УДК 621.371.39:681.322.01

АНАЛИЗ И СИНТЕЗ УСТРОЙСТВ КОДИРОВАНИЯ РЕЧЕВОГО СИГНАЛА НА ОСНОВЕ АНТРОПОМОРФИЧЕСКОЙ ОБРАБОТКИ И СИНУСОИДАЛЬНЫХ МОДЕЛЕЙ

Д.С. ЛИХАЧЕВ, А.А. ПЕТРОВСКИЙ

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Поступила в редакцию 16 июня 2006

В данной работе предлагается метод построения синусоидального вокодера с использованием антропоморфического анализа речи. Особенностью данной системы является то, что речь как на вокализованных, так и на невокализованных участках представляется в виде ограниченного числа так называемых "доминирующих" синусоидальных компонент.

Ключевые слова: синусоидальный вокодер, антропоморфическая обработка.

Введение

Интенсивное развитие информационных технологий в современном мире делает особенно важным решение проблемы быстрой и качественной передачи различного рода информации по цифровым каналам связи. Несмотря на разнообразие применяемых для этого средств, основным видом коммуникации между людьми остается передача информации посредством речи. В связи с этим в настоящее время продолжают интенсивно развиваться и совершенствоваться методы цифровой обработки и передачи речи. Использование цифрового представления данных позволяет обеспечить надежность и экономичность связи, возможность гарантированной защиты от несанкционированного доступа [1]. При этом большое значение приобретает решение проблемы минимизации числа бит, необходимых для передачи сигнала, т.е. проблема компрессии и кодирования речи. Актуальной задачей обработки речи становится создание систем низкоскоростной передачи с высоким качеством восприятия сигнала, способных функционировать в реальных условиях окружающей среды [2, 3].

Для высококачественного кодирования речи при скоростях передачи выше 8 кбит/с разработаны достаточно хорошие методы, основанные на линейном предсказании и кодировании во временной области. В настоящее время ведутся разработки вокодеров, которые обеспечили бы коммерческое качество восстанавливаемой речи при скоростях около 2 кбит/с. Для этого диапазона кодирование речи в частотной области [4] с применением техники синусоидального анализа [5] и использованием различных моделей слуха человека [6–10] показывает потенциально большие возможности.

Все методы обработки сигналов, основанные на синусоидальном представлении, базируются на предположении о том, что любой сигнал можно представить в виде суммы синусоидальных компонент с изменяющимися во времени параметрами: амплитудами, частотами и фазами [5]. Основными недостатками синусоидальных систем является то, что для достижения хорошего качества восстанавливаемой речи речевой сигнал должен быть представлен достаточно большим количеством речевых параметров, что непригодно для кодирования с низкими скоростями. При ограничении количества синусоидальных компонент резко деградирует каче-

ство речи. Для решения этой проблемы в настоящее время предложен ряд речевых синусоидальных кодеров, где тем или иным способом ограничивается объем кодируемой информации [10–15]. Наиболее перспективными в этой области являются синусоидальные кодеры с антропоморфической обработкой речевого сигнала, в которых объем кодируемой речевой информации уменьшается за счет применения различных слуховых моделей человека и психоакустических принципов [9]. Типичную схему анализа речи в таких системах представлена на рис. 1.

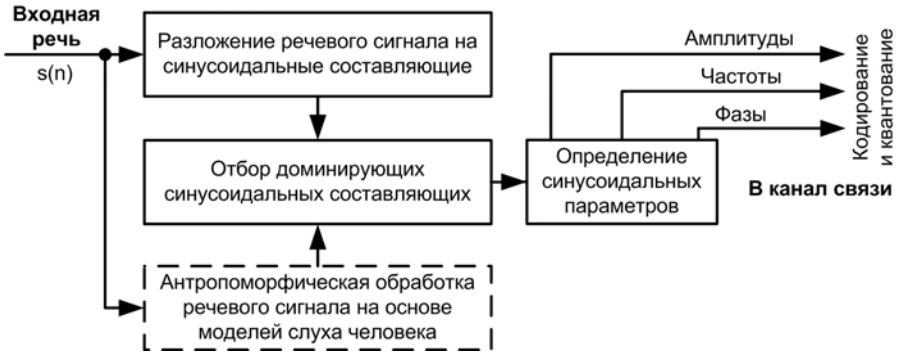

Рис. 1. Схема синусоидального анализа с антропоморфической обработкой

В данной работе предлагается метод компрессии речевого сигнала на основе синусоидальной модели с антропоморфической обработкой. Особенностью данного подхода является применение моделей слуха человека для отбора наиболее важной информации.

Система кодирования речевого сигнала на основе синусоидальной модели с антропоморфической обработкой

Структура предлагаемого синусоидального кодера речи с антропоморфической обработкой речевого сигнала представлена на рис. 2. Входной оцифрованный речевой сигнал анализируется с помощью спектрального анализа на основе преобразования Фурье, совмещенного с антропоморфической обработкой. В процессе анализа в исходном фрагменте речевого сигнала с помощью моделей слуха человека [16–21] выделяется несколько наиболее "важных" для человеческого слуха доминирующих синусоидальных компонент (СК), для каждой из которых в дальнейшем определяется амплитуда, частота и фаза.

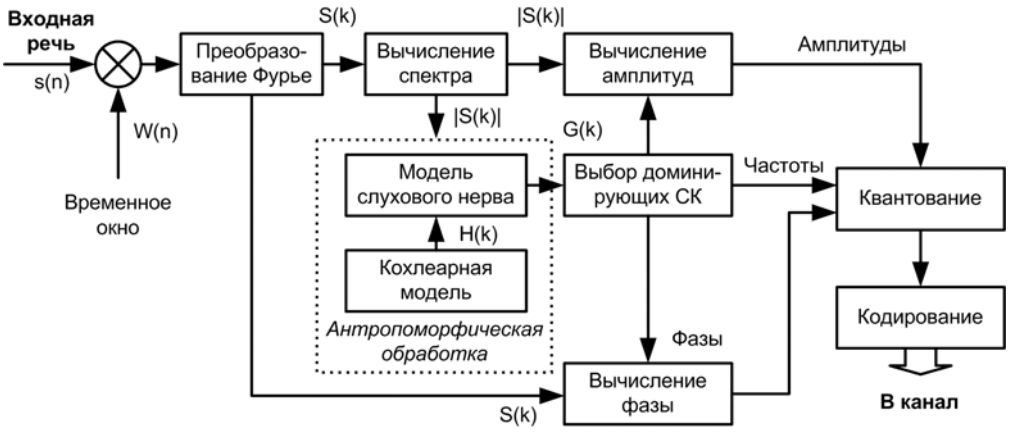

Рис. 2. Структура кодера

Для передачи по линии связи синусоидальные параметры соответствующим образом квантуются и кодируются [22]. На стороне декодера принятые параметры деквантуются и декодируются, а сам процесс синтеза речи сводится к суммированию сгенерированных синусоидальных компонент с найденными в процессе анализа амплитудами, фазами и частотами. При этом для получения в процессе синтеза приемлемого качества речи применяется частотное упорядочивание синусоид и интерполяция их параметров от фрейма к фрейму [16].

Задача выделения доминирующих частотных составляющих

Применение принципа "антропоморфической обработки сигнала" предполагает использование таких устройств и алгоритмов обработки информации, когда вычислительный процесс организовывается по "образу и подобию" человека, т.е. применяемые способы и алгоритмы моделируют какие-либо процессы, происходящие в его слуховых и речеобразующих системах. При этом предполагается, что при достаточно точном моделировании используемые системы будут иметь те же полезные свойства, что и их физиологический аналог.

Слуховая система человека представляет собой достаточно сложную систему с множеством составных частей. При моделировании такой сложной системы возникает проблема дифференциации ее элементов по значимости с точки зрения механизма обработки звуковой информации и применимости для задач компрессии. Моделирование "второстепенных" элементов не только сильно увеличит алгоритмическую и вычислительную сложность, но и усложнит процесс интерпретации полученных результатов.

При этом вне зависимости от применяемых алгоритмов модель слуховой системы человека обычно делят на две взаимозависимые части: модель периферической части (внешнее, среднее и внутреннее ухо) и модель слухового нерва (частично волосковые клетки, непосредственно слуховой нерв и участки головного мозга, отвечающие за обработку импульсов от слухового анализатора). При моделировании внутреннего уха (улитки) математически описывается движение базилярной мембраны под воздействием акустических колебаний и тем или иным способом выполняется спектральное разложение, которое является основной функцией примыкающих к базилярной мембране внутренних волосковых клеток. Здесь важно отметить, что именно от степени соответствия реального и смоделированного процесса обработки на уровне слухового нерва сильно зависит адекватность всей модели слуха.

В предлагаемой системе компрессии в качестве основной задачи для антропоморфической обработки ставится выработка такого критерия, который позволял бы отбирать наиболее важные для восприятия человеком "доминирующие" частотные компоненты. В этом случае при использовании синусоидальной модели речь будет синтезироваться как сумма синусоид с параметрами, которые определяются на "доминирующих" частотах.

Для успешного решения поставленной выше задачи необходимо не просто смоделировать процессы, происходящие в периферической части слуховой системы человека, но проанализировать информацию, которая циркулирует и на уровне слухового нерва. Также следует учитывать, что результаты работы слуховых моделей нужно представить в таком виде, который позволял бы их эффективно использовать совместно с синусоидальной моделью.

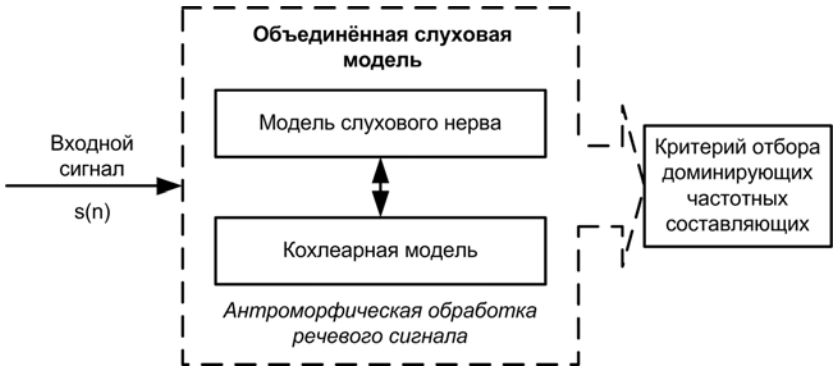

Рис. 3. Обобщенная схема выделения доминирующих частотных составляющих на основе объединенной слуховой модели

С этой целью предлагается объединить две модели слуха человека: кохлеарную модель, которая описывает функционирование улитки уха человека, и модель слухового нерва (рис. 3).

Кохлеарная модель

В качестве модели периферической части слуховой системы человека предлагается использовать так называемую SDCM-модель – Second order Difference Cochlea Model (разностная кохлеарная модель второго порядка) [19]. Согласно данной модели функционирование улитки

уха описывается работой банка цифровых фильтров второго порядка с высокой степенью перекрытия полос пропускания:

$$y_k(n) + b_{1k} y_k(n-1) + b_{2k} y_k(n-2) = A_k a_{0k} [u_s(n) - u_s(n-2)], \quad (1)$$

где $y_k(n)$ — перемещение или так называемая пучность базилярной мембраны в позиции x_k; b_{1k}, b_{2k}, A_k и a_{0k} — параметры, определяемые физическими свойствами базилярной мембраны в позиции x_k; $u_s(n)$ — входной синусоидальный сигнал, характеризующий скорость перемещения стремечка.

Соответствующую (1) передаточную функцию модели улитки в дискретном пространстве и времени можно записать в следующем виде:

$$H_k(z) = A_k \frac{a_{0k}\left(1 - z^{-2}\right)}{1 + b_{1k} z^{-1} + b_{2k} z^{-2}}.$$

Модель слухового нерва

Хорошей степенью адекватности реальным физиологическим процессам обладает модель слуха человека, представленная в работах [6, 10]. Результатом работы модели является так называемая слуховая гистограмма *G*(*f*,*t*), которая позволяет получить представление об акустической информации, циркулирующей на уровне слухового нерва. С ее помощью можно дифференцировать частотные составляющие анализируемого речевого сигнала по степени их "важности" для человеческого слуха. Однако непосредственное применение этой модели в данном случае затруднительно из-за ее плохой "совместимости" с синусоидальной моделью. Кроме того, она обладает достаточно высокой вычислительной сложностью, что также усложняет обработку речевого сигнала на ее основе в реальном масштабе времени.

Поэтому в данном случае предлагается таким образом модифицировать процесс вычисления гистограммы G (*t*,*f*), чтобы устранить вышеперечисленные недостатки и в то же время сохранить ее полезные свойства. Одним из возможных решений является перенос большинства действий из временной области в частотную, что позволит не только резко снизить вычислительную сложность алгоритма анализа, но и даст возможность корректно совместить его с применяемым для определения синусоидальных параметров спектральным анализом.

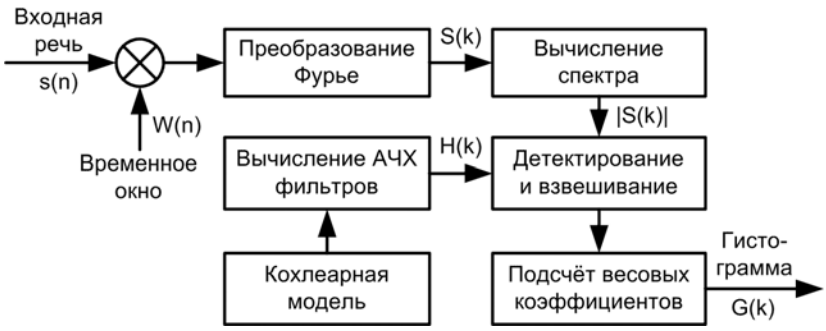

Рис. 4. Вычисление модифицированной слуховой гистограммы

Применяя обработку входного речевого сигнала с помощью кохлеарных фильтров в частотной области, слуховую гистограмму G(t,f) можно представить в виде дискретной функции частоты *G*(*k*) (рис. 4).

Гистограмма *G*(*k*) вычисляется с помощью следующего выражения:

$$G(k) = \sum_{m=1}^{M} G_m(k),$$

где *m* — номер обрабатываемого в текущий момент времени кохлеарного канала; M — число кохлеарных фильтров; $G_m(k)$ — *k*-й элемент гистограммы для m-го кохлеарного фильтра, он может быть вычислен по формуле:

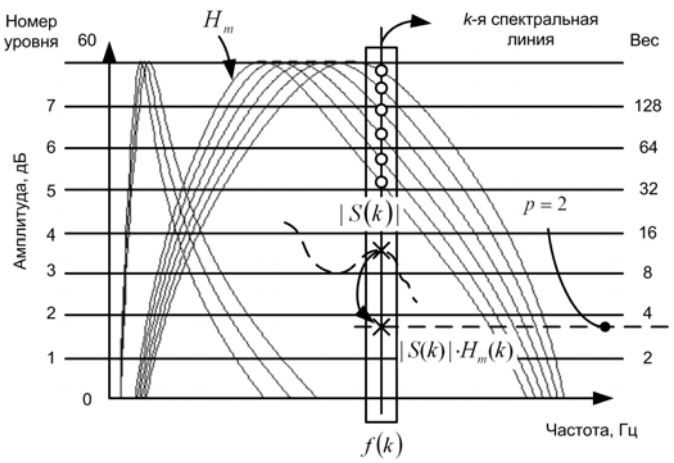

Рис. 5. Вычисление *k*-го элемента гистограммы

$$G_m(k) = |S(k)| H_m(k) 2^p ,$$

где p — номер уровня; $H_m(k)$ амплитудно-частотная характеристика m-го кохлеарного фильтра. Пример вычисления одного элемента гистограммы показан на рис. 5.

Из рис. 5 видно, что номер уровня p выбирается в зависимости от величины произведения $|S(k)| H_m(k)$. Все значения уровней логарифмически распределены по всему амплитудному диапазону спектра сигнала.

Модифицированная модель слуха на основе преобразования Фурье с неравным интервалом дискретизации по оси частот

Как было описано выше, в ходе антропоморфического анализа спектр сигнала определенным образом перемножается с коэффициентами амплитудно-частотных характеристик кохлеарных фильтров. Таким образом, анализ сигнала происходит преимущественно в частотной области. Основным преимуществом такого подхода является резкое снижение вычислительной и алгоритмической сложности применяемого алгоритма, однако точность такого анализа будет сильно зависеть от частотного разрешения. Например, в используемом банке кохлеарных фильтров [17] наименьшая полоса пропускания фильтра $\Delta f_{\min} \approx 40$ Гц. При использовании дискретного преобразования Фурье (ДПФ) длиной 1024 отсчета на частотный диапазон $\Delta f_{\min}$ приходится всего 4–5 отсчетов, что недостаточно для полноценного анализа. С другой стороны на самую большую полосу пропускания приходится более 10 частотных отсчетов. Таким образом, нельзя достаточно корректно провести антропоморфический анализ.

Для решения данной проблемы предлагается использовать ДПФ с неравномерной частотной шкалой (с неравным интервалом дискретизации по оси частот [18]):

$$\hat{X}(z_k) = X(\hat{z}_k) = \sum_{n=0}^{N} x[n] \hat{z}_k^{-n} ,$$

где $\hat{z}_k$ — образ равноудаленных точек на единичной окружности в z-плоскости, получаемых в результате преобразования:

$$z_k^{-1} = e^{-j\frac{2\pi k}{N}} \longrightarrow \hat{z}_k^{-1} = A(z_k) \quad k = \overline{0, N-1} ,$$

где A(z) — произвольная передаточная функция всепропускающего фильтра:

$$z^{-1} \to A(z) = \frac{z^{-1} - a}{1 - az^{-1}}, \quad |a| < 1 .$$

Размещение коэффициентов зависит от используемой характеристики A(z):

$$\hat{\omega} = \omega + 2\arctan\left(\frac{a\sin\omega}{1 - a\cos\omega}\right) \quad \text{для} \begin{cases} z = j\omega, \\ \hat{z} = j\hat{\omega}. \end{cases} \qquad (2)$$

Знак параметра a в (2) определяет, какой частотный диапазон будет вытягиваться: при $a>0$ — область нижних частот, при $a < 0$ — область верхних частот. Противоположная часть коэффициентов будет соответственно сжиматься.

В ходе экспериментов было выяснено, что при использовании ДПФ с неравномерной частотной шкалой длиной 1024 отсчета для оцифрованного с частотой дискретизации 8000 Гц речевого сигнала оптимальное значение $a\approx 0{,}4$. При этом на полосу пропускания каждого фильтра приходится примерно одинаковое количество отсчетов в частотной области.

На рис. 6 и 7 изображены амплитудно-частотные характеристики кохлеарного фильтра с центральной частотой $f_0 \approx 47$ Гц полосой пропускания $\Delta f \approx 40$ Гц, которые используются для анализа на основе обычного ДПФ и ДПФ с неравномерной частотной шкалой соответственно (кружками условно показаны отсчеты, на которых вычислялась АЧХ). Таким образом, при использовании ДПФ с неравномерной частотной шкалой на анализируемый частотный диапазон приходится 14 отсчетов, что значительно больше, чем в случае обычного дискретного преобразования Фурье.

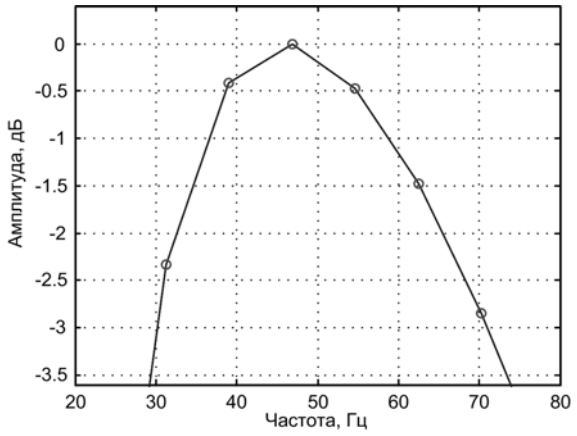

Рис. 6. АЧХ первого фильтра из банка (обычное ДПФ)

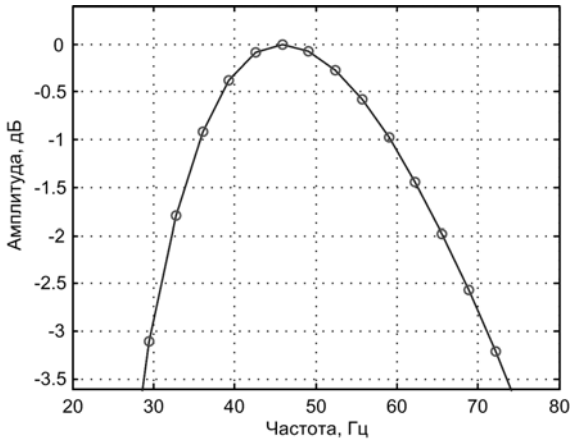

Рис. 7. АЧХ первого фильтра из банка (ДПФ с неравномерной частотной шкалой)

Квантование параметров

В ходе моделирования работы вышеописанной системы анализа экспериментально исследовались законы распределения параметров модели [22]. Как показали проведенные эксперименты, фазы имеют гауссово распределение, а амплитуды и частоты — нет. Поэтому для кодирования синусоидальных параметров предлагается следующий подход. Поскольку частоты и амплитуды определяются по спектральной характеристике речевого сигнала, для их кодирования целесообразно использовать векторное квантование в пространстве амплитуда-частота. На сторону декодера в качестве параметра вместо непосредственных значений амплитуды и частоты передается только индекс найденного элемента в кодовой книге. Фазы кодируются с использованием скалярного квантования. Амплитуды кодируются в логарифмическом диапазоне от 0 до 60 дБ с равномерным шагом. Этого вполне достаточно для кодирования речевого сигнала с хорошим качеством [2]. Поскольку применяется двумерное векторное квантование, то производится согласование шага квантования по амплитуде и частоте [22].

Процесс квантования амплитуды и частоты для одной синусоидальной составляющей может быть описан следующим образом. Из определенных в процессе анализа значений амплитуды A и частоты k для каждой синусоидальной компоненты формируется входной вектор. В заранее сформированной кодовой книге C с набором векторов y_i осуществляется поиск, используя критерий минимальной ошибки квантования между входным вектором и вектором из кодовой книги. Когда такой вектор найден, его индекс кодируется и передается на сторону декодера, который содержит копию кодовой книги из кодера.

Кодовая книга представляет собой набор из L предопределенных выходных векторов y_i:

$$C = \{y_i\}, \quad i = \overline{1, L} \quad ,$$

где

$$y_i = \{A_i, k_i\}.$$

Для поиска оптимального элемента в кодовой книге используется критерий ближайшего элемента (nearest neighbor condition) [23].

Ошибка квантования определяется как среднеквадратическая ошибка [23]. В данном случае ошибка квантования $d(x, y)$ для входного вектора x и вектора из кодовой книги y определяется как квадрат разности координат (т.е. соответствующих значений амплитуды и частоты) этих векторов и вычисляется по следующей формуле:

$$d(x, y) = \frac{1}{2}\left(A^x - A^y\right)^2 + \frac{1}{2}\left(k^x - k^y\right)^2 ,$$

где A^x, A^y — амплитуды синусоид для входного вектора и ближайшего вектора из кодовой книги соответственно; k^x, k^y — индексы частотных отсчетов синусоид для входного вектора и ближайшего вектора из кодовой книги соответственно.

Как показывают экспериментальные результаты, описанный выше подход для квантования синусоидальных параметров дает хорошие результаты, если длина кодовой книги равна 4096 или больше. Однако в этом случае данный алгоритм квантования параметров имеет высокую вычислительную сложность, что создает трудности для его использования в системах реального времени. Также если длина кодовой книги очень велика, то возникают некоторые трудности в процессе тренировки кодовой книги [22].

С другой стороны, если кодовая книга слишком мала, то выходная синтезированная речь имеет плохое качество, появляются значительные искажения. Кроме того, экспериментальным образом установлено, что именно ошибка по частоте чрезвычайно сильно влияет на качество синтезируемой речи. Поэтому предлагается скомбинировать векторное квантование амплитуд и частот с дополнительным скалярным квантованием ошибки по частоте [22].

Исследование качества реконструированного речевого сигнала

В качестве примера рассмотрим процесс компрессии и декомпрессии отрезка оцифрованного речевого сигнала, которому соответствует слово на русском языке "недорогой" — рис. 8. Частота дискретизации сигнала F_S =8000 Гц, амплитуда – 16 бит на отсчет, количество каналов — один, голос — мужской. Параметры кодера: количество синусоидальных составляющих — 7; длина преобразования Фурье — 1024; длительность окна анализа — 32 мс; длительность фрейма синтеза — 22,5 мс.

При указанных выше параметрах скорость передачи составляет около 6 кбит/с. По субъективным ощущениям разборчивость речи и узнаваемость диктора – хорошие, однако в восстановленной речи присутствуют небольшие тональные артефакты.

В соответствии с методикой, приведенной в [24], для синусоидального вокодера с антропоморфической обработкой при скорости передачи параметров 8000 бит/с были проведены следующие тесты: 1) измерение разборчивости речи артикуляционным методом; 2) измерение качества речи и узнаваемости диктора методом парных сравнений с контрольным трактом; 3) измерение степени узнаваемости голоса диктора.

Для измерения разборчивости речи артикуляционным методом в качестве речевых образцов использовались предварительно надиктованные и записанные в формате . WAV наборы слогов из специальных слоговых артикуляционных таблиц [24]. Были записаны голоса для трех

дикторов (два мужских и один женский). Для прослушивания были задействованы три аудитора.

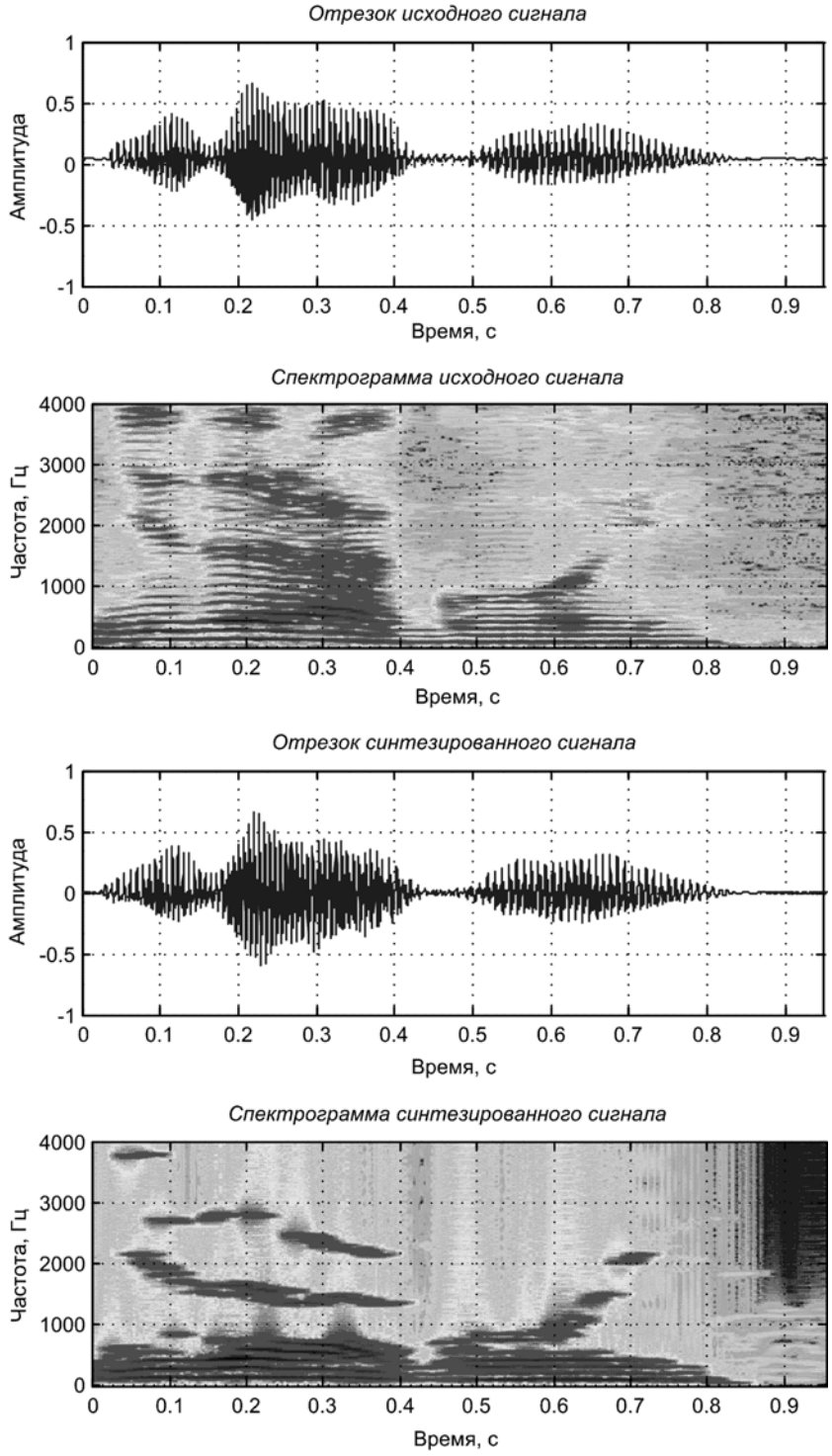

Рис. 8. Спектрограмма синтезированного сигнала

Для измерения качества речи и узнаваемости диктора методом парных сравнений с контрольным трактом и измерения степени узнаваемости голоса диктора в качестве речевых образцов использовались предварительно надиктованные и записанные в формате .WAV фразы. Были записаны и прослушаны голоса для пяти дикторов (три мужских и два женских).

Получены следующие результаты [25]:

1) слоговая разборчивость речи — 96 %;

2) субъективное качество речи — 3,8 баллов;

3) характеристика восстановленной в декодере речи — некоторое нарушение естественности и узнаваемости, иногда присутствуют подзванивание и дребезжание;

4) узнаваемость голоса диктора — 95 %.

Заключение

Проведенные эксперименты показывают, что при использовании предложенного метода компрессии сигнала восстановленная в декодере речь отличается довольно высокой степенью разборчивости и хорошей узнаваемостью диктора даже при ограниченном числе синусоидальных компонент.

ANALYSIS AND SYNTHESIS OF SPEECH CODING DEVICES ON BASIS OF ANTHROPOMORPHICAL PROCESSING AND SINUSOIDAL MODELS

D.S. LIKHACHEV, A.A. PETROVSKY

Abstract

A building method of sinusoidal vocoder with using anthropomorphical speech analysis is proposed in this paper. According to the conception both voiced and unvoiced speech components are represented by a finite set of the so-called "dominating" sinusoidal waves.

Литература

1. *Graf M., Truong H.L.* // Computer networks. 1999. Vol. 31, Issue 3. P. 273.
2. *Kondoz A.M.* Digital speech: coding for low bit rate communication systems. John Wiley & Sons, Inc., N.Y., 1996.
3. *Серков В.В., Петровский А.А.* // Докл. 3-й междунар. конф. "Цифровая обработка сигналов и ее применения" (DSPA`2000). М., 2000. С. 241–244.
4. *Das A., Gersho A.* // Int. J. of Speech Technology. 1999. Vol. 2. P. 317–327.
5. *McAulay R.J., Quatieri T.F.* // IEEE Trans. on Acoust., Speech and Signal Processing. 1986. Vol. ASSP-34. P. 744–754.
6. *Ghitza O.* Advances in Speech Signal Processing. Editors: S. Furui and M.M. Sondhi. N.Y., Marcel Dekker. 1992. P. 453.
7. *Ghitza O.* // IEEE Trans. on Speech and Audio Processing. 1994. Vol. 2, № 1. Pt 2.
8. *Ghitza O.* // J. Acoust. Soc. America. 1993. Vol. 93, № 4. P. 2160–2171.
9. *Лихачев Д.С., Петровский А.А.* // Изв. Белорус. инж. акад. 2002. №2(14)/1 С. 159–162.
10. *Chitza O.* // IEEE Transactions on Speech and Audio Processing. 1987. Vol. ASSP-35. № 6.
11. *Wanggen Wan, Oscar C. Au., Cyan L. Keung, Chi H. Yim.* // Proc. EUROSPEECH'99. 1999. P. 1555–1558.
12. *Oscar C. Au, Wanggen Wan, Cyan L. Keung, Chi H. Yim.* // Proc. EUROSPEECH'99. 1999. P. 2287–2290.
13. *Edler B.* // ISO/IEC, JTC1/SC29/WG11 MPEG95/MO414. Oct. 1995. P. 3617–3620.
14. *Edler B., Purnhagen H.* // ISO/IEC, JTC1/SC29/WG11 MPEG96/MO632. Jan.1996.
15. *Purnhagen H., Edler B., Ferekidis C.* // Proc. 104th Conv. Aud. Eng. Soc., May 1998. P. 4747.
16. *Лихачев Д.С., Петровский А.А.* // Доклады 5-ой междунар. конф. "Цифровая обработка сигналов и ее применение" (DSPA'2003). Москва, Россия. 12–17 марта 2003 г. Т. 2. М., С. 379–382.
17. *Petrovsky A.A., Likhachov D.S., Wan W.* // International Scientific Journal of Computing. 2004. Vol. 3, Issue 1. P. 75–83.
18. *Лихачев Д.С.* Антропоморфический анализ на основе дискретного преобразования Фурье с неравномерной частотной шкалой // Изв. Белорус. инж. акад. 2005. №1(19)/2 С. 177–180.
19. *Petrovsky A.A. , Likhachov D.S.* // The proc. of the III International Conference on Neural Networks and Artificial Intelligence (ICNNAI'2003). 12–14 November 2003 г., Minsk, Belarus. Minsk, 2003. P. 126–131.
20. *Likhachov D.S., Petrovsky A.A.* // Computer Information Systems and Industrial Management Applications. 26–28 June 2003. Elk, Poland. P. 11–19.
21. *Ivanov A.V., Likhachev D.S., Petrovsky A.A.* // The proc. of the 9th Intern. Workshop on Systems, Signals and Image Processing (IWSSIP'02). 7–8 Nov. 2002. Manchester, UK. P. 231–236.
22. *Likhachov D.S., Petrovsky A.A.* // The proc. of the 9th Intern. conference "Speech and Computer" (SPECOM'2004). 20–22 September 2004. St. Petersburg, Russia. P. 195–202.
23. *Makhoul J., Roucos S., Gish H.* // Proc. IEEE. Nov. 1985. Vol. 73. P. 1551–1588.
24. СТБ ГОСТ Р 50840-2000. Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости.
25. Отчет о НИР "Разработать процедуры сжатия речевой информации, обеспечивающие коммерческое качество восстановленной речи". БГУИР. Минск, 2005.