

НИЗКОСКОРОСТНОЙ КОДЕР РЕЧИ НА ОСНОВЕ МОДЕЛИ СЛУХА ЧЕЛОВЕКА

Д.С. Лихачёв, А.А. Петровский

Белорусский государственный университет информатики и радиоэлектроники
220027, Минск, ул. П.Бровки, 6 (Беларусь)
E-mail: palex@bsuir.by

Реферат. В данной работе предлагается метод построения низкоскоростной вокодерной системы на основе кохлеарной модели с синусоидальным представлением речи. Использование в предлагаемой системе цифровой модели слуховой системы человека позволяет получить представление о циркулирующей на уровне слухового нерва акустической информации и значительно уменьшить объём кодируемой речевой информации без значительных потерь в качестве синтезируемой речи. В работе описываются основные подходы реализации кодера речи и даны результаты моделирования работы системы.

1. Введение

В настоящее время наблюдается интенсивное развитие и использование цифровых каналов связи для передачи речи. Использование цифрового представления речевых данных позволяет обеспечить не только надежность связи, но и возможность гарантированной защиты от несанкционированного доступа. Для повышения экономической эффективности этого вида связи необходимо решить проблему минимизации количества бит, необходимых для передачи речи с коммерческим качеством. Для скоростей передачи сигнала около 2 Кбит/с до сих пор не разработаны достаточно хорошие методы кодирования речи с высоким качеством восприятия синтезированного сигнала в реальных условиях окружающей среды.

В данной работе изложены основные подходы построения низкоскоростного кодера речи со скоростью передачи от 2 до 16 Кбит/с. Предлагаемый кодер использует синусоидальное представление речи [1]. Для улучшения перцептуального качества синтезируемой речи предлагается использовать модель слуха человека [2].

2. Синусоидальная модель речеобразования

В основе предлагаемой модели лежат идеи, изложенные в работе [1]. Синтезируемый речевой сигнал в этом случае рассматривается как выход линейной системы, представляющей характеристики речевого тракта при поступлении на нее сигнала возбуждения от голосовых связок. Согласно модели речеобразования [1] синтезируемый речевой сигнал при длительности анализируемого фрейма речи около 20-25 мс имеет вид:

$$s(n) = \sum_{l=1}^L A_l \cos(\omega_l n + \varphi_l) \quad (1)$$

где n – номер отсчета; L – количество синусоид; A_l – амплитуда l -й синусоиды; ω_l – частота l -й синусоиды; φ_l – фаза l -й синусоиды.

3. Модель слуха человека.

Очевидно, что при кодировании речевого сигнала на основе модели (1) синтезированная речь будет тем более качественной и близкой по восприятию к оригиналу, чем больше будет использовано синусоидальных компонент. Необходимое для корректного представления сигнала количество синусоид может быть значительно уменьшено при использовании кохлеарной модели [3]. На основе этой модели вычисляется гистограмма [2], которая позволяет получить представление об

акустической информации, циркулирующей на уровне слухового нерва и дифференцировать частотные составляющие анализируемого речевого сигнала по степени их “важности” и информативности для человеческого слуха – рис.1а.

Согласно модели, представленной в работе [3] функционирование улитки уха на электрическом уровне описывается работой как банка цифровых кохлеарных фильтров с высокой степенью перекрытия полос пропускания:

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

где u_s - входной синусоидальный сигнал, характеризующий скорость перемещения стремечка в ухе человека; $y_k(n)$ – 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. Амлитудно-частотные характеристики для 32 кохлеарных фильтров приведены на рис.1б. Процессы, происходящие при анализе акустической информации в слуховом нерве, имитируются массивом детекторов пересечения уровня $I..P$, расположенных на выходе каждого кохлеарного фильтра. Силу нервного возбуждения на физиологическом уровне характеризует определенное пороговое значение уровня пересечения. По интервалу между пересечениями можно определить частоту входного сигнала в данный момент времени. В зависимости от количества пересечений каждой фиксируемой частоте назначается определеннный весовой коэффициент. Для построения гистограммы подсчитываются все весовые значения с выходов всех каналов для каждой частоты.

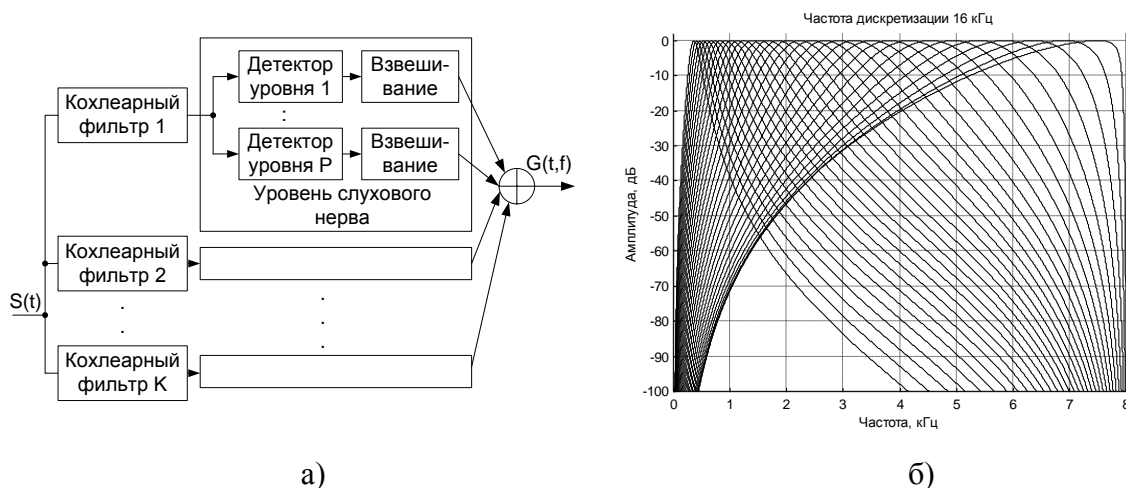


Рис.1. Групповая интервальная гистограмма: а) вычисление групповой интервальной гистограммы, б) АЧХ для 32 кохлеарных фильтров

4. Анализ и синтез речи на основе модели слуха

Процесс анализа речи при использовании предлагаемой слуховой модели можно представить следующим образом – рис.2. Входной дискретизированный речевой сигнал $S(n)$ периодически анализируется с помощью временного окна Хэмминга $W(n)$. На основе гистограммы $G(f)$ по модулю спектра $|S(f)|$ осуществляется отбор пиков, наиболее подходящих с точки зрения наилучшего перцептуального качества синтезированной речи. По положению выбранных пиков определяются частоты синусоид, а по их значению – амплитуды. Фазы синусоид определяются по действительной и мнимой компонентам спектра $S(f)$ на соответствующих частотах. Процесс синтеза речи сводится к суммированию сгенерированных синусоидальных компонентов с найденными в процессе анализа амплитудами, фазами и частотами – рис.2. При этом для получения в процессе синтеза приемлемого качества речи необходимо генерировать синусоиды, непрерывно эволюционирующие во времени. С

этой целью применяется частотное упорядочивание синусоид и интерполяция их параметров от фрейма к фрейму [1].

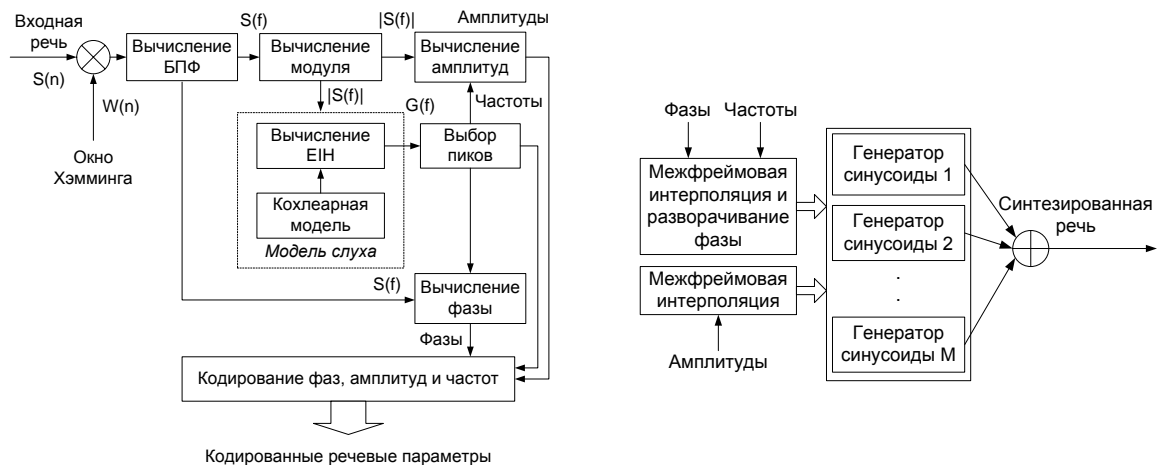


Рис.2. Анализ и синтез речевого сигнала

Процесс анализа-синтеза речи, используя вышеописанную модель речеобразования, был промоделирован в системе Matlab 5.3. Полученные результаты представлены на рис.3. При обработке речи использовалось временное окно Хэмминга длительностью около 30 мс с перекрытием около 10 мс при частоте дискретизации сигнала 8 кГц. Длина преобразования Фурье – 1024 отсчётов. Число синусоидальных компонент равно 7. Число кохлеарных фильтров – 64.

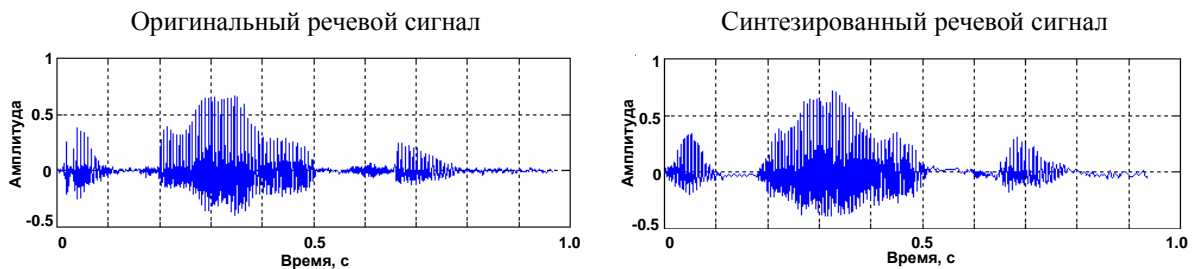


Рис.3. Результаты моделирования

5. Заключение

Предложенная система анализа-синтеза речи имеет относительно простую алгоритмическую реализацию – обрабатываемый речевой сигнал не разделяется на вокализованные и невокализованные фреймы, не требуется процедура определения частоты основного тона. При анализе речи используется кохлеарная модель, которая хорошо согласуется с особенностями человеческого слуха [3] и позволяет значительно улучшить качество синтезируемой речи. Результаты моделирования показали, что предлагаемая система отличается довольно высокой степенью разборчивости и хорошей узнаваемостью диктора даже при ограниченном числе синусоидальных компонент (от 7 до 16).

Библиография

1. McAulay R.J., Quatieri T.F, "Speech analysis/synthesis based on a sinusoidal representation", IEEE Trans. on Acoust., Speech and Signal Processing. – 1988. - Vol. ASSP-34. - P. 744-754.
2. O. Ghitza, "Auditory Nerve Representation as a Basis for Speech Processing", Advances in Speech Signal Processing, edited by Sadaki Furui, Tokyo, Japan, pp. 453 – 485.
3. W. G. Wan and C. X. Fan, "A new solution for cochlear macromechanics", *Acustica*, vol. 75, pp. 79–82, 1991.