

УДК 621.39

А. Н. Ткаченко, к. т. н., доц.; Л. В. Крупельницкий, к. т. н., доц.; С. В. Хрущак**АДАПТИВНОЕ ОПРЕДЕЛЕНИЕ ГОЛОСОВОЙ АКТИВНОСТИ ПРИ
ДЕЛЬТА-СЖАТИИ РЕЧЕВЫХ СИГНАЛОВ**

Разработан метод дельта-сжатия с прогнозированием следующего индекса квантованного значения в кодовой книге и усовершенствован метод определения голосовой активности, описанный в рекомендации G.729, Annex B. Проведено экспериментальное исследование предложенных методов.

Ключевые слова: сжатие речевых сигналов, коэффициенты линейного прогнозирования, линейные спектральные частоты, векторное квантование, определение голосовой активности.

Актуальность

При передачи речи цифровыми каналами одним из наиболее важных вопросов является объем информации, который необходимо передавать за единицу времени для обеспечения качественной голосовой связи. Сжатие речи уменьшает объемы данных, которые передаются, и аппаратные затраты.

В данной статье рассмотрен процесс сжатия речевых сигналов с помощью вокодера LPC и предложен ряд подходов для дальнейшего уменьшения объемов данных.

Введение

подавляющее большинство алгоритмов сжатия речи используют модель линейного прогнозирования для описания и кодирования сигнала, согласно которой спектральная информация о сигнале описывается 10-ю коэффициентами линейного прогнозирования (LPC). Для более эффективного квантования и дальнейшей интерполяции коэффициенты LPC, как правило, превращают в линейные спектральные пары (LSP) и квантуют с помощью кодовых книг.

Коэффициенты LSP, полученные в результате кодирования речи по методу линейного прогнозирования, имеют зависимости между элементами как внутри фрейма, так и между соседними фреймами, что свидетельствует о высоком уровне избыточности при использовании данных коэффициентов [1].

Учет корреляции между коэффициентами позволяет уменьшить объем данных, необходимых для описания параметров речевого сигнала, и таким образом уменьшить требования к пропускной способности канала передачи информации. В данном исследовании рассматривается параметрическая система сжатия речи, которая учитывает корреляцию между коэффициентами внутри фрейма сигнала и между последовательными фреймами и таким образом позволяет достичь большей степени сжатия.

Дальнейшим этапом уменьшения объемов данных, которые необходимо передавать в канал связи, является использование дискретной передачи (Discontinuous transmission – DTX) – отключение передатчика или передача комфортного шума (Comfort noise – CN) на приемочную сторону во время пауз в разговоре. Такой подход, например, используется в сети GSM. Основной задачей при этом является определение голосовой активности (Voice Activity Detection – VAD). Большинство подходов определения голосовой активности базируются на энергетических порогах, определении периода основного тона, спектральном анализе, частоте переходов через ноль или комбинации нескольких методов. Сложность задачи состоит в том, что постоянной высокой точности определения вокализованности фрейма сигнала невозможно достичь, поскольку большинство алгоритмов базируется на

определенном пороговом значении, которое является фиксированным или вычисляется из сигнала на шумовых сегментах. Так, например, при использовании для VAD метода наименьших квадратов шумовые сегменты используются в фильтре с конечной импульсной характеристикой для его тренировки. Но в случаях, когда фоновый шум не является стационарным или когда речевой сигнал является преимущественно вокализированным, для достижения высокой точности алгоритма VAD необходимо постоянно корректировать пороговое значение независимо от типа фрейма. Для этого необходим метод, который мог бы эффективно адаптироваться к изменению фонового шума.

В данной работе усовершенствован метод определения голосовой активности, предложенный международной ассоциацией ITU-T – G.729 Annex B [3], который разрабатывался для мультимедиа и IP-телефонии. Данный метод обеспечивает адаптацию к высокому уровню фонового шума и использует данные, которые получают в процессе кодирования, что уменьшает его вычислительную сложность. Но, несмотря на это, он имеет низкое быстродействие по сравнению с другими стандартами (ETSI AMR, статистические методы), особенно при низком соотношении сигнал/шум. Поэтому предложены некоторые изменения в данном методе для улучшения его быстродействия и проведено его исследование с использованием предложенных изменений.

Учет корреляции между коэффициентами

Учет корреляции между коэффициентами одного фрейма происходит за счет использования векторного квантования коэффициентов LSP. При векторном квантовании набор коэффициентов LSP рассматривается как один вектор, для которого происходит поиск ближайшего квантованного значения в специальной таблице – векторной кодовой книге [4].

Для учета корреляции между последовательными фреймами в [4] предложен метод упорядочивания векторов согласно отношению мажоризации, что позволяет осуществить переход к передаче разности между индексами (дельты) последовательных фреймов. Недостатком предложенного метода является значительная задержка (100 мс), необходимая для достижения качественного сигнала, при уменьшении объема информации для описания параметров речевого сигнала с 24 на 20 бит, что недопустимо в некоторых системах связи. Учитывая это, предложен метод дельта-сжатия с прогнозированием следующего отсчета, позволяющий получить меньшую задержку (20 мс), при незначительно меньшей точности передачи индекса.

Для уменьшения окна дельты между индексами соседних фреймов при довольно малой задержке, по сравнению с методом предложенным в [5], предлагается прогнозирование индекса фрейма на основе предыдущих значений. Прогнозирование происходит путем создания экстраполирующей функции для индексов нескольких фреймов на приемочной и передающей стороне. При сжатии на приемочную сторону также передается дельта между прогнозируемым и реальным значениями. Иллюстрация предложенного метода представлена на рис. 1.

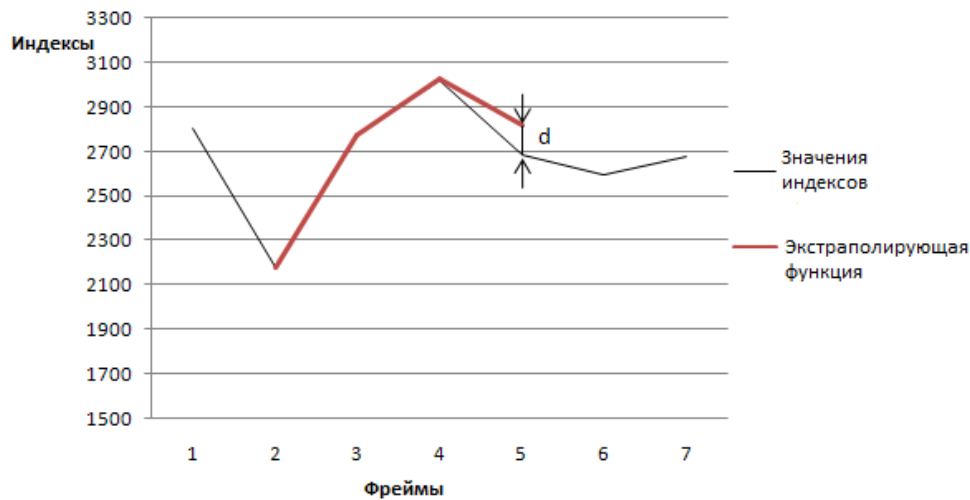


Рис. 1. Пример работы метода сжатия с экстраполяцией

Экстраполирующая функция строится по методу наименьших квадратов. Суть данного метода заключается в том, что функция для описания экспериментальных данных g строится как линейная комбинация M базисных функций F_j [6]:

$$g = \sum_{j=0}^{M-1} c_j F_j(x). \quad (1)$$

При этом коэффициенты c_j выбираются таким образом, чтобы минимизировать критерий $C_{\text{мнк}}$ – сумму квадратов отклонений экстраполирующей функции от экспериментальных значений:

$$C_{\text{мнк}} = \sum_{i=1}^n (y_i - g_i)^2 \rightarrow \min C, \quad (2)$$

где y_i – экспериментальная функция.

Исследовались также альтернативные методы – метод наименьших модулей (МНМ) и метод Чебышева. Для МНМ критерием минимизации является сумма отклонения модулей (3):

$$C_{\text{мнм}} = \sum_{i=1}^n |y_i - g_i|. \quad (3)$$

Для метода Чебышева критерием минимизации является максимальное отклонение [6]:

$$C_{\text{мч}} = \max |y_i - g_i|. \quad (4)$$

Однако только метод наименьших квадратов позволяет найти наилучшие c_j за конечное число операций, сводя задачу к решению системы линейных уравнений. Это означает, что он является наиболее простым для вычисления, к тому же он точнее описывает экспериментальную функцию.

Введение задержки позволяет корректировать значение индексов, которые передаются; это необходимо в ситуациях, когда прогноз окажется неточным (рис. 2). Например, в данном случае мы должны спрогнозировать индекс 5-го фрейма на основе предыдущих – 2-го, 3-го и 4-го. Однако прогнозирование только ухудшило результат – значение дельты возросло.

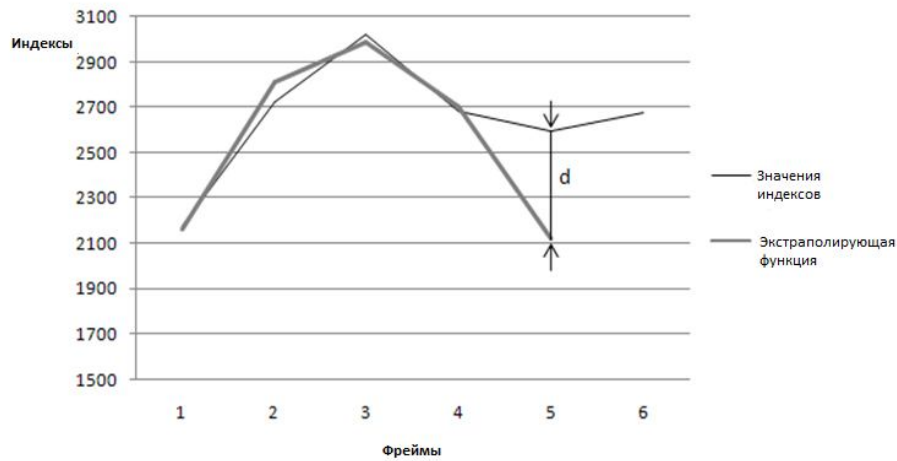


Рис. 2. Пример неточного прогноза

В случаях, когда при передачи параметров i -го фрейма выполняется неравенство (5):

$$\Delta_{i+1} > S_w \quad (5)$$

где Δ_{i+1} – значение дельты для $i+1$ фрейма, S_w – размер окна, применяется такой алгоритм:

1. Осуществляется аппроксимация значений экспериментальной функции, учитывая следующие значения индексов $i+1, i+2, \dots, i+k$, где k – величина задержки в фреймах. Для аппроксимации также используется метод наименьших квадратов.

2. Осуществляется коррекция значения i -го индекса согласно аппроксимирующей функции, чтобы получить наилучший прогноз для следующего фрейма.

3. Осуществляется прогноз $i+1$ фрейма на основе предыдущих фреймов $i, i-1, i-2, \dots, i-p$, где p – количество фреймов для создания прогноза, которое определяется порядком модели прогнозирования.

4. Если различие между прогнозируемым и реальным значениям выходит за пределы окна, то коррекция проводится снова: происходит переход к п. 2. В предельном случае, когда при максимально допустимом отклонении i -го фрейма прогноз является неточным, то происходит переход к п. 5.

5. Различие между скорректированным значением и прогнозируемым, учитывая значения индексов для $i-1, i-2, \dots, i-(p+1)$ фреймов, передается в канал связи.

Результат работы данного алгоритма можно видеть на рис. 3. После коррекции значения индекса для 4-го фрейма прогноз происходит успешно – значение дельты остается незначительным.

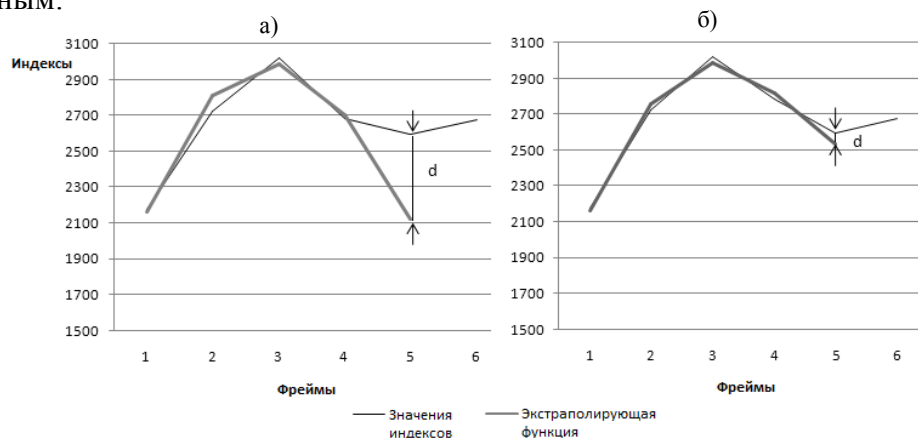


Рис. 3. Графики экспериментального и экстраполированного значения индексов:

- а) до выполнения коррекции;
- б) после выполнения коррекции

Улучшение алгоритма определения голосовой активности

Рекомендация ITU-T G.729 Annex B разрабатывалась как дополнение к вокодеру G.729. Решения о вокализованности принимаются для фреймов продолжительностью 10 мс на основе таких характеристик:

1. Дифференциальная энергия сигнала для всего диапазона спектра – ΔE_f ;
2. Дифференциальная энергия сигнала для низкочастотного спектра (0 – 1 кГц) – ΔE_l ;
3. Дифференциальная частота переходов через ноль – ΔZCR ;
4. Спектральное искажение – ΔLSF .

Схема данного VAD приведена на рис. 4.

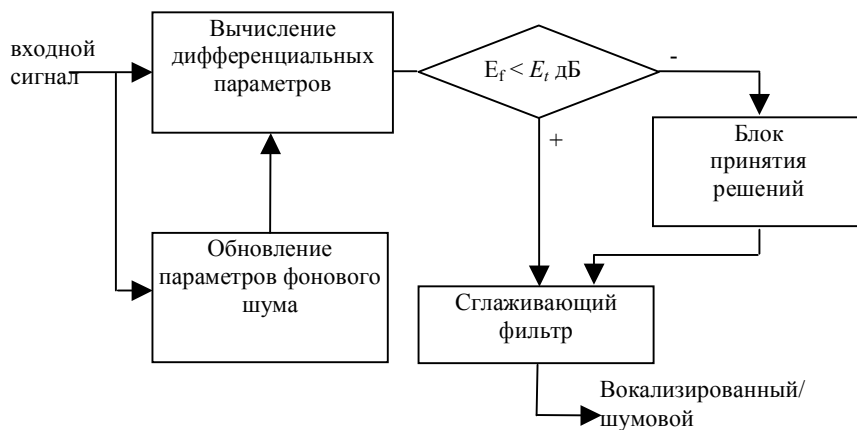


Рис. 4. Блок-схема алгоритма ITU-T G.729 Annex B

Входные параметры получают из речевого сигнала, используя авторегрессионную модель первого порядка, если выполняется условие $\Delta E_f < E_t$, где E_t – заданное пороговое значение. Однако при расчёте данных параметров возникают дополнительные сложности. А использование частоты переходов через ноль и энергетического порога вносит погрешности при высокой зашумленности сигнала.

Для улучшения качества предложенного метода предлагается ускорить оценку спектральной обводящей за счет сглаживания с помощью МНК. За счет этого значительно легче вычислить пологость спектра, которая является ключевой при определении вокализованности фрейма сигнала (для невокализованного фрагмента – спектральная обводящая будет более пологой). Пример такого сглаживания спектральной обводящей изображен на рис. 5.

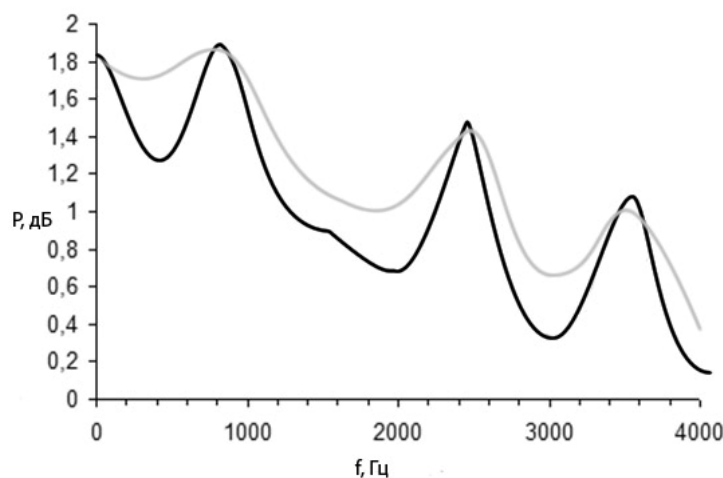


Рис. 5. Пример сглаживания спектральной обводящей сигнала

Таким образом, это позволяет уменьшить количество вычислений, поскольку пологость спектра считается только по нескольким точкам, и повысить точность алгоритма, особенно в низкочастотной полосе сигнала.

Для вычисления оптимального порогового значения вокализованности сигнала применяется новый подход – геометрически адаптивный энергетический порог, предложенный Езером и Таньером [7]. Данный подход базируется на использовании вероятностного распределения амплитуд и позволяет устанавливать пороговое значение адаптивно, используя не только шумовые сигналы, но и вокализованные. Данный метод стойкий к нестационарному шуму, но иногда может давать неправильный результат при резких и коротких скачках шума. Благодаря простоте вычисления адаптивного энергетического порога также уменьшается сложность алгоритма VAD.

Исследование метода дельта-сжатия

Тестирование метода ускоренного поиска векторов происходило на основе фонетического материала, составленного из двух фонетически полных текстов [4]. При использовании метода дельта-сжатия с прогнозированием значения индекса в кодовой книге для следующего фрейма прогнозирование происходит путем создания экстраполирующей функции для индексов нескольких фреймов. Исследовались несколько методов построения экстраполирующей функции, а именно: метод наименьших квадратов (МНК), метод Чебышева и метод наименьших модулей (МНМ).

В качестве критериев оценивания методов использовались дисперсия распределения вероятности дельты D между последовательными фреймами и спектральное искажение SD . Результаты оценивания приведены в табл. 1.

Таблица 1

Результаты исследования методов экстраполяции для кодовой книги размером 4096 векторов

Название метода	D	Среднее SD , дБ	Процент фреймов, для которых $SD > 2$ дБ, %	Процент фреймов, для которых $SD > 4$ дБ, %
МНК	21531	0,882	1,78	0
Метод Чебышева	31095	0,912	1,96	0
МНМ	141950	0,956	2,15	0

Как можно видеть из таблицы 1, наилучшим по показателям дисперсии конечного распределения и спектральным искажением является МНК. Для МНМ количество фреймов со спектральным искажением в пределах 2 – 4 дБ превышает допустимое значение в 2%, принятое для систем низкоскоростного сжатия речи.

Исследование метода определения голосовой активности

Общепринятой практикой оценивания качества методов VAD является использование одного из двух видов тестов – объективных или субъективных. К субъективным относят методы, у которых работа алгоритма определяется при помощи оценок слушателей. Однако такая схема не позволяет определить количество неточных срабатываний при высоком уровне шума, поскольку это будет выражаться в увеличении количества данных в канале, а не качестве результирующей речи, поэтому был избран объективный метод оценивания VAD. Для тестов была избрана речевая база TIMIT, в запись которой добавлен случайный шум, далее запись была разделена на фреймы, которые потом были размечены вручную на вокализованные и шумовые, то есть фактически сравнение проводилось с идеальным вариантом. Данные файлы обрабатывались с помощью разных алгоритмов определения активности – предложенного, оригинального, ETSI AMR 1 и ETSI AMR 2, после чего

подсчитывалось количество фрагментов, на которых было обрезание речи и на которых шум был распознан как речь. Результаты тестирования приведены в таблице 2.

Таблица 2

Результаты тестирования качества

Название метода	Фреймов с обрезанием речи, %	Шумовых фреймов распознанных как вокализированные, %
Улучшенный ITU-T G.729B	83%	2%
ITU-T G.729B	68%	5%
AMR 2	87%	0,3%
AMR 1	76%	5,5%

Результаты тестирования показали, что улучшенный метод ITU-T G.729B позволяет получить лучшее качество распознавания речи на зашумленных фреймах, чем обычный метод, но худшее, чем достигается в методе с использованием статистики высокого порядка, который, однако, является более вычислительно сложным. Если учитывать, что в среднем при двухсторонней беседе каждый из собеседников разговаривает около 40% времени [2], то данный метод может уменьшить нагрузку на канал связи до 60%, хотя в реальных условиях цифра не превышает 45%.

Для исследования быстродействия предложенного улучшенного метода определения голосовой активности использовались материалы из набора TIMIT, поскольку это позволяет сравнивать полученные результаты с зарубежными аналогами. Результаты сравнения улучшенного метода ITU-T G.729 Annex B и существующих методов для разного размера анализируемых блоков приведены в таблице 3.

Как видно из таблицы, усовершенствованный метод работает быстрее, чем оригинальный, и быстрее, чем метод, базирующийся на статистике высокого порядка. Метод с использованием адаптивных энергетических уровней наиболее быстрый, но и наименее точный.

Таблица 3

Среднее количество операций для предложенного и известных методов при разном размере анализируемого блока

Название метода	Количество математических операций для заданного размера анализируемого блока, мс			
	10	20	50	100
Улучшенный ITU-T G.729B	1,7·10 ⁴	2,4·10 ⁴	5,2·10 ⁴	14,3·10 ⁴
ITU-T G.729B	2,2·10 ⁴	2,7·10 ⁴	6,1·10 ⁴	18,6·10 ⁴
AMR 2	2,9·10 ⁴	2,6·10 ⁴	4,8·10 ⁴	9,8·10 ⁴
AMR 1	1,6·10 ⁴	2,0·10 ⁴	3,9·10 ⁴	7,2·10 ⁴

Выводы

Исследование метода дельта-сжатия с прогнозированием показало, что его использование позволяет уменьшить количество бит для передачи речевых параметров с 24 до 20 при задержке в 1 фрейм и незначительном ($SD = 0,882$ дБ) искажении сигнала. Применение дискретной передачи пакетов позволяет дополнительно уменьшить нагрузку на канал на 45%. Предложенный усовершенствованный метод ITU-T G.729B позволяет получить лучшее качество распознавания зашумленных фреймов, чем обычный, и меньшую вычислительную сложность, чем метод ETSI AMR 2.

СПИСОК ЛИТЕРАТУРЫ

1. Chu W. C. Speech Coding Algorithms – Foundation and Evolution of Standardized Coders. / W. C. Chu. New Jersey: Wiley. – 2003. – 553 p.
2. Ramirez J. Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, Robust Speech Recognition and Understanding. / J. Ramirez, J. M. Gorriz and J. C. Segura. Vienna: I-Tech Education and Publishing. – 2007. – 460 p.
3. Benyassine A. ITU-T recommendation G.729 annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data application. / A. Benyassine, E. Shlomot, and H.-Y. Su // IEEE Communications. – 1997. – № 35. – P. 64 – 73.
4. Ткаченко О. М. LSF-вокодер на основі векторного квантування / О. М. Ткаченко, Н. О. Біліченко, О. Д. Феферман, С. В. Хрущак // Реєстрація, зберігання і обробка даних. – 2007. – № 1. – С. 35 – 41.
5. Ткаченко О. М. Метод дельта-ущільнення мовних сигналів / О. М. Ткаченко, О. Д. Феферман, С. В. Хрущак // Інформаційні технології та комп'ютерна інженерія. – 2008. – № 1(11). – С. 8 – 13.
6. Коваленко И.Н. Теория вероятностей и математическая статистика: Учебное пособие. / И. Н. Коваленко, А. А. Филиппова. – М.: Высшая школа, 1982. – 256 с.
7. Özer H. A geometric algorithm for voice activity detection in nonstationary Gaussian noise / H. Özer and S. G. Tanyer // EUSIPCO, – 1998. – № 1. – P. 23 – 26.

Ткаченко Александр Николаевич – доцент кафедры вычислительной техники.

Крупельницький Леонид Витальевич – доцент кафедры вычислительной техники.

Хрущак Сергей Викторович — аспирант кафедры вычислительной техники.
Винницкий национальный технический университет.