УДК 612.382

СИСТЕМА АНАЛИЗА–СИНТЕЗА ГОЛОСА НА ОСНОВЕ ПЕРИОДИЧЕСКИ-АПЕРИОДИЧЕСКОЙ ДЕКОМПОЗИЦИИ С АВТОРЕГРЕССИОННОЙ ПАРАМЕТРИЗАЦИЕЙ

ТХАЙ ЧУНГ КИЕН

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

Поступила в редакцию 28 апреля 2007

В статье представлена система анализа–синтеза голоса, основанной на периодически-апериодической декомпозиции с авторегрессивной параметризацией. Чтобы разложить вокальный сегмент на периодическую и апериодическую компоненты используется дискретное преобразование Фурье с частотой основного тона (TVDFT). Периодическая компонента — сумма синусоидальных составляющих, которые представлены основным тоном, амплитудами и фазами. Мы использовали авторегрессивный метод для моделирования огибающих спектров и получения параметров амплитуды. Результаты исследования показали, что предложенный метод соответствует речевой модели в работе [2, 3] и он удобен для системы конверсии голоса.

Ключевые слова: периодически-апериодическая декомпозиция, авторегрессионная параметризация.

Введение

В последние десятилетия было предложено достаточно большое количество моделей анализа–синтеза речи, например синусоидальная модель, модель гармоники плюс шум. Синусоидальные модели, основанные на известном предположении о том, что речевой сигнал может быть представлен как сумма синусоидальных сигналов со своими амплитудами и частотами, были предложены в работах [4]. Кроме этого, модель гармоники плюс шум (Harmonic plus Noise Model — HNM) была предложена в работе [5]. Ее фундаментальный принцип основан на концепции определения максимальной частоты вокализованного сигнала, которая оценивается для каждого сегмента. В вокализованных и переходных сегментах, гармонические компоненты могут быть получены только до определенной частоты. Более высокие частотные компоненты рассматриваются как шум. Синтез выполняется с перекрытием и накоплением. Перекрывающиеся окна центрированы относительно сегмента с целью обеспечения когерентной фазы [7]. Шумовая составляющая, измеренная в течение интервала анализа, формируется при помощи окрашивания белого гауссова шума кратковременным фильтром-предсказателем (LPC). HNM использовалась как для синтезатора текста в речь (TTS), так и в системе конверсии голоса. Результаты показали, что для TTS HNM превосходит TD-PSOLA по всем вышеописанным характеристикам за исключением вычислительной сложности [6]. Тем не менее максимальная частота вокализованности сигнала, как правило, ограничена значением 4 кГц, причем речевой сигнал оцифровывается с частотой дискретизации 16 кГц.

В проводимом тестировании система анализа–синтеза, основанная на периодически-апериодической декомпозиции, использовалась для кодирования речи [2, 3]. Это позволило синтезировать высококачественную речь без сглаживания на границах фреймов. Однако при

использовании этой модели для конверсии голоса в процессе обучения и трансформации имеет место проблема высокой размерности вектора признаков, что не позволяет обеспечить выравнивание временной длительности, так как число гармоник отличается между фреймами. Для того чтобы преодолеть эти проблемы, рассмотрим речевой синтез, основанный на периодически-апериодической декомпозиции с авторегрессивной параметризацией.

Метод декомпозиции речевого сигнала

Предположим, что речевой сигнал состоит из двух частей: периодической (гармоническая) части и апериодической (шумовая) части как в [2, 3]:

$$s(n) = h(n) + r(n). \quad (1)$$

Квазипериодический компонент речевого сигнала считается как периодическая часть, а апериодический компонент — шум. Предположим, что для каждого речевого сегмента возбуждающий сигнал состоит из совокупности синусоидальных сигналов и результирующий сигнал имеет вид

$$h(n) = \sum_{k=0}^{K(n)-1} A_n(k)\cos(k\omega_0 n + \varphi_k) = \sum_{k=0}^{K(n)-1} A_n(k)\cos(\theta(n,k) + \varphi_k),\ \theta(n,k) = nk\omega_0, \quad (2)$$

где $A_n(k)$ и $\varphi_k(k)$ — амплитуда и фаза для k-й гармоники; ω_0 — фундаментальная частота, $K(n)$ — номер гармоники.

Сигнал является периодическим, то он может быть представлен множеством дискретных отсчетов. Для определения параметров синусоидальных составляющих (проведения гармонического анализа) может быть использовано дискретное преобразование Фурье, которое описывается математическим уравнением (3):

$$H(k) = \sum_{n=0}^{N-1} s(n)e^{-jn\omega_0 k} \quad (3)$$

Амплитуды и фазы могут быть представлены следующим образом: $A_n=|H(k)|$ и $\varphi_{k=}\angle H(k)$. В уравнении (3) принято, что фундаментальная частота f_0 является постоянной на интервале анализа, однако она изменяется по времени. Решить эту проблему можно, применив ДПФ согласованное с частотой основного тона (Time Varying Discrete Fourier Transform — TVDFT), предложенное в [2, 3], которое обеспечивает более точную периодически-апериодическую декомпозицию речи. $\theta(n,k)$ определяется следующим образом:

$$\theta(n,k) = nk\omega_0 + \frac{\Delta\omega_0 n}{2N}, \quad (4)$$

где $\Delta\omega_0$ — отклонение между двумя смежными фреймами речи, N — длина фрейма. Перепишем уравнение (3) в виде

$$H(k) = \sum_{n=0}^{N-1} s(n)e^{-j\left(nk\omega_0 + \frac{\Delta\omega_0 n}{2N}\right)}. \quad (5)$$

Главное преимущество анализа на основе введенного в данной работе ДПФ согласованного с частотой основного тона (TVDFT) в отличие от скачущего ДПФ — частота основного тона точно совпадает с частотным отсчетом соответствующего коэффициента преобразования. Спектральный анализ проводится в области гармоник частоты ω_0. Шумовая составляющая вычисляется следующим уравнением:

$$r(n) = s(n) - h(n). \quad (6)$$

Авторегрессивная параметризация

Периодически-апериодическая модель декомпозиции может достигать высококачественной речевой обработки, особенно в областях модификации частоты основного тона и масштаба времени в синтезе речи. Кроме того, уравнение (2) почти отличается от оригинального речевого сигнала (рис. 2, 3). Однако данная модель имеет высокий порядок. Например, если фрейм речевого сигнала имеет фундаментальную частоту F_0=100 Гц, и частота дискретизации F_s=16000, то требуется K=160 амплитудных параметров. Число амплитудных параметров также должно зависеть от фундаментальной частоты, которые различные между фреймами. В области конверсии голоса, чтобы трансформировать спектральную огибающую число параметров каждого фрейма должно быть одинаково. По этой причине мы должны параметризовать периодически-апериодическую модель декомпозиции эквивалентной моделью речи. Исходная модель фильтра выбрана. Модель исходного фильтра представлена параметрами, описывающими передаточную функцию вокальной модели тракта. Два типа модели исходного фильтра полезны для обработки речи: модель полюсного фильтра, известная как авторегрессивная модель, и модель фильтра с полюсами и нулями, известная как авторегрессивная модель скользящего среднего (ARMA). Авторегрессивная модель вокального тракта известна в обработке речи как модель линейного предсказания (LPC), или модель полюсного фильтра вокального тракта с бесконечной импульсной характеристикой (БИХ-фильтр) [1]:

$$P(z) = \frac{G}{1 + \sum_{k=1}^{N_A} a_k z^{-k}}, \qquad (7)$$

где N_A — порядок модели, коэффициент G и коэффициенты a_k являются параметрами авторегрессивной модели или параметрами линейного предсказания. Такая модель имеет частотную характеристику, задаваемую уравнением:

$$P(e^{j\omega}) = \frac{G}{1 + \sum_{k=1}^{N_A} a_k \exp(-jk\omega)}. \qquad (8)$$

Амплитуду в уравнении (2) можно выразить $A(k)=|P(k)|$, т.е.

$$P(e^{j\omega}) \cong H(k). \qquad (9)$$

Как показано на рис. 1, форма пересинтезированной гармоники речевого сигнала, с использованием оригинальной амплитуды совпадает с исходной формой сигнала, а форма сигнала, пересинтезированной гармоники не совпадает (их амплитуды отличны). Таким образом, необходимо добавить коэффициент масштабирования в уравнение (2), которое перепишется следующим образом:

$$h_{AR}(n) = g \sum_{k=0}^{K(n)-1} A_t(k) \cos(k\omega_0 n + \varphi_k), \qquad (10)$$

где амплитуды и фазы $A_k=|P(k)|$ и $\varphi_{k=} \angle H(k)$, и коэффициент усиления определяется выражением

$$g = \frac{\sqrt{\sum (|H(k)|^2)}}{\sqrt{\sum (|P(k)|^2)}}. \qquad (11)$$

Результат на рис. 2 показывает, что форма пересинтезированной гармоники речевого сигнала на основе авторегрессивной модели может совпадать с оригинальной формой речевого сигнала и ее периодической составляющей.

Экспериментальные результаты

Этап синтеза может быть изображен автокорреляционым коэффициентом и коэффициентами масштабирования с использованием уравнения (1) и параметров, которые получаются на этапе анализа. На рис. 3 представлена схема анализа речи с использованием предложенной модели. Периодическая и шумовая части синтезируются отдельно, после этого с помощью сложения синтезированной гармонической и шумовой частей синтезируемая речь получается, как показано на рис. 4. Каждый вокализованный сегмент описывается фундаментальной частотой, числом гармоник, коэффициентами LPC и фазами, а невокализованный сегмент — только коэффициентами LPC. Для оценки использовались сигналы мужской и женской речи, которые были дискретизированы с частотой 16 000 Гц. Речевой сигнал обрабатывался фреймами длиной 480 отсчетов с перекрытием 240 отсчетов. Порядок авторегрессивной модели был выбран в диапазоне 10–40. Оценка частоты основного тона и решение о вокализованности фрейма базируются на методе, изложенном в [8].

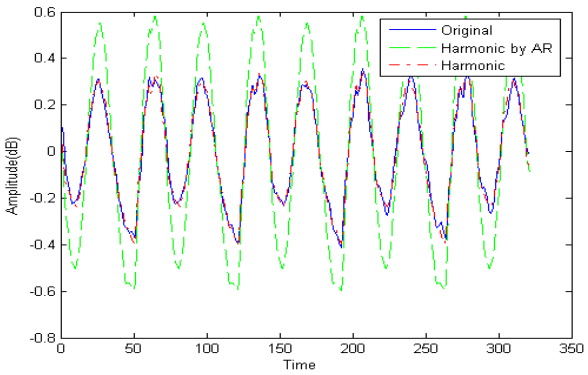

Рис. 1. Форма речевого сигнала, пересинтезированной гармоники и форма волны с использованием авторегрессионной модели (сплошная линия) и оригинальной амплитуды

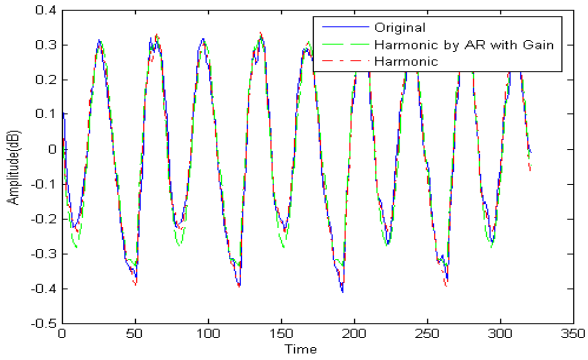

Рис. 2. Форма речевого сигнала, пересинтезированной гармоники и форма волны с использованием авторегрессивной модели (сплошная линия) с коэффициентом увеличения и оригинальной амплитуды

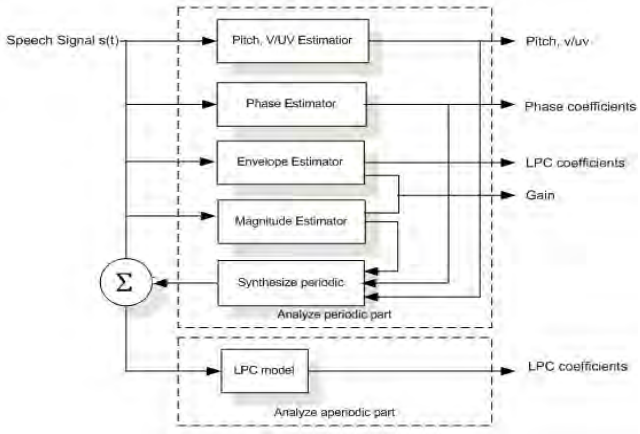

Рис. 3. Схема анализа речи

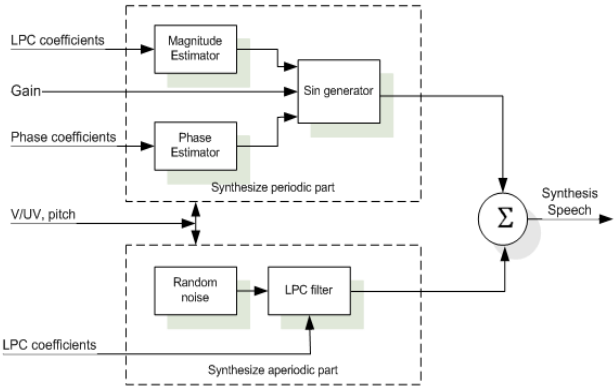

Рис. 4. Схема синтеза речи

Для определения ошибки или различия между периодическими компонентами использовалась среднеквадратическое значение отклонения (Root Mean Square - RMS) и логарифм спектрального искажения (Spectral Distortion — SD). Осуществлялась оценка трех подходов:

периодическая часть, которая синтезируется уравнением (2).

периодическая часть авторегрессивной модели

периодическая часть авторегрессивной модели с коэффициентом увеличения.

Кроме этого, логарифм спектрального искажения вычисляется для трех этих видов периодической части.

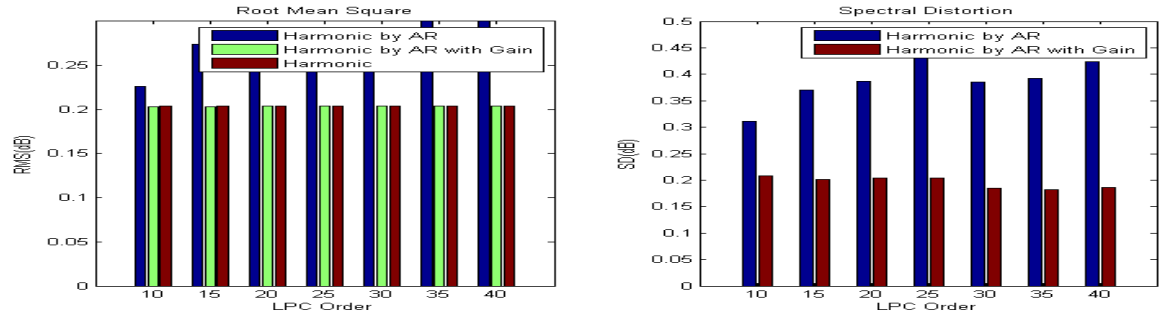

Рис. 5. RMS и логарифм спектрального искажения для женского голоса

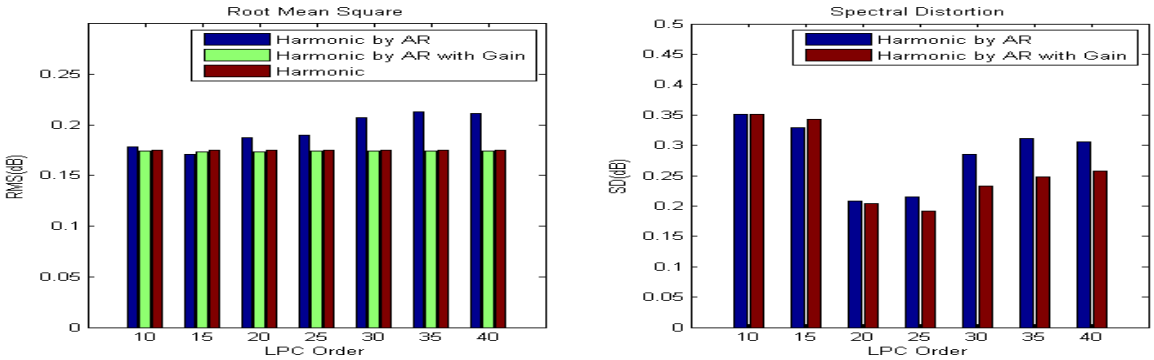

Рис. 6. RMS и логарифм спектрального искажения для мужского голоса

Как показано на рис. 5 и 6, авторегрессивная модель с коэффициентом масштабирования дает лучшие результаты, чем авторегрессионная модель. Тестирование показывает, что качество синтезируемой речи очень высоко, и слушатели не могут отличить оригинальный сигнал от синтезированного моделью на основе периодически-апериодической декомпозиции. В нашем эксперименте порядок фильтра предсказания равен 20, т.е. спектральная огибающая была представлена 20 коэффициентами. Данные в таблице показывают различие между периодически-апериодической моделью декомпозиции и предложенной моделью при изменении фундаментальной частоты речевого сигнала, оцифрованного с частотой дискретизации 16 кГц.

Требуемый порядок модели линейного предсказания для различных значений частоты основного тона

F_0 (Гц)	100	110	120	130	140	150	160	170	180
Старая модель	40	36	33	30	28	26	25	23	22
Предложенная модель	20	20	20	20	20	20	20	20	20

Заключение

Целью данной работы является исследование возможности комбинирования системы анализа–синтеза, основанной на периодически-апериодической модели декомпозиции с моделью исходного фильтра. Представлен результат периодически-апериодической модели декомпозиции с авторегрессионной параметризацией и коэффициентом масштабирования. Результат исследования показал, что качество синтезируемой речи очень высоко, и предложенная модель совпадает с периодически-апериодической декомпозицией, в то же время модель имеет меньший порядок и более стабильна. Кроме того, предложенная модель может быть применена в системе конверсии голоса.

THE VOICE ANALYSIS–SYNTHESIS SYSTEM BASED ON THE PERIODIC-APERIODIC DECOMPOSITION OF SPEECH AND AUTOREGRESSIVE PARAMETRIZATION

THAI TRUNG KIEN

Abstract

The voice analysis-synthesis system based on the periodic-aperiodic decomposition with autoregressive parameterization is represented in this paper. Time varying discrete Fourier transform is used to decompose voiced segment into two parts: periodic part and aperiodic part. The periodic part is sum of sinusoidal components, which are represented by magnitudes and phases. We used autoregressive method to model spectral envelope, and to obtain magnitude parameters. In synthesis process, the periodic and aperiodic part are synthesized separately, and added together. The result shows that, proposed model is fitted with speech model in [2, 3] and it is comfortable speech model for voice conversion system.

Литература

1. *Rabiner, Lawrence R, and Schafer, Ronald W*. // Digital Processing of Speech Signals. Bell Laboratories, 1978.
2. *Sercov V., Petrovsky A. //* An improved speech model with allowance for time-varying pitch harmonic amplitudes and frequencies in low bit-rate MBE coders. Proc. of the European Conf. on Speech Communication and Technology EUROSPEECH. Budapest, Hungary, 1999. P. 1479–1482.
3. *Piotr Zubrycki, Alexander Pavlovec, Alexander Petrovsky*. // XI Symposium AES "New trends in Video and Audio", Białystok, Poland, September 20–22. 2006.
4. *M. W. Macon and M. A. Clements*. // *Proc. ICASSP'96.* Vol. I. Atlanta, GA, 1996. P. 361–364.
5. *Yannis Styliano* // IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 1, January 2001.
6. *Syrdal, A., Y. Stylianou, L. Garisson, A. Conkie and J. Schroeter*.// Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP98). Seattle, USA, 1998. P. 273–276.
7. *Stylianou* // 3rd ESCA Speech Synthesis Workshop, Nov. 1998.
8. *Thai Trung Kien //* PRIP 200, Minks, Belarus. May, 2007.