УДК338.984:519.6

В. А. Нестеров – магистрант кафедры моделирования вычислительных и электронных систем
О. О. Жаринов (канд. техн. наук, доц.) – научный руководитель

АНАЛИЗ МЕТОДОВ СЖАТИЯ РЕЧЕВОГО СИГНАЛА

Основные объемы передаваемой информации на сегодня приходятся на речь – это и проводная телефония, и системы сотовой и спутниковой связи, и т.д. Поэтому эффективному кодированию, или сжатию речи в системах связи уделяется исключительное внимание.

Стоит отметить, что речь обладает такими существенными отличиями от обобщенного звукового сигнала, как более узкая полоса частот (около 4 кГц) и наличие значительного числа неинформативных пауз в сигнале, что позволило создать ряд алгоритмов сжатия ориентированных только на речевые сигналы. На сегодняшний день, наибольшее распространение алгоритмы сжатия речи с потерями получили в системах IP-телефонии. VoIP (Voice-over-IP) – IP-телефония – система связи, при которой аналоговый звуковой сигнал от одного абонента дискретизируется (кодируется в цифровой) вид, компрессируется и пересылается по цифровым каналам связи до второго абонента, где производится обратная операция – декомпрессия, декодирование и воспроизведение аналогового сигнала.

Источником информационных данных является речевой сигнал, возможной моделью которого является нестационарный случайный процесс. В первом приближении можно выделить следующие типы сигнальных фрагментов: вокализированные, невокализированные, переходные и паузы. При передаче речи в цифровой форме каждый тип сигнала при одной и той же длительности и одинаковом качестве требует различного числа бит для кодирования и передачи. Следовательно, скорость передачи разных типов сигнала также может быть различной, что обусловливает применение кодеков с переменной скоростью.

На сегодняшний день применяемые методы сжатия речевых сигналов, можно условно разделить на 3 класса: кодеры формы сигнала, вокодеры и гибридные кодеры, которые сочетают в себе достоинства двух предыдущих классов.

Кодеры формы сигнала являются простейшими кодерами речи, вообще не использующими информацию о том как был сформирован сигнал, а просто старающиеся максимально приблизить декодированный сигнал по форме к оригиналу. Теоретически они не зависят от характера сигнала, подаваемого на вход, и могут использоваться для кодирования любых, в том числе и не речевых сигналов.

Самым простым способом кодирования формы сигнала является импульсно-кодовая модуляция – ИКМ (или PCM – Pulse Code Modulation), при использовании которой производится просто дискретизация и равномерное квантование входного сигнала, а далее преобразование полученного результата в равномерный двоичный код.При этом $f_{\max} = 3.4 кГц$; частота дискретизации $f_д = 8 кГц$. После равномерного квантования при числе уровней $L = 2^{12}$ и предварительного кодирования производится цифровая компрессия, в результате чего длина кодовой комбинации уменьшается до $n = 8$ разрядов. Результатом преобразования является двоичная последовательность, передаваемая со скоростью 64 кбит/с.

Типичным вокдером является стандарт транкинговой радиосвязи APCO 25, описывающий структуру цифровой транкинговой системы и некоторые ее интерфейсы. Для цифровой передачи речи стандарт APCO 25 предусматривает использование кодера IMBE (Improved MultiBand Excitation, модифицированный метод многополосного возбуждения). Кодер формирует цифровой поток со скоростью 4,4 кбит/с. Для исправления ошибок в цифровом речевом сигнале используется избыточное кодирование, порождающее дополнительный цифровой поток со скоростью 2,8 кбит/с. Цифровой речевой сигнал передается кадрами длительностью 180 мс. Два речевых кадра образует суперкадр длительностью 360 мс.

Речевой IMBE-кодер основан на модели речи, которая относится к моделям с многополосным возбуждением (MBE). Основная идея работы кодера состоит в разделении цифрового речевого входного сигнала на перекрывающиеся речевые сегменты (или фреймы) с использованием окна Кайзера.

Затем для определенного фрейма оценивается набор параметров. Блок-схема алгоритма анализа показана на рис 1.

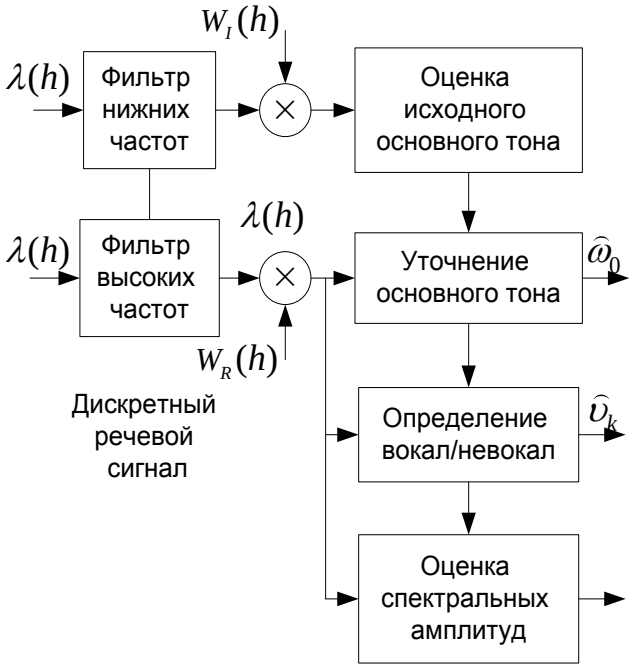

Рис 1.

В речевом IMBE-кодере параметры возбуждения и огибающей спектра оцениваются одновременно так, что синтезированный спектр является самым близким к исходному речевому спектру. Параметры MBE модели речи, которые должны быть оценены для каждого речевого фрейма следующие:

- период основного тона (или основная частота);
- решение вокал/невокал;
- спектральные амплитуды, характеризующие огибающую спектра.

В декодере вокализированная и невокализированная компоненты синтезируются отдельно и на заключительной стадии объединяются для получения полного речевого сигнала. Алгоритмы, которые используются для синтеза вокализированных и невокализированных частей речи, основаны на двух различных способах.

Невокализованная часть речи генерируется из гармоник, которые объявлены невокализованными. Для каждого фрейма речи блок случайного шума взвешивается и преобразуется с помощью быстрого преобразования Фурье. Области спектра, которые соответствуют вокализованным гармоникам, принимаются равными нулю.

Так как вокализованная речь моделируется ее индивидуальными гармониками в частотной области, на стороне декодера она восстанавливается как совокупный сигнал регулируемых генераторов. Каждой гармонике вокализованной области фрейма поставлен в соответствие генератор, который характеризуется частотой и фазой. Однако из-за того, что вокализованная часть речи не является периодической на интервалах, состоящих нескольких фреймов анализа, отклонения от ожидаемых параметров соседних фреймов могут вызвать скачки по концам фреймов, что приведет к значительному ухудшению качества речи. Для разрешения этой проблемы во время синтеза проверяются параметры текущего и предыдущего фреймов для уверенности, что на границе фреймов происходит плавный переход. Это делается для того, чтобы на границах фреймов вокализированная речь была непрерывной. Для обеспечения непрерывности в начале и конце фрейма речи функция амплитуды линейно интерполируется между значениями оценок для текущего и предыдущего фреймов.

Синтез речи в IMBE-декодере требует информации об основной частоте, решении вокал/невокал, величине спектральных составляющих и фазе вокализованных гармоник. Так как фазы вокализованных гармоник можно предсказать, информация о фазе не передается между кодером и декодером. Основная частота (основой тон) квантуется с половинной точностью выборки во временной области, при-

чем возможный диапазон тона перекрывается восемью битами. Решение вокал/невокал является двоичным числом и не требует квантования.

Число полос, на которые разбивается речевой фрейм в частотной области, зависит от основного тона фрейма, но не превышает 12.

Таким образом, в кодере IMBE фрейм речи имеет длительность 20 мс, содержит 144 бита, из которых 56 используются для канального кодирования, 88 – для кодирования параметров речевой модели. Кодер работает на скорости 4,4 кбит/с. Скорость передачи в канале – 7,2 кбит/с.

Промежуточным звеном между кодерами формы сигнала, в которых сохраняется форма колебания речевого сигнала в процессе его дискретизации и квантования, и параметрическими вокодерами, основанными на процедурах оценки и кодирования небольшого числа параметров речи, объединяющим преимущества каждого из них, являются гибридные кодеки, к числу которых относится кодер CELP (Code Excited Linear Prediction). Метод кодирования CELP основан на линейной авторегрессионной модели процесса формирования и восприятия речи и входит в группу методов анализа через синтез, реализующих современные и эффективные алгоритмы информационного сжатия речевых сигналов.

Линейная авторегрессионная модель процесса формирования речевых сигналов с локально постоянными на интервалах 10. . .30 мс параметрами получила в настоящее время наибольшее распространение. Эта модель описывается уравнением:

$$\lambda(h) = \sum_{m=1}^{M} a(m)\lambda(h-m) + x(h),$$

где M – порядок модели; $\lambda(h)$ – последовательность отсчетов речевого сигнала; $a(m)$ – коэффициенты линейного предсказания, характеризующие свойства голосового тракта; $x(h)$ – порождающая последовательность или сигнал возбуждения голосового тракта.

Авторегрессионная модель речевого сигнала описывает его с достаточно высокой степенью точности и позволяет применять развитый математический аппарат линейного предсказания. При этом обеспечивается более высокое качество декодированной речи, устойчивость к входному акустическому шуму и ошибкам в канале связи по сравнению с системами с иными принципами кодирования.

В рамках данной модели наиболее перспективными методами кодирования считаются методы «анализа через синтез» с использованием многоимпульсного возбуждения. Новизна многоимпульсного возбуждения заключается в том, что в сигнале остатка линейного предсказания выбираются такие его значения, которые наиболее важны для повышения качества синтезированной речи. При этом используемая в процедуре анализа через синтез схема кодирования, помимо учета ошибок квантования, включает критерии субъективной оценки качества речевого сигнала, что обеспечивает естественное звучание синтезированной речи.

При многоимпульсном возбуждении сигнал остатка линейного предсказания представляется в виде последовательности импульсов с неравномерно распределенными интервалами и с различными амплитудами (около 8-10 импульсов за 10 мс). Амплитуды и положение этих импульсов определяются на покадровой основе (кадр за кадром). Основным преимуществом многоимпульсного возбуждения является то, что она определяется для любого речевого сегмента и при этом не требуется знаний ни о вокализованности данного сегмента, ни о периоде основного тона.

Методы анализа через синтез используют синтезатор (декодер) речевого сигнала как составную часть устройства кодирования. При этом задача анализа сводится к процедуре оценки передаваемых в канал связи параметров речи, проводимой в соответствии с некоторым критерием рассогласования между исходным и декодированным сигналами. Для учета специфики слухового восприятия в качестве критерия рассогласования обычно используется взвешенная по частоте квадратическая ошибка:

$$\varepsilon_{\omega} = \int_{0}^{F/2} \left| S(f) - S_q(f) \right|^2 w(f) df,$$

где $S(f)$ и $S_q(f)$ – преобразование Фурье исходного и синтезированного речевых сигналов; $w(f)$ – весовая функция. Принимая во внимание важность для восприятия речи не только формант, но и

межформантных областей, для алгоритмов анализа речи через синтез в качестве эталонной была предложена весовая функция следующего вида:

$$w(z) = A(z)A^{-1}(z/\gamma),$$

где $A^{-1}(z)$ – передаточная характеристика синтезирующего фильтра; γ – параметр, регулирующий энергию ошибки или шум квантования. Фактически при таком окне взвешивания подчеркивается ошибка в межформантных областях и тем самым обеспечивается более равномерное по частоте распределение отношения мощности полезного сигнала к мощности ошибки кодирования.

В алгоритмах кодирования с «анализом через синтез» повышение эффективности информационного уплотнения речевых сигналов производится, преимущественно, за счет сокращения избыточности последовательности *x(h)*, которая осуществляет возбуждение синтезирующего фильтра $A^{-1}(z)$ линейного предсказания, формирующего огибающую сигнала, с коэффициентом передачи

$$A^{-1}(z) = \left(1 - \sum_{m=1}^{M} a(m)z^{-m}\right)^{-1}$$

Для этой цели применяется также дополнительный фильтр с характеристикой $P^{-1}(z) = (1 - g_p z^{-T})^{-1}$ с коэффициентом предсказания g_p и задержкой на период основного тона *T*. Фильтр выполняет функции генератора квазипериодических колебаний голосовых связок при произношении вокализованных звуков.

Экспериментально установлено, что кодовое возбуждение обеспечивает наиболее высокое качество кодирования речевого сигнала, в том числе и при наличии входных акустических помех.

CELP наиболее эффективно применяется при передаче речевого сигала в диапазоне скоростей от 4 до 6 кбит/с.

Вычислительные эксперименты. Для более подробного анализа речевых кодеров были проведены вычислительные эксперименты, которые состояли из следующих этапов:

1) оцифровка и сохранение в формате .wav исходного речевого сигнала;
2) кодирование исходного сигнала при помощи наиболее современных алгоритмов сжатия речевых данных с потерями на типичных для данного алгоритма битрейтах;
3) оценка субъективного качества восстановленных сигналов группой экспертов по традиционной 5-ти бальной шкале, где наилучшему качеству звучания соответствует наибольший бал и вычисление средней оценки для алгоритма.

В таблице представлены наиболее типичные результаты экспериментов.

Результаты оценки субъективного качества восстановленных сигналов [1]

Кодек	Тип кодека	Битрейт, Кбит/с	Оценка
G.711	ИКМ	64	4,1
G.726	АДИКМ	32	3,85
G.728	LD – CELP	16	3,61
Skype	AMR-WB	14	4,1
G.729	CS – ACELP (без VAD)	8	3,92
G.729	2-х кратное кодирование	8	3,27
G.729	3-х кратное кодирование	8	2,68
G.729a	CS – ACELP	8	3,7
G.723.1	MP –MLQ	6,3	3,9
G.723.1	ACELP	5,3	3,65

Низкоскоростным кодекам свойственны определенные ухудшения параметров, влияющие на качество передачи речи, по сравнению со стандартным кодеком ИКМ. Важно, что эти ухудшения накапливаются при тандемном включении как однородных, так и разнородных низкоскоростных кодеков.

Следует отметить следующие основные факторы, влияющие на качество передачи речи при использовании кодеков:

- искажения квантования;
- временная задержка;
- амплитудно-частотные искажения;
- битовые ошибки;
- проскальзывания;
- потеря кадров;
- потеря пакетов.

Планирование речевых соединений требует обязательного учета ухудшений, вносимых каждым переходом А-Ц и Ц-А, и определения на этой основе допустимого количества таких переходов.

Библиографический список

1. [1] И.И.Чижов Т.Н. Созонова И. В. Деев «Об оптимизации процедур сжатия речевых данных» БГУ 2004.г
2. 4 стр.
3. [2] Шульгин В.И. «Основы теории связи» Харьков, Уч. Пособ/ХАИ 2005г. 194 стр.
4. [3] Безрук В.М., Скорик Ю.В., 2009 «Методология выбора речевых кодеков с учетом совокупности показателей качества на основе метода анализа иерархий» Сб. докл/ХНУР 2009 129 стр.
5. [4] Феннерман М. «Вопросы качественной передачи голоса по IP-сетям: Сжатие, задержка и эхо. Часть 1» электронные компоненты №11 2008 стр. 83-85