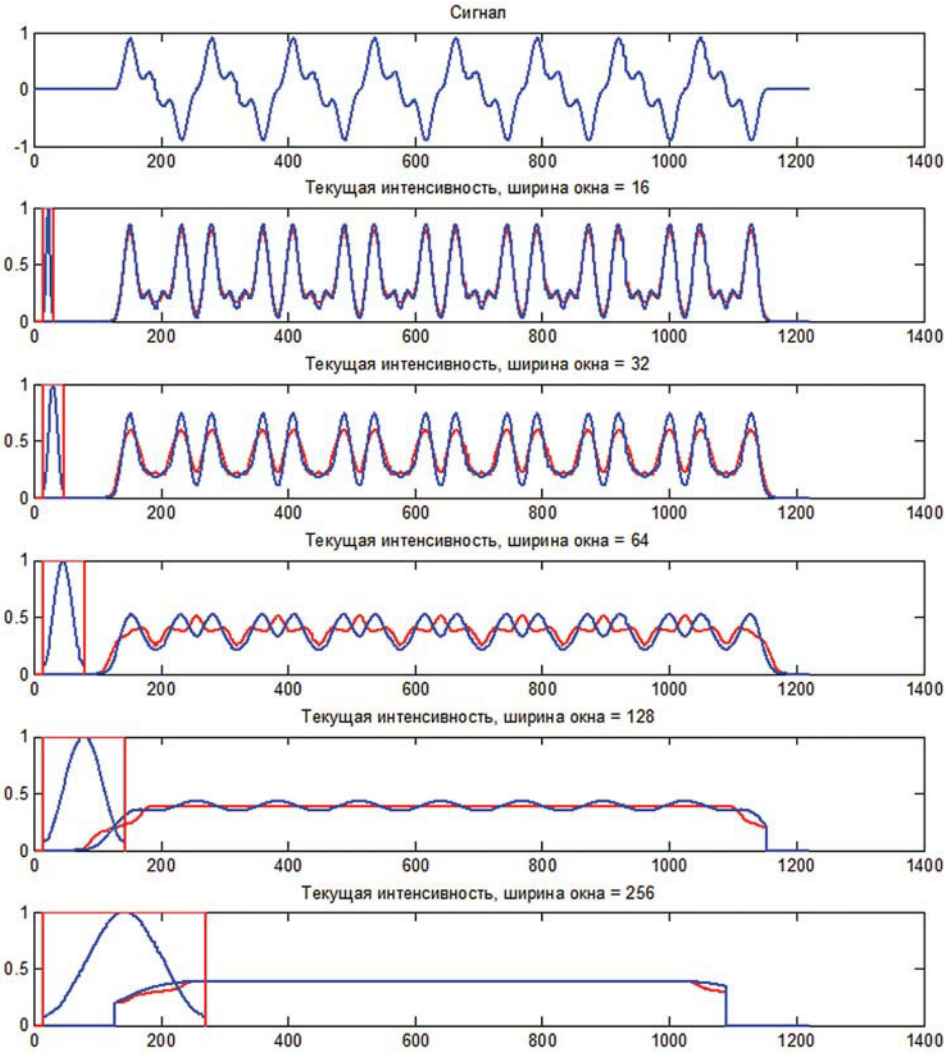

Рис. 1. Оценки интенсивности при различной ширине окна анализа. Красные линии — прямоугольное окно, синие точки – окно Хэмминга

В качестве примера на рисунке 1 приведены оценки текущей интенсивности тестового сигнала, представляющего сумму трёх синусоид с отношением частот 1 / 2 / 4 и амплитудами 0.6, 0.2, и 0.2 соответственно. Оценка выполнялась согласно выражению: $s = k \cdot \sum_{i=1}^{N} w_i \cdot abs(x_i)$, где x_i — отсчёт сигнала, w_i — значение весовой функции окна, N — ширина окна в числе отсчётов, k — коэффициент, учитывающий площадь весовой функции, $k = 1 \Big/ \sum_{i=1}^{N} w_i$.

Следует отметить, что приведённый пример отражает поведение оценок, полученных со сдвигом окна на один отсчёт, т.е. в скользящем режиме. Блочный

подход предполагает перемещение окна скачками, с частичным перекрытием. Характер оценок зависит при этом от степени перекрытия, что отражено на рисунке 2, где приведены те же оценки при постоянной ширине окна в 160 отсчётов, но с различной степенью перекрытия. Целесообразно принять за меру перекрытия отношение *(ширина–шаг) / шаг,* тогда в примере перекрытие меняется от 159 до 1.

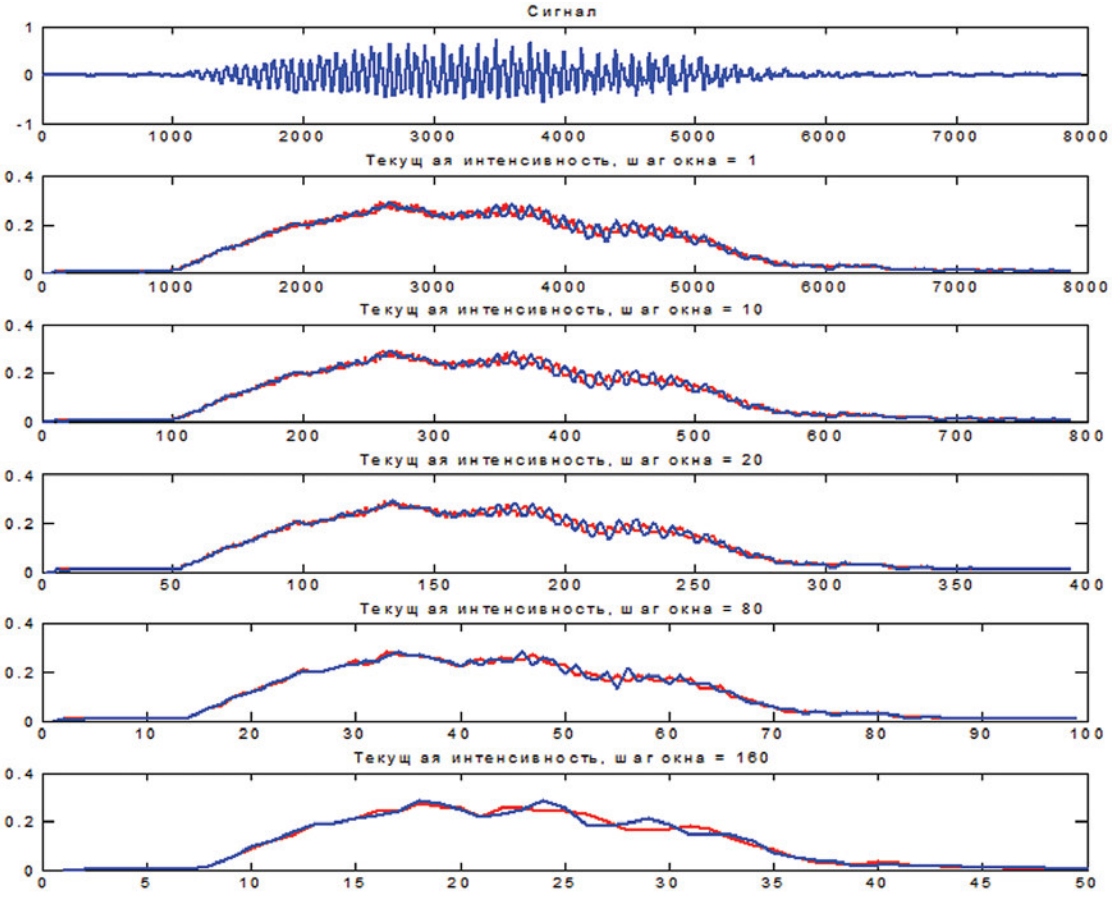

Рис. 2. Оценки интенсивности при различном шаге окна анализа. Красные линии — прямоугольное окно, синие точки — окно Хэмминга. Ширина окон 160 отсчётов

Очевидно, что ширина окна и его форма должны выбираться по-разному при решении различных задач. Если оцениваются параметры для распознавания речи, окно должно сглаживать периодичность основного тона, в то время как для оценивания самого основного тона важно эту периодичность выделять, подавляя высшие гармоники.

Рекуррентные методы обеспечивают обновление параметров модели на каждый вновь поступивший отсчёт речевого сигнала. Такое описание может показаться избыточным, и для реализации процедур распознавания речи это справедливо, поскольку шаг дискретизации по времени информативных параметров обычно принимается 10…40 мс, минимум 1 мс. Однако рекуррентные методы имеют некоторые преимущества, компенсирующие кажущийся недостаток. При синтезе речи в любом случае выходные отсчёты необходимо получать в том же темпе, что и входные при анализе, поэтому применение рекуррентных методов предпочтительно в коммуникационных приложениях [12]. «Скользящий» режим эффективен и при вычислении корреляционных, спектральных и других характеристик сигналов, позволяет с точностью до одного отсчёта определять границы речевых элементов и акустических событий в произнесении, а усредне-

ние оцениваемых параметров позволяет формировать описания речевых реализаций с требуемым, в том числе неравномерным по времени, шагом. Наконец, рекуррентный анализ в большей степени адекватен процессам слухового восприятия речи человеком, происходящим в непрерывном времени с использованием аналоговых или аналого-дискретных «устройств». Наличие параметров, обновляемых в темпе обновления входного речевого сигнала, позволяет реализовывать сложные комплексные процедуры обработки речевых сигналов, использующие разнотемповость, конвейеризацию и распараллеливание процедур анализа и принятия решений. При этом избыточность рекуррентных процедур компенсируется повторным использованием полученных оценок в различных ветвях и подзадачах алгоритма.

Выбор между блочным и рекуррентным подходом непрост и неоднозначен. Решение должно учитывать характер решаемой задачи, требования к вычислительным ресурсам и методам реализации алгоритмов. Здесь необходимо отметить, что современные возможности микроэлектронных технологий практически снимают ограничения по программной и аппаратной реализации алгоритмов любой сложности в реальном времени, если нет ограничений по стоимости, энергоёмкости, массогабаритным и иным параметрам. Во всяком случае, эти ограничения не должны приниматься во внимание при проведении фундаментальных исследований, имеющих целью достижение новых, наивысших результатов в речевых технологиях. В связи с этим возникает потребность в ревизии применяемых методов и алгоритмов как в теоретическом плане, так и в плане проведения сравнительно-экспериментальных исследований существующих и новых (или хорошо забытых старых) подходов.

Инварианты и разложения вероятностных характеристик

Как отмечалось выше, для нахождения авторегрессионных параметров используются вероятностные характеристики анализируемых процессов – выборочная автокорреляционная $R_n(k)$ либо ковариационная $C_n(i,k)$ последовательность, вычисляемые по последовательности отсчётов анализируемого сигнала в соответствии с соотношениями:

$$R_n(k) = \sum_{m=0}^{N-1-k} x_n(m) \cdot x_n(m+k), \tag{4}$$

$$C_n(i,k) = \sum_{m=p}^{N-1} x_n(m-i) \cdot x_n(m-k), \tag{5}$$

$$0 \le i \le p, \quad 1 \le k \le p,$$

где p — порядок модели.

Не затрагивая вопросов реализации хорошо известных классических методов решения авторегрессионных уравнений, отметим возможности получения дополнительной информации о речевых сигналах, которую можно извлечь из этих характеристик. Эта информация может оказаться полезной для решения смежных задач, обычно возникающих в процессе формирования параметрического описания речевых сигналов, например, классификации «пауза–тон–шум», определения границ слов либо слитно произносимых фраз, сегментация речевого потока на слоги, фонемы и другие речевые элементы, включая интервалы смыкания-размыкания голосовых

складок. Для решения указанных задач требуется обобщённая либо детализированная информация об энергетических, частотных, временных характеристиках речевых сигналов.

Поскольку выборочные оценки корреляционных свойств получены на интервале, из них можно извлечь усреднённую на этом интервале информацию, например, о текущей энергии либо интенсивности. Известно, что эта информация содержится в нулевом отсчёте автокорреляционной функции. Автокорреляционная функция часто используется для решения задачи выделения основного тона. В работе [42] предлагается использовать для выделения параметров речевых сигналов инварианты автокорреляционной матрицы. Например, указывается, что значение её определителя позволяет определить интервалы смыкания-размыкания голосовых складок.

На рисунке 3 кроме поведения определителя показаны ещё два графика – след матрицы собственных значений и величина, обратная обусловленности ковариационной матрицы, определяемой как отношение максимального собственного значения к минимальному. Выбранные параметры окна анализа позволили сохранить и даже подчеркнуть периодичность основного тона.

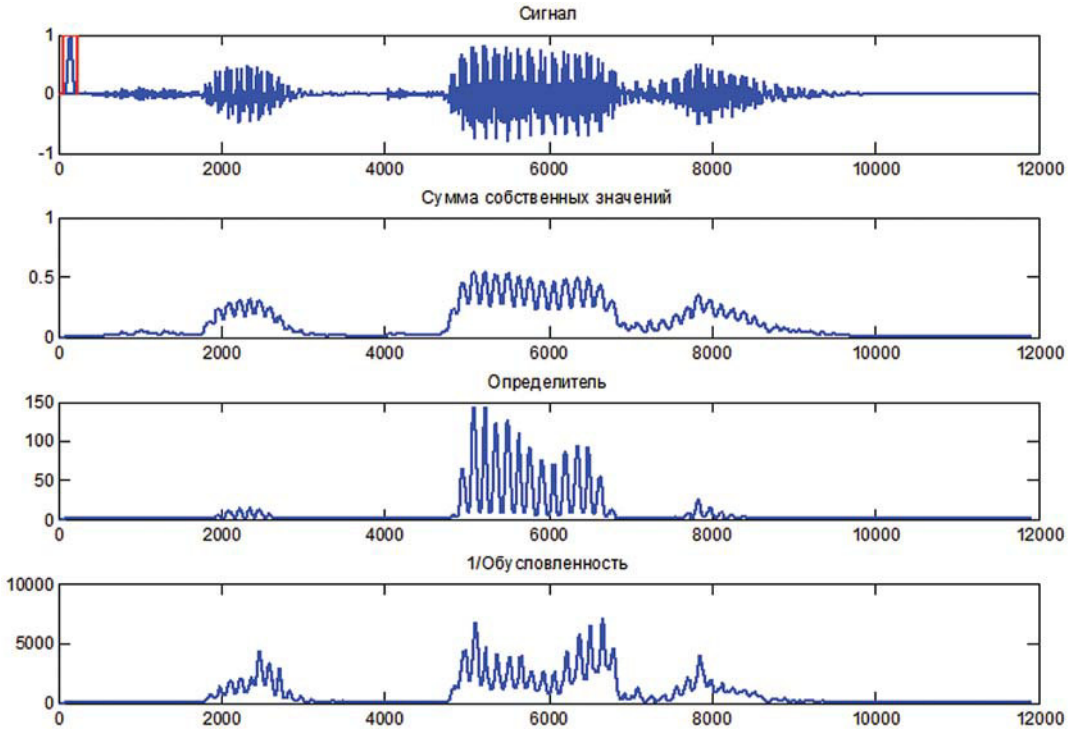

Рис. 3. Инварианты ковариационной матрицы сигнала размерности 4x4. Реализация слова «четыре», параметры дискретизации — 16 кГц, 16 бит. Окно анализа 160 отсчётов (10 мс), шаг 1 отсчёт

Формирование векторов информативных параметров на основе сингулярного разложения автоковариационных матриц с возможностью адаптации к характеристикам шумов

В последние десятилетия всё большее внимание исследователей уделяется методам обработки речевых сигналов на основе разложения матриц по собственным значениям и собственным векторам (РСЗ): Eigen Value Decomposition — EVD, Singular Value Decomposition — SVD, Generalised Singular Value Decomposition — GSVD и другие модификации

этого подхода. Математические основы метода известны давно и применялись в статистическом анализе и распознавании образов (разложение Карунена-Лоэва — KLT, метод главных компонент [24, 44]) для сжатия данных и анализа их структуры, а также сокращения размерности пространства признаков. Для спектрального оценивания применяется гармоническое разложение Писаренко, метод Прони, методы MUSIC и EV [27], также основанные на разложении по собственным значениям. Представляет интерес возможность применения сингулярных разложений для исследования динамики систем, в том числе нелинейных, с распределёнными параметрами, к которым относится и речевой тракт [28, 37].

Несмотря на достаточно длительный период развития и очевидную перспективность рассматриваемого подхода, разработку его нельзя считать полностью завершённой, о чём свидетельствует не снижающееся число публикаций. Первые работы по применению сингулярных разложений для шумоочистки речи появились в 1991 году [2, 3]. К 1995 году были сформированы основные аспекты решения проблемы [4]. В работе [7] представлено большинство основных этапов и приёмов реализации похода. Последующие публикации посвящены в различной степени либо обобщению и уточнению процедур [20], либо развитию дополняющих подход методов [1, 6]. В частности, представляет интерес направление, развивающее рекуррентный (в зарубежных публикациях чаще называемый адаптивным) подход к реализации процедур [13, 1, 16, 19, 22]

Идея подхода состоит в разделении сигналов по корреляционным свойствам на два подпространства — сигнала (с шумом) и шума. Это даёт основания предполагать возможность очистки сигналов от шума и на основе этого получать решения задач речевых технологий, работоспособные в реальной, чаще всего зашумлённой обстановке.

Речевой сигнал при блочной цифровой обработке обычно разбивается на перекрывающиеся отрезки — блоки, или фреймы, представляемые в виде последовательности отсчётов, содержащих компоненты собственно речевого сигнала и наложенного на него аддитивного шума:

$$\mathbf{x} = \left(x_1, x_2, x_3, ..x_k\right)^T = \mathbf{s} + \mathbf{n}, \quad (6)$$

где k — размер блока, **s** и **n** — векторы соответственно сигнала и шума, той же размерности k, что **x**.

Сигнал **x** может быть отображён в многомерное пространство «встроенной» размерности m:

$$\mathbf{H} = \left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..\mathbf{x}_n\right), \quad (7)$$

где $\mathbf{x}_i = \left(x_i, x_{i+1}, x_{i+2}, x_{i+3}, ..x_{i+m-1}\right)^T$ — часть последовательности **x** длины m, $i = 1..n$, $n = k - p + 1$, **H** — Ганкелева траекторная матрица.

Сингулярное разложение траекторной матрицы определяется как произведение трёх матриц:

$$\mathbf{H} = \mathbf{USV}^T \quad (8)$$

Здесь **S** — матрица m x n, на главной диагонали которой расположены сингулярные значения, матрицы **U** размером m x m и **V** размером n x n составлены соответственно из левых и правых собственных векторов.

Предполагается, что речь и шум занимают различные подпространства — подпространства сигнала и шума размерностью p и q соответственно, причём сигнал отображается на измерения, соответствующие наибольшим, а шум — наименьшим сингулярным значениям.

Если тем или иным способом изменить веса компонентов пространств сигнала и шума, задаваемых собственными значениями, обратное преобразование (9) позволит получить улучшенный речевой сигнал.

$$\mathbf{H}_e = \mathbf{UGSV}^T \quad (9)$$

Здесь **G** — весовая матрица, определяющая увеличение либо уменьшение вклада каждого (или некоторых) измерений в выходной сигнал, восстанавливаемый усреднением антидиагоналей выходной Ганкелевой матрицы.

Для каждого блока процедура повторяется, выходной сигнал объединяется путём суммирования перекрывающихся блоков. При блочной обработке наиболее подходящим для взвешивания входных векторов **x** (6) является использование окна Ханна.

В том случае, когда нет необходимости в получении очищенного от шума речевого сигнала (например, при формировании параметрических описаний для распознавания речи либо дикторов) есть возможность получения улучшенных оценок вероятностных характеристик сигнала. Для этого можно использовать тот факт, что левые собственные векторы сингулярного разложения точно совпадают с матрицей собственных векторов автокорреляционной матрицы. Поэтому улучшенную оценку этой матрицы можно получить, используя преобразование

$$\mathbf{R}_e = \mathbf{UGSU}^T \quad (10)$$

После этого можно получить улучшенный вариант авторегрессионных параметров классическим методом.

Остаётся решить, как выбирать веса для рассмотренного преобразования. Примитивное решение — обнулить те сингулярные значения, которые отвечают за подпространство шума. Однако это приводит к существенным искажениям речевого сигнала, а также появлению характерных «музыкальных тонов». Дело в том, что размерность сигнального пространства достаточно сильно варьирует при произнесении различных звуков и довольно заметно увеличивается для шумных звуков — в несколько раз по сравнению с гласными. Поэтому необходимо правильно оценивать размерности подпространств и рационально выбирать весовую матрицу.

Кроме этого, рассмотренный подход справедлив в предположении о том, что шум — некоррелированный случайный процесс с равномерным распределением, то есть белый шум. В реальности шумы чаще всего окрашены, поэтому возникает необходимость решения задачи предварительного «обеления» (prewhitening) траекторной матрицы перед сингулярным разложением с последующим «окрашиванием» (dewhitening) после обратного преобразования. Соответственно необходимо реализовать процедуры оценивания характера шумов для решения о необходимости применения указанных преобразований. Это требуется и для случая, когда характер шумов изменяется в процессе обработки, то есть существует необходимость постоянной адаптации параметров алгоритма.

Ещё одно направление совершенствования рассматриваемого подхода состоит во включении в него процедур перцептивной обработки, то есть учёта особенностей слухового восприятия речи, доказавших свою эффективность и в других подходах к обработке речевых сигналов [31].

Последовательное решение задач шумоочистки и первичного анализа позволяет получать устойчивые к шумам оценки параметров. Но более эффективным решением представляется совмещение процедур, используемых в процессе шумоочистки, с получением устойчивых к шумам информативных параметров речевого сигнала. Поскольку в обоих случаях исходными данными служат спектральные либо корреляционные характеристики сигналов, такое решение можно признать рациональным, особенно в случае формирования описаний речевых сигналов для их распознавания, без необходимости восстановления очищенного от шума сигнала.

Реализация подхода к оцениванию информативных параметров речи на основе использования сингулярных разложений требует построения комплексной процедуры анализа, включающей, помимо основного процесса, ряд смежных процедур (выделение признаков «тон-шум-пауза», классификация и идентификация шумов и др.). Представляется целесообразным выбрать рекуррентный вариант построения алгоритмов оценивания автоковариационных матриц, основного тона, собственных значений и собственных векторов, размерностей подпространств сигнала и шума.

Экспериментальное исследование влияния алгоритмов шумоочистки на надёжность распознавания изолированно произносимых слов

В дополнение к оценкам улучшения разборчивости восприятия речи алгоритмами, описанными в [30, 31], был поставлен эксперимент по оценке влияния шумоочистки на надёжность распознавания изолированно произносимых слов. Речевой материал, использованный в эксперименте, получен от одного диктора, произносившего цифры от ноля до девяти сериями по пять произнесений каждого слова в различные периоды времени (с 2005 по 2013 г.), с использованием разных микрофонов и в различной акустической обстановке, но при отношении сигнал/шум (ОСШ, SNR) не хуже 40 дб. Таким образом, получено шесть серий по пять реализаций десяти слов, т.е. в общей сложности триста реализаций с частотой дискретизации 8 КГц и разрядностью 16 в wav - формате. Для удобства планирования экспериментов имена файлов сформированы по схеме

s<*s*>r<*r*>w<*ww*>n<*nn*>.wav,
где: s — от «speaker»;
s — номер диктора, от 0 до 5;
r — от «realization»;
r — номер реализации, от 0 до 4;
w — от «word»;
ww — номер слова от 0 до 9, для цифр соответствует их значению;
n — от «noise»;
nn — значение ОСШ: 00, 06, 12, 20, 40.

Распознавание осуществлялось сравнением с эталонами на основе алгоритма динамического программирования по методу «скользящего экзамена», т.е. каждая из 300 реализаций поочерёдно сравнивалась со всеми остальными, выступавшими в качестве эталонов.

В качестве эталонов, помимо исходных реализаций, использовались эти же реализации с наложенным на них белым шумом для ОСШ 0, 6, 12, 20 и 40 дБ, то есть ещё 5*300 = 1500 реализаций. Зашумлённые реализации подвергались обработке двумя алгоритмами шумопонижения [31] — на основе спектрального вычитания NRS и на основе обработки сигнала в подпространствах PCSS. Таким образом, получено ещё 3000 реализаций «очищенных» сигналов.

Описания реализаций получены с использованием метода MFCC в его классической форме — через БПФ, мел-преобразование, логарифмирование и косинусное преобразование. Отличие от классики заключается в том, что вычислялись не 13, а 16 коэффициентов мел-кепстра, а первая и вторая производные вектора MFCC-параметров рассчитывались по 5-точечной схеме численного дифференцирования со сглаживанием:

$$x'(n) = \{8 * [x(n+1) - x(n-1)] + x(n-2) - x(n+2)\}/12,$$

где $x'(n)$ — n-й отсчёт вектора параметров.

После дифференцирования последовательность векторов производных подвергалась дополнительному сглаживанию по 5-точечной схеме:

$$\tilde{x}'(n) = \{6 * x'(n) + 4 * [x'(n+1) + x'(n-1)] + x'(n-2) + x'(n+2)\}/16.$$

Для каждой реализации матрицы параметров, их первой и второй производной записывалась в отдельные файлы с теми же именами, но с расширениями «.mf0», «.mf1» и «.mf2» соответственно. Это обеспечивает возможность изменять объём описания, используя каждую матрицу параметров отдельно либо с сочетании с другими и получая при этом размерность пространства параметров от 16 до 48.

Эксперимент по распознаванию организуется следующим образом. В память компьютера загружаются матрицы реализаций, образуя массив эталонов и массив реализаций. В процессе загрузки формируется индексный массив, содержащий адреса начала каждой реализации. Это обеспечивает возможность выделить реализацию из массива, использовав два индекса: текущей реализации и следующей, уменьшенный на единицу.

Возможны две схемы организации тестирования. В первом случае один и тот же набор реализаций используется и как эталонный, и как проверочный, поэтому формируется общий массив реализаций. Сравнение реализаций с эталонами организуется в шести вложенных циклах: по номерам эталонных дикторов, номерам произнесений эталонных слов и номерам эталонных слов в словаре, а затем по номерам дикторов, произнесений и слов, предъявляемых в качестве тестовых. При этом сравнение тестовой реализации со «своим» эталоном не производится, а оценке расстояния в этой паре присваивается заведомо большое значение. Все расстояния вычисляются и заносятся в таблицу расстояний, имеющую размерность 10 эталонов*30 реализаций. Результат распознавания определяется индексом эталона минимального элемента таблицы расстояний. Поскольку процедура сравнения регулярна, то известен индекс тестового слова, и содержимое адресуемой этими индексами ячейки таблицы результатов (размерности 10*10) увеличивается на единицу. Если ячейка не находится на диагонали квадратной матрицы результатов, фиксируется ошибка распознавания. При 100% правильности распознавания таблица результатов диагональна с содержимым диагональных ячеек 30. Процент правильно распознанных слов равен сумме диагональных элементов, делённой на 300 и умноженной на 100 (т.е. сумма/3). Таблицы расстояний и результатов записываются в текстовый файл, что позволяет детально проанализировать ошибки распознавания.

Пример таблицы результатов для сравнения исходных реализаций с исходными эталонами приведён ниже.

Таблица 1

r\e:	0	1	2	3	4	5	6	7	8	9
0	30	0	0	0	0	0	0	0	0	0
1	0	29	0	1	0	0	0	0	0	0
2	0	0	30	0	0	0	0	0	0	0
3	0	0	0	30	0	0	0	0	0	0
4	0	0	0	0	30	0	0	0	0	0
5	0	0	0	0	0	30	0	0	0	0
6	0	0	0	0	0	0	30	0	0	0
7	0	0	0	0	0	0	0	30	0	0
8	0	0	0	0	0	0	0	0	30	0
9	0	0	0	0	0	0	0	0	0	30

Процент = 99.7

Второй вариант тестирования предусматривает сравнение реализаций из двух различных наборов, например, исходных и зашумлённых или очищенных реализаций для различных значений SNR. В остальном схемы процедур аналогичны, только тестовая реализация сравнивается со всеми эталонами.

Тестирование включает семь групп экспериментов, для удобства обзора результатов обозначенных нижеприведёнными аббревиатурами:

OEOR: исходные (original) эталоны — исходные реализации;
OENR: исходные эталоны — зашумлённые (noised) реализации;
NENR: зашумлённые эталоны и реализации с одинаковым уровнем шума;
OENRB: исходные эталоны — зашумлённые реализации с границами, определёнными на исходном сигнале (искусственный пример);
NENRB: аналогично предыдущему варианту, но с шумными эталонами;
OECN: исходные эталоны — реализации, очищенные от шума алгоритмом NRS;
OECP: то же с применением алгоритма PCSS.

Результаты тестирования сведены в таблицу 2.

Таблица 2

Результаты распознавания в условиях различных уровней шумов и применения алгоритмов шумоочистки

SNR	OENR	OENRB	NENR	NENRB	OECN	OECP	OEOR
0	31,3	79	62,3	95,3	87,7	74	99.7
6	82,7	89,3	93	98,3	94,3	87,7	
12	92,3	93,2	95,3	99	96	97,7	
20	97	97	98	99,7	96	98	
40	99,7	98,7	98,7	99,7	98,3	99	

Для наглядности приводятся графики зависимостей надёжности распознавания шумных реализаций по исходным и шумным эталонам (рисунок 4).

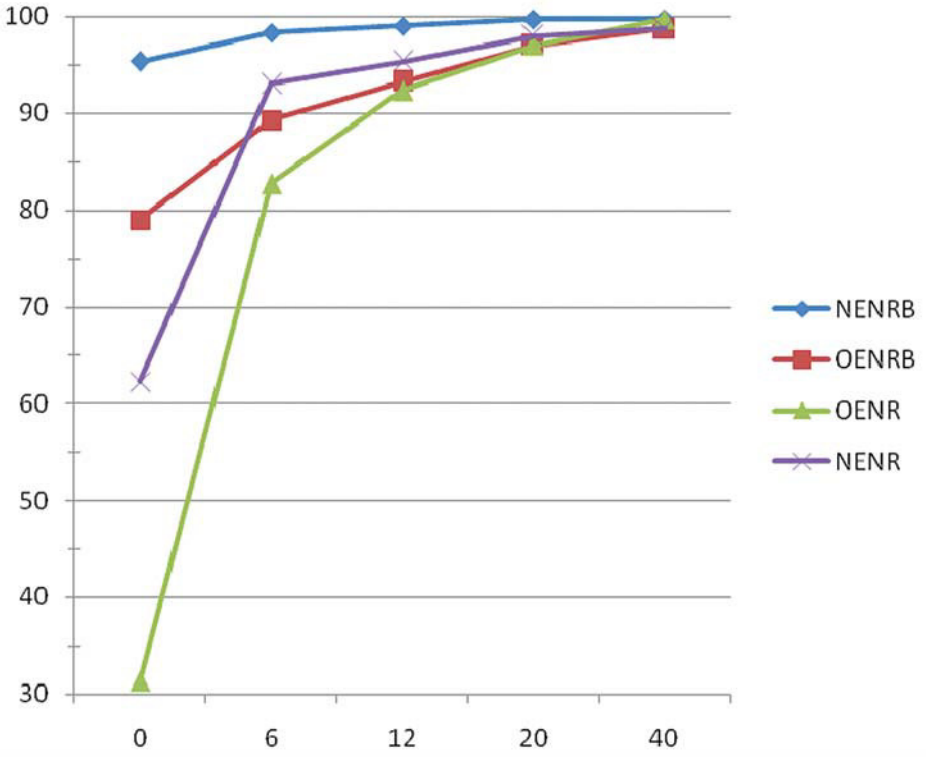

Рис. 4. Влияние погрешностей выделения границ произнесения под действием шума

Сюда включены результаты искусственного примера, когда описания зашумлённых реализаций сформированы в точных границах, оценённых на исходных сигналах. Как и следовало ожидать, в этом случае результаты существенно лучше. Подтверждено также, что результаты распознавания в условиях шума выше в случае использования эталонов, полученных в тех же условиях, и здесь основную долю ошибок также следует отнести к погрешностям определения границ произнесения. Следует при этом отметить, что в экспериментах использовался простейший алгоритм выделения границ на основе сравнения с порогами текущих энергии и частоты переходов через ноль.

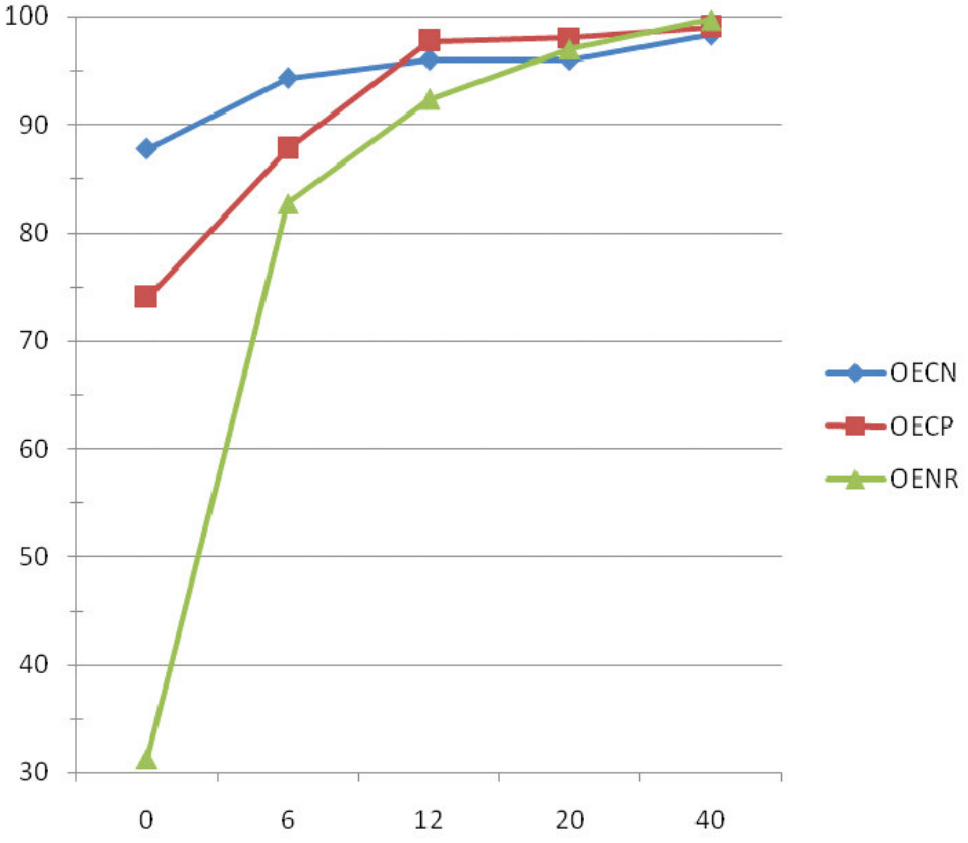

Рис. 5. Эффект применения алгоритмов шумопонижения

Не вызывает удивления и результат применения алгоритмов шумоочистки (рис. 5). К сожалению, приемлемая точность распознавания в случае шумоочистки использованными алгоритмами достигается лишь при ОСШ больше 10. Очевидно, этот порог можно понизить комплексным решением, предусматривающим использование более совершенного алгоритма выделения границ произнесения, а также совершенствование алгоритмов шумопонижения и методик их применения, в частности, комбинирования нескольких алгоритмов в одной процедуре [30].

Литература

1. *Afzalian A., Karami mollaei M.R., Ghasemi J*. A Combined Voice Activity Detector Based On Singular Value Decomposition and Fourier Transform. — Signal Processing: An International Journal (SPIJ), Volume(4): Issue(1) P. 54–61.

2. *Bakamidis S., Dendrinos M., Carayannis G*. SVD Analysis by Synthesis of Harmonic Signals. — *Acoustics Speech and Signal Processing (ASSP)*, IEEE, Vol. 39, №.2, P. 472–477. Feb. 1991.

3. *Burg J.P*. A New Analysis Technique for Time Series Data. — Proc. NATO Advanced Study Institute on Signal Proc, Enschede Netherlands, 1968.

4. *Ephraim Y., Van Trees H.L*. A signal subspace approach for speech enhancement — IEEE Trans. Speech and Audio Processing, vol. 3, P. 251–266, July 1995.

5. *Fattah S.A., Zhu W-P., Omair Ahmad M.O*. A Ramp Cosine Cepstrum Model for the Parameter Estimation of Autoregressive Systems at Low SNR. — EURASIP Journal on Advances in Signal Processing, vol. 2010, 15 pages, 2010.

6. *Ghasemi Jamal, Mollaei Mohammad Reza Karami*. A New Approach for Speech Enhancement Based on Eigenvalue Spectral Subtraction. — Signal Processing: An International Journal, (SPIJ) Volume(3), Issue(4) 34-41.

7. *Hansen P.S.K*. Signal Subspace Methods for Speech Enhancement. — Ph.D. Thesis LYNGBY 1997 IMM-PHD-1997-42.

8. *Hermansky H*. Perceptual linear predictive (PLP) analysis of speech // Journ. Acoust. Soc. Am. 1990. V. 87. № 4. P. 1738—1752.

9. *Hermansky H., Morgan N., Baya A., Kohn P*. RASTA-PLP speech analysis technique — IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., vol.1. P. 121–124.

10. *Hermansky H., Morgan N*. RASTRA of processing of speech. IEEE Trans. Speech Audio Process. 1994, 2 (4), 578–589.

11. *Hwang J.J., Cho S.G., Moon J., Lee J.W*. Nonuniform DFT based on non-equispaced sampling. — Proceedings of the 5th WSEAS Int. Conf. on Signal, Speech And Image Processing, Corfu, Greece, August 17–19, 2005 (p. 11–16).

12. *Itakura F., Saito S*. Speech Analysis-Synthesis System Based on the Partial Autocorrelation Coefficient. — Acoust. Soc. Jap. Meeting, 1969.

13. *Jensen S.H., Jeppesen M., Rodbro C.A*. Recursively Updated Eigenfilterbank For Speech Enhancement. — Center for Person Kommunikation (CPK), Aalborg University, DK-9220-Aalborg, Denmark.

14. *Karsmakers P*. Perceptual Speech Enhancement with SVD-based Subspace Filtering. — Project Report: Katholieke Universiteit Leuven, May 2004.

15. *Matrouf D., Gauvain J.-L*. Using AR HMM state-dependent filtering for speech enhancement Proceeding ICASSP '99. Vol. 2 1999.

16. *Moonen M., Spriet A., Wouters M.J*. A multichannel subband GSVD based approach for speech enhancement in hearing aids. — European Transactions on Telecommunications, Volume 13, № 2, 2001.

17. *Moonen M., van Waterschoot T*. Comparison of Linear Prediction Models for Audio Signals. —EURASIP Journal on Audio, Speech, and Music Processing, vol. 2008, Article ID 706935, 24 pages, 2008. doi:10.1155/2008/706935.

18. *Rahman M.S. Shimamura T.* Linear Prediction Using Refined Autocorrelation Function. — EURASIP Journal on Audio, Speech, and Music Processing, vol. 2007, Article ID 45962, 9 pages.

19. *Rezayee A., Gazor S.* An adaptive KLT approach for speech enhancement. IEEE Trans. Speech Audio Processing, vol. 9. P. 87–95, Feb. 2001.

20. *Soumya T.S., Soumya V.J, Soman K.P., Vidyapeetham A.V.* Singular Value Decomposition A Classroom Approach. International Journal of Recent Trends in Engineering. Vol. 1, № 2, May 2009.

21. *Tokuda K., Kobayashi T., Masuko T., Imai S.* Mel-Generalized Cepstral Analysis — A Unified Approach to Speech Spectral Estimation. — Third International Conference on Spoken Language Processing (ICSLP 94) Yokohama, Japan September 18-22, 1994.

22. *Uhl C., Lieb M.* Experiments With An Extended Adaptive SVD Enhancement Scheme For Speech Recognition In Noise. Philips Research Laboratories.

23. *Wiener N.* Extrapolation Interpolation and Smoothing of Stationary TimeSeries (M. I. T. Press, Cambridge, Massachusetts, 1966).

24. *Андерсон Т.* Статистический анализ временных рядов: Пер. с англ. / Под.ред. Ю.К. Беляева. М.: Мир, 1976. 755 с.

25. *Колмогоров А.Н.* Стационарные последовательности в гильбертовом пространстве. Бюллетень М!ГУ. 1941. № 6. С. 1–40.

26. *Маркел Дж.Д., Грей А.Х.* Линейное предсказание речи. М.: Связь, 1980. 308 с.

27. *Марпл С.Л.* Цифровой спектральный анализ и его приложения. М: Мир, 1990.

28. *Перервенко Ю.С.* Исследование инвариантов нелинейной динамики речи и принципы построения системы аудиоанализа психофизиологического состояния. Дисс. к.т.н. Таганрог: Южный федеральный университет, 2009.

29. *Петровский А.А., Иванов А.В.* Моделирование аудиторной суппрессии в частотной области на основе СДПФ.

30. *Петровский А.А., Азаров И.С., Лихачев Д.С., Ромашкин Ю.Н., Жигулёвцев Ю.Н., Харламов А.А.* Фильтрация речи на фоне полигармонических и стохастических помех // Речевые технологии. 2012. № 3. С. 45–57.

31. *Петровский А.А., Азаров И.С., Лихачев Д.С., Ромашкин Ю.Н., Жигулёвцев Ю.Н., Харламов А.А.* Шумоподавление на основе перцептивных алгоритмов спектрального вычитания и обработки сигналов в подпространствах // Речевые технологии. 2012. № 4.

32. *Плотников В.Н., Суханов В.А., Жигулёвцев Ю.Н.* Речевой диалог в системах управления. М.: Машиностроение, 1988. 224 с.

33. *Плотников В.Н., Суханов В.А., Жигулёвцев Ю.Н., Белинский А.В.* Цифровые анализаторы спектра. М.: Радио и связь, 1990. 184 с.

34. *Прохоров Ю.Н.* Статистические модели и рекуррентное предсказание речевых сигналов. М.: Радио и связь, 1984. 240 с.

35. *Рао С.Р.* Линейные статистические методы и их применения / Пер. с англ. М., 1968.

36. *Робинсон Э.А.* История развития теории спектрального оценивания // ТИИЭР. Т. 70. № 9. С. 6–32.

37. *Свиридов А.А.* Прогрессивное кодирование аудио с помощью сингулярного разложения. // Наука и образование: Электронный научно-технический журнал. М.: МГТУ им. Н.Э. Баумана.

38. *Себер Дж*. Линейный регрессионный анализ: Пер. с англ. / Под ред. М.Б. Малютова. М.: Мир, 1980. 456 с.

39. *Сейдж Э.П., Мелс Дж.Л*. Идентификация систем управления: Пер. с англ. / Под ред. Н.С. Райбмана. М.: Наука, 1974. 246 с.

40. *Сейдж Э., Мелс Дж*. Теория оценивания и ее применение в связи и управлении: Пер. с англ. / Под ред. Б.Р. Левина. М.: Связь, 1976. 495 с.

41. *Смирнов Н.В., Дунин-Барковский И.В*. Курс теории вероятностей и математической статистики для технических приложений. 3 изд., М., 1969.

42. *Собакин А.Н*. Анализ артикуляционных характеристик речи на базе корреляционной матрицы // X сессия РАО: Сб. тр. М.: ГЕОС, 2000. С. 268–270.

43. *Сорокин В.Н*. Синтез речи. М.: Наука, 1992. 392 с.

44. *Фукунага К*. Введение в статистическую теорию распознавания образов: пер. с англ. / под ред. А.А. Дорофеюка. М.: Наука, 1979. 368 с.

45. *Цыплихин А.И*. Анализ и автоматическая сегментация речевого сигнала. Автореферат дисс. к.т.н. М., 2006 г.

46. *Эйкхофф П*. Основы идентификации систем управления. М.: Мир, 1975.

Сведения об авторах

Жигулёвцев Юрий Николаевич —
кандидат технических наук, старший научный сотрудник, доцент МГТУ им. Н.Э. Баумана. Окончил в 1969 г. МГТУ им. Н.Э. Баумана по специальности "Системы автоматического управления". Автор более 80 научных публикаций, 6 авторских свидетельств на изобретения, соавтор 2 монографий. Область научных интересов: методы и средства построения систем речевого взаимодействия.