УДК 681.3

Чистиков Павел Геннадьевич,
Рыбин Сергей Витальевич

ПРОБЛЕМЫ ЕСТЕСТВЕННОСТИ РЕЧЕВОГО СИГНАЛА В СИСТЕМАХ СИНТЕЗА

Аннотация

В данной статье рассматривается подход к синтезу речи, основанный на конкатенации звуковых элементов, как наиболее распространённый и широко применяемый в современных системах для получения естественного речевого сигнала. Описаны проблемы, возникающие при реализации данного подхода и пути их решения, основанные на модификации сигнала. Представлены три метода модификации основного тона: TD-PSOLA, SPECINT и LP-PSOLA. Рассмотрены недостатки и достоинства каждого из алгоритмов, и на основании экспериментальных данных рекомендован лучший.

Ключевые слова: модификация частоты основного тона, синтез речи.

1. ВВЕДЕНИЕ

Процесс изучения и построения систем синтеза речи становится всё более и более популярен в последнее время. Много подходов и алгоритмов предложено в этой области. Если в первых подобных системах акцент делался на разборчивость речи, то теперь особое внимание уделяется её естественности, интонационной насыщенности, тембральной окраске. Голос довольно точно сообщает окружающим о текущем состоянии человека, о его переживаниях, отношении к фактам, самочувствии, а нередко – и о темпераменте, о чертах характера. Уловить эмоции позволяет тон голоса. А для понимания сообщения важны как сила, так и его высота [1].

Тембр придает характерную особенность звуку, индивидуальную окраску, связанную с одновременным воздействием различных звуковых частот, он напрямую зависит от присоединения к основному тону добавочных тонов, возникающих в резонаторной части голосового аппарата.

При компилятивном синтезе из множества речевых единиц (аллофонов) трудно обеспечить оптимальное сочетание этих параметров на протяжении всей фразы даже при наличии большого количества звукового материала. Практически невозможно собрать из аллофонов фразу, обеспечив гладкий основной тон, плавный темп и мелодику без отклонений от естественной речи. Для приближения синтезированной фразы к реально произнесенной необходимо прибегать к модификации, исправлять основной тон во избежание слышимых скачков. Также неизбежна модификация при моделировании требуемого интонационного контура.

Нередко возникает и самостоятельная задача относительной модификации тона (увеличение/уменьшение существующей высоты голоса). В своей речи диктор без

© П.Г. Чистиков, С.В. Рыбин, 2011

особых усилий может изменить основной тон до 2 раз, профессиональные ораторы и певцы способны изменять тон голоса до 5 раз. Здесь важно отметить, что после таких изменений диктор остается узнаваемым, что обеспечивается постоянством тембра голоса. Получается, для воспроизведения определенного интонационного контура может потребоваться модификация тона в 2 раза вверх или вниз.

Таким образом, естественность сигнала зависит от объёма речевой базы, которая содержит всевозможные звуковые единицы с различной частотой основного тона. В современных системах [2] такие базы достигают десяти часов речи. Однако даже такого количества недостаточно, и приходится прибегать к модификации.

В данной статье рассматриваются основные подходы для модификации частоты основного тона, применяемые в современных системах синтеза речи, и основные проблемы, которые возникают при реализации каждой из них. В разделе 2 содержится описание технологий TD-PSOLA, SPECINT и LP-PSOLA. Экспериментальные результаты и сравнения работы каждого из подходов представлены в разделе 3. Выводы и идеи дальнейших разработок вынесены в раздел 4.

2. ОБЗОР ОСНОВНЫХ ПОДХОДОВ

A. Алгоритм TD-PSOLA

Широко распространены алгоритмы, работающие во временной области, наиболее популярным из которых является технология PSOLA [3]. Данный алгоритм работает периодосинхронно, то есть каждый обрабатываемый фрагмент представляет собой один период. Обязательным условием для этого является возможность определить частоту основного тона сигнала с высокой точностью, так как от этого напрямую зависит качество работы этого алгоритма. Границами периодов основного тона служат места закрытия гортани. Далее сигнал разбивается на фрагменты, взвешенные окном Хеннинга, которое захватывает два соседних периода с перекрытием в один период, как показано на рис. 1. Эти взвешенные фрагменты затем могут быть перекомбинированы путём перемещения их центров и наложением с добавлением перекрывающихся частей (отсюда и название overlap and add – перекрытие и добавление). Несмотря на то, что после выполнения данных операций форма результирующего сигнала становится не в точности такой, какая была прежде, процедура перекрытия с добавлением позволяет получить достаточно близкий результат, чтобы различия не были заметны.

Непосредственная модификация частоты основного тона выполняется путём распределения полученных взвешенных фреймов на новые значения частоты, представляющей собой множество расстояний между окнами, им соответствующими. Для примера рассмотрим участок речи с частотой основного тона 100Гц, границы периодов будут лежать с интервалом в 10мс. Взяв эти периоды за основу, проанализируем их и разделим на описанные выше периодосинхронные фрагменты, взвешенные окнами Хеннинга. Далее создадим новое множество периодов, границы которых будут располагаться ближе друг к другу, скажем, через каждые 9 мс. Далее, если перераспределить подготовленные фреймы путём перекрытия с наложением, мы получим сигнал, который будет иметь частоту основного тона, равную 1.0/0.009 = 111 Гц. Если производить обратную операцию – создать множество периодов, границы которых будут располагаться дальше друг от друга, и перераспределить фреймы с перекрытием, мы получим синтезированный сигнал с более низкой частотой основного тона. Процедура уменьшения частоты основного тона частично объясняет причину использования двух периодов во взвешенных фреймах; это делается для того, чтобы не оставалось пустых мест в результирующем сигнале при увеличении расстояния между центрами фреймов.

При сохранении длительности фонограммы в целом слушатели не замечают неестественностей в сигнале при небольших модификациях частоты основного тона [4].

Когда алгоритм применяется для модификации хорошо размеченной на периоды основного тона речи, качество его работы чрезвычайно высоко, и, пока степень изменения частоты основного тона не слишком значительна (скажем +/– 10 % от оригинала), качество речи может быть «идеальным» в том смысле, что слушатель не может заметить в речи какой-то неестественности. С точки зрения вычислительной нагрузки на аппаратные ресурсы, сложно представить какой-либо алгоритм, работающий быстрее. Поэтому зачастую TD-PSOLA рассматривается как приемлемое решение для проблемы модификации частоты основного тона. Однако, конечно, алгоритм не идеален во многих ситуациях, не потому что он не выполняет поставленную задачу, а потому, что на практике, как минимум, нам приходится модифицировать частоту основного тона более чем на 10 %, например, чтобы гарантировать гладкость интонационного контура в синтезированной речи в случаях отсутствия звуковых элементов с требуемой частотой в базе данных. Также, работая во временной области, он вносит неконтролируемые искажения в сигнал, и при уменьшении частоты основного тона существенно редуцируется энергия на границах «склеек» фреймов.

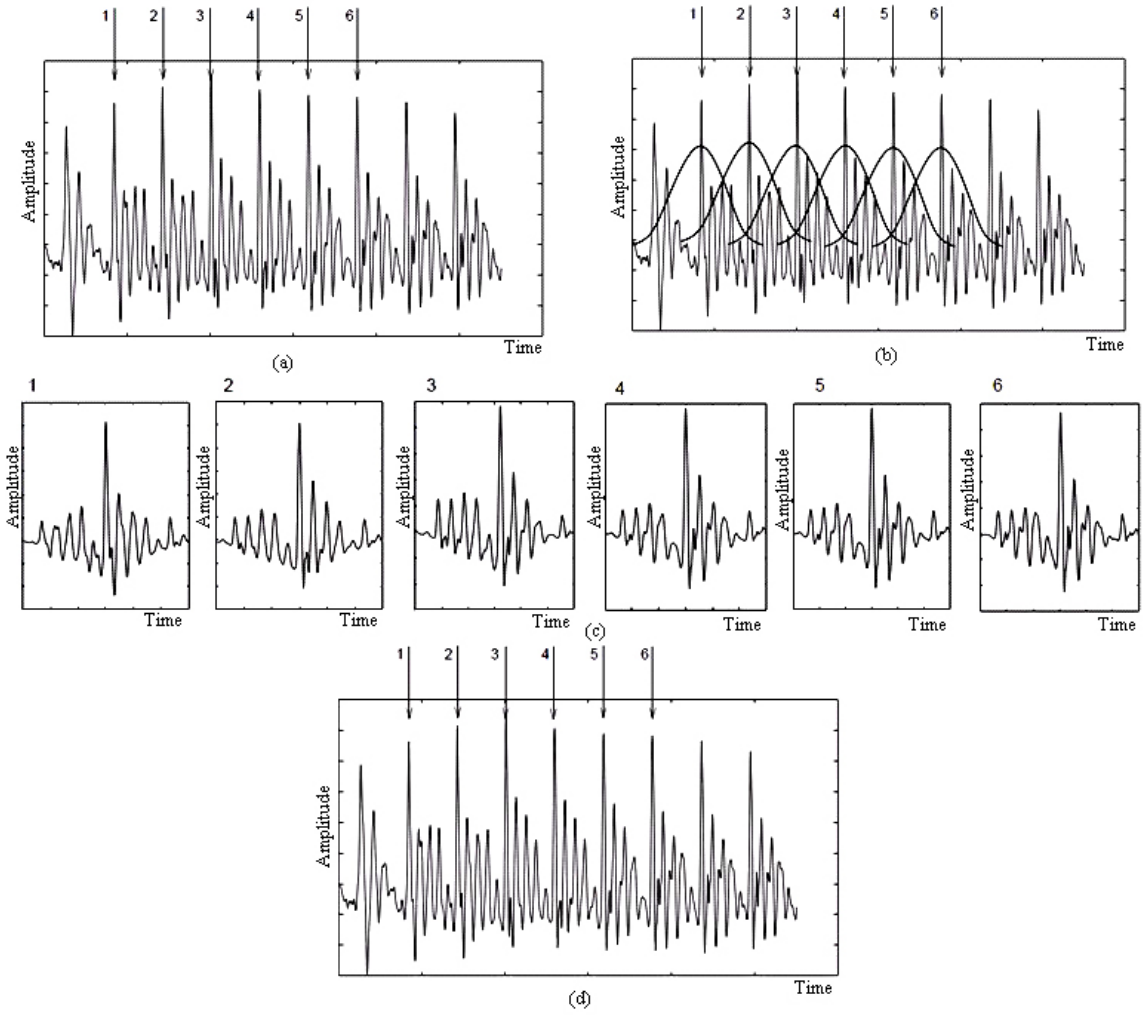

Рис. 1. Основные операции алгоритма PSOLA:
a) участок вокализованного сигнала, размеченный на периоды основного тона,
b) взвешивающие окна Хеннинга, центрированные на каждом периоде,
c) полученная последовательность пар периодов после процедуры взвешивания окном,
d) ресинтезированный путём перекрытия с добавлением сигнал

B. Алгоритм SPECINT

В связи с психоакустическими эффектами малейшие искажения в относительном положении формант и изменения огибающей основного тона ведут к побочным эффектам, из-за которых речь становится неестественной, непривычной для нашего восприятия, как следствие, человек при её прослушивании быстро утомляется и не может длительное время внимательно её воспринимать. Поэтому одним из основополагающих действий является получение огибающей основного тона исходного сигнала и её воспроизведение на сигнале новой длины.

Немаловажно сохранение энергетической огибающей, поскольку при увеличении или уменьшении частоты основного тона появляются неизбежные её искажения, что также приводит к снижению естественности речи.

Перед тем как понизить или повысить основной тон, увеличить или уменьшить длительность, необходимо получить значения основного тона на всём модифицируемом участке. При модификации нужно изменить требуемые характеристики аллофонов так, чтобы траектория основного тона осталась прежней, то есть измениться должен только масштаб (частоты и времени), иначе при малейшем изменении спектральной картины мы услышим режущие слух новые интонации в речи даже при незначительных модификациях. Для этого анализируется сигнал с целью получения вектора значений частоты основного тона на всём его протяжении. В системе синтеза русской речи это аллофон. То есть на каждом периоде аллофона вычисляется значение его основного тона, заполняется некоторый массив данных (вектор значений). Далее полученная огибающая изменяется по тону (поднимается или опускается), затем путём сплайн-интерполяции она растягивается или сжимается на требуемую длину. В итоге получаем модель аллофона после модификации, под которую мы должны модифицировать исходный аллофон.

Модификация сигнала под требуемую модель происходит следующим образом. Каждый период модифицируется под параметры, смоделированные выше. Рассмотрим этот процесс на примере некоторого периода. Путём дискретного преобразования Фурье (ДПФ) получаем спектр сигнала, рассматриваем отдельно вещественные и мнимые его составляющие.

Очевидно, что в спектральной области мы получим пики на частотах, кратных частоте периода. Далее мы интерполируем пики на весь диапазон частот, равный половине частоты дискретизации, и вычисляем значения сплайнов в точках, соответствующих пикам нового периода. Далее, выполнив обратное ДПФ, мы получим период с требуемой частотой.

Однако при таком подходе без дополнений мы не можем контролировать амплитуду результирующего сигнала. Точнее, огибающая амплитуды у нас сохранится, но абсолютное её значение будет отличным от исходного, что сделает сигнал громче или тише, так как этот параметр напрямую зависит от того, повышается или понижается основной тон. С увеличением частоты основного тона амплитуда уменьшается, с уменьшением – увеличивается [5].

Для сохранения исходных величин амплитуды вычисляется нормирующий коэффициент, на который домножаются значения коэффициентов вещественной и мнимой части. В результате получаются пики, находящиеся на огибающей, которая нормирована таким образом, чтобы после обратного ДПФ получились те же значения амплитуд, как и в исходном сигнале.

Данный алгоритм позволяет получать хорошее качество модификации при увеличении или уменьшении частоты основного тона до двух раз. Особенно хорошие результаты получаются в случаях, когда сигнал уже имеет естественную огибающую частоты основного тона. Хотя стоит заметить, что для высоких частот основного тона мы имеем малое количество гармоник для точного формирования данной огибающей, что сказывается на качестве результата. Также существенным недостатком для применения данного метода является потребление огромного количе-

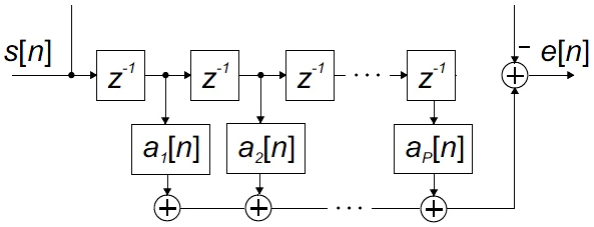

Рис. 2. Структурная схема блока LP фильтра

ства вычислительных ресурсов, так как выполняются сложные математические операции, такие как, например ДПФ.

C. Алгоритм TD-PSOLA

Данный подход [6] комбинирует в себе основные идеи методов TD-PSOLA и SPECINT. Применяется LP модель, изображённая на рис. 2, для получения сигнала ошибки $e[n]$. Далее он модифицируется методом, представленным в разделе 2-А. И в заключение, полученная модифицированная функция ошибки предсказания $e'[n]$ используется для восстановления исходного сигнала с новой частотой основного тона [7].

Формула для вычисления $e[n]$ в (1):

$$e[n] = s[n] - \bar{a}^T \cdot \bar{s}[n-1], \qquad (1)$$

где

$$\bar{s}[n-1] = [s[n-1], s[n-2], ..., s[n-P]]^T, \qquad (2)$$

$$\bar{a} = [a_1, a_2, ..., a_P]^T. \qquad (3)$$

Результирующий вектор коэффициентов линейного предсказания вычисляется по формуле

$$\bar{a}_n = \bar{R}^{-1}[n-1] \cdot \bar{p}[n], \qquad (4)$$

где

$$\bar{R}^{-1}[n-1] = \sum_{i=0}^{n} \bar{s}[i-1] \cdot \bar{s}[i-1], \qquad (5)$$

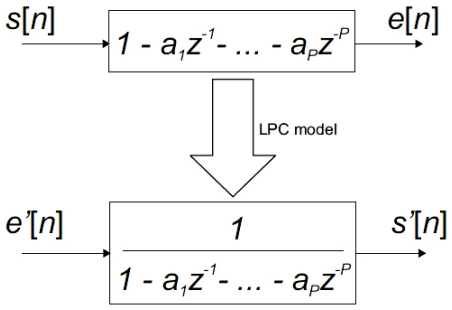

Рис. 3. Схема анализа и синтеза нового сигнала с использованием LP-модели

$$\bar{p}[n] = \sum_{i=0}^{n} \bar{s}[i-1] \cdot s[i]. \qquad (6)$$

Значения $\bar{p}[n]$ в выражении (4) можно вычислить рекурсивно, для того чтобы избежать дополнительных накладных вычислительных расходов, как показано в формуле (7):

$$\bar{p}[n] = \bar{s}[n-1] \cdot s[n] + \bar{p}[n-1]. \qquad (7)$$

Далее, полученные LP коэффициенты на этапе анализа сигнала, применяются в обратном LP фильтре к модифицированной функции ошибки $e'[n]$, для того чтобы получить модифицированный сигнал $s'[n]$ с желаемой частотой основного тона, как показано в (8):

$$s'[n] = e'[n] + \bar{a}^T \cdot \bar{s}'[n-1]. \qquad (8)$$

Общая схема алгоритма представлена на рис. 3.

Модифицированную функцию ошибки $e'[n]$ можно получить из $e[n]$, используя TD-PSOLA, как показано далее. LP модель определяется для каждого отсчёта сигнала n, что позволяет добиться плавных переходов с соседними моделями. Качество результирующей модели зависит от выбора её порядка P.

После правильного определения меток частоты основного тона $p_m[n]$ и периодов частоты основного тона $p[n]$ в исходном сигнале [8–9] контур частоты основного тона может быть модифицирован желаемым образом. С этой целью определяются новые метки $p'_m[n]$, соответствующие значениям новых периодов основного тона $p'[n]$, так что

$$p'[n] = \beta[n] \cdot p[n], \qquad (9)$$

где $\beta[n]$ – это степень модификации периода основного тона, который может быть различным для естественной просодической модификации, автоматической коррекции частоты основного тона и т. д. Новые метки частоты основного тона $p'_m[n]$ определяются путём добавления в интервал $p'[n]$ отсчётов между двумя соседними метками, так что метка частоты основного тона будет перемещена в позицию $n + p'[n]$, если n содержит метку (то есть $p'_m[n + p'[n]] = 1$, если $p'_m[n] = 1$, где

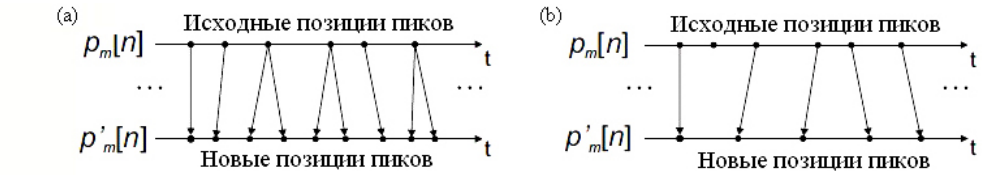

Рис. 4. Распределение меток частоты основного тона при модификации сигнала: *a)* увеличение частоты, *b)* уменьшение частоты

позиция метки частоты основного тона равна 1). На следующем шаге необходимо соединить каждую новую метку частоты основного тона $p'_m[n]$ с соответствующим ей ближайшим пиком в оригинальном сигнале $p_m[n]$. Это делается путём непосредственного сравнения временных индексов $p_m[n]$ и $p'_m[n]$, как показано на рис. 4.

В заключение, для генерации итоговой функции ошибки сигнал разбивается на взвешенные окном Хеннинга фрагменты по парам периодов с перекрытием в один период, то есть для каждой метки, начиная с предыдущей и заканчивая следующей. Данные сегменты соединяются согласно процедуре перекрытия с наложением, так чтобы соответствовать новым периодам частоты основного тона $p'[n]$, полученным ранее, как показано на рис. 5.

Данный подход также обладает рядом недостатков: во-первых, это аппаратная сложность вычисления LP коэффициентов и обратного LP-фильтра, во-вторых, в сигнале возбуждения остаётся информация о голосовом тракте, ведущая к появлению непонятных тембральных артефактов при восстановлении сигнала по модифицированной функции ошибки.

3. ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

В данном разделе приводятся некоторые практические эксперименты модификации частоты основного тона представленными алгоритмами и сравнение результатов их работы.

Пример 1. Участок синтезированной женским голосом речи был модифицирован с $\beta[n] = 2$ и $\beta[n] = 0.5$. Рис. 6 демонстрирует небольшие участки исходного и модифицированного сигналов методами TD-PSOLA, SPECINT и LP-PSOLA. На рис. 7 приведены соответствующие им спектрограммы.

Исходя из данных на рис. 6, можно сделать вывод, что все алгоритмы работают корректно – пики периодов основного тона становятся дальше или ближе друг от друга при модификации с $\beta[n] = 2$ и $\beta[n] = 0.5$ соответственно. Аналогичный вывод можно сделать и проанализировав частотные полосы, представленные спектрограммами на рис. 7.

Пример 2. Вновь характеристики частоты основного тона были модифицированы с $\beta[n] = 2$ и $\beta[n] = 0.5$. Однако в данном примере использовалась речь, синтезированная мужским голосом. Вид результатов во временной и частотной области практически аналогичен тем, что изображены на рис. 6 и рис. 7 соответственно.

Так же, как и в примере 1, по результатам не сложно определить, что частота основного тона модифицирована должным образом, в то время как спектральная огибающая осталась неизменной.

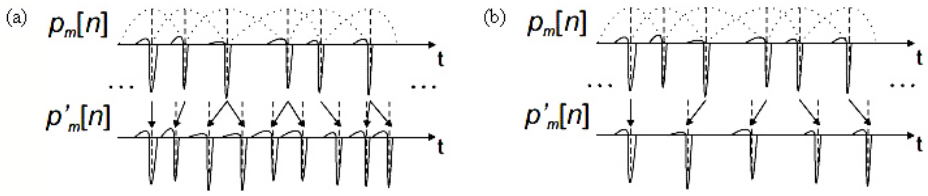

Рис. 5. Получение модифицированного сигнала: *a)* увеличение частоты, *b)* уменьшение частоты

Пример 3. На рис. 8 приведено сравнение результатов представленных методов во временной области с $\beta[n] = 2$ на сигнале из примера 1. Данный пример иллюстрирует значительные недостатки алгоритма TD-PSOLA, которые заключаются в существенной редукции энергии сигнала между соединяемыми пиками периодов частоты основного тона при $\beta[n] > 1$. Хотя большие окна анализа помогли бы исправить эту проблему, они могли бы стать причиной появления ложных пиков в модифицированном сигнале, так как они могут не полностью редуцироваться окнами анализа. Такие ложные пики ведут к грубостям и неестественностям в сигнале.

Стоит отметить, что, вместо непосредственного перекрытия периодов сигнала, как в TD-PSOLA, SPECINT и LP-PSOLA позволяют сохранить индивидуальность откликов на импульсы возбуждения с новой частотой основного тона. Именно поэтому, несмотря на их недостатки, развитие идёт именно в этом направлении.

4. ВЫВОДЫ

В данной статье авторы исследовали основной принцип генерации речевого сигнала в современных системах синтеза речи, основанный на конкатенации звуковых элементов из базы данных. Описаны проблемы, возникающие при данном подходе, связанные с ограничением объёма звуковых баз. Приведено подробное описание алгоритмов, основанных на модификации частоты основного тона исход-

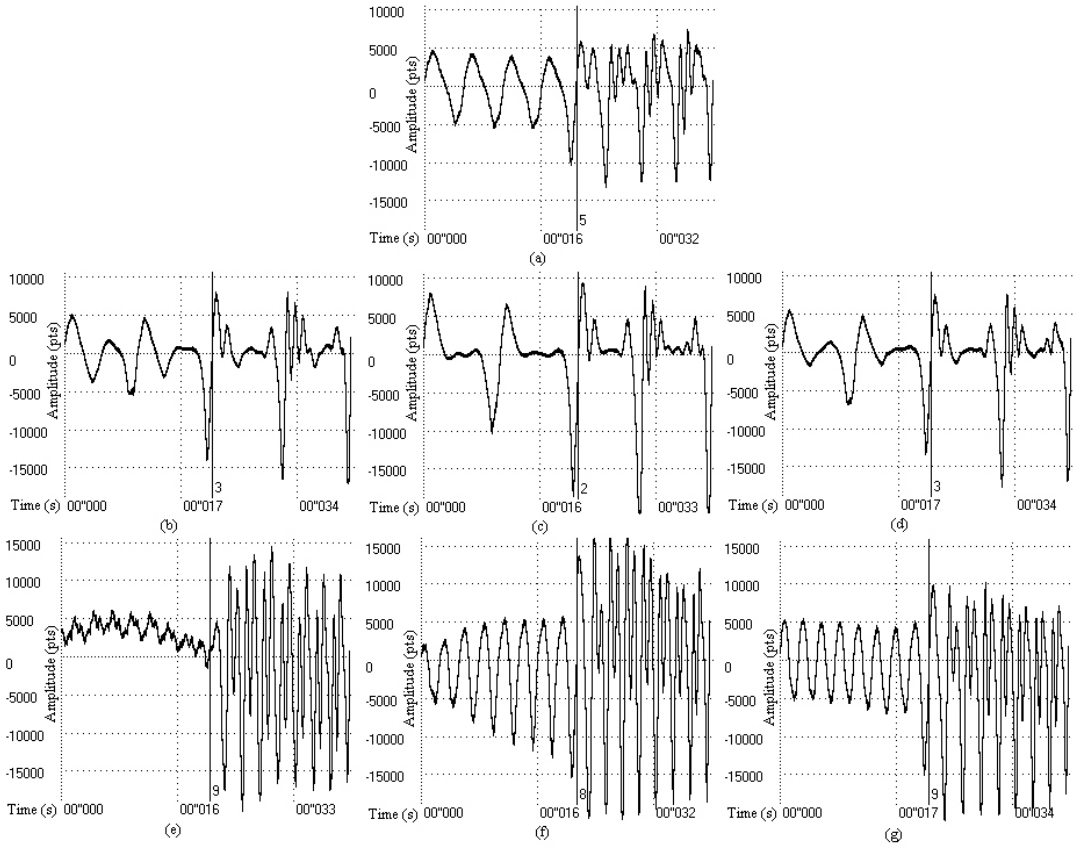

Рис. 6. Фрагменты сигналов в примере 1:
a) исходный сигнал,
b) модифицированный сигнал методом LP-PSOLA с $\beta[n] = 2$,
c) модифицированный сигнал методом SPECINT с $\beta[n] = 2$,
d) модифицированный сигнал методом TD-PSOLA с $\beta[n] = 2$,
e) модифицированный сигнал методом LP-PSOLA с $\beta[n] = 0.5$,
f) модифицированный сигнал методом SPECINT с $\beta[n] = 0.5$,
g) модифицированный сигнал методом TD-PSOLA с $\beta[n] = 0.5$

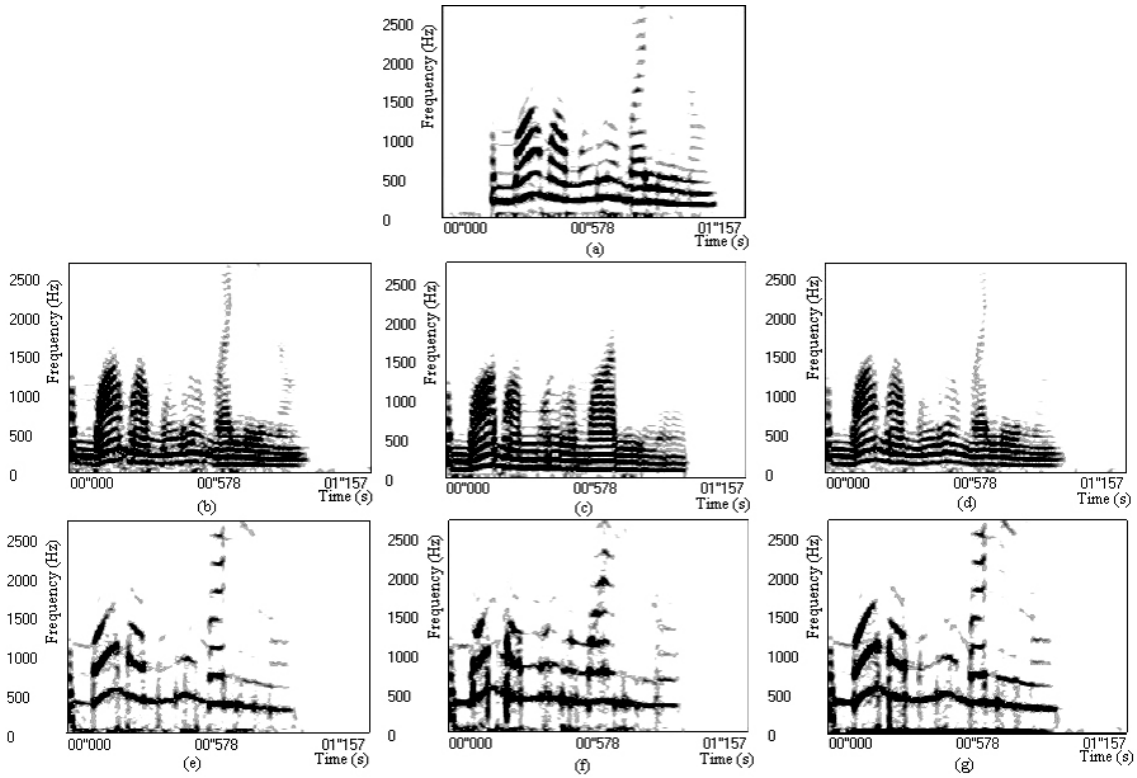

Рис. 7. Спектрограммы сигналов в примере 1:
a) исходный сигнал,
b) модифицированный сигнал методом LP-PSOLA с $\beta[n] = 2$,
c) модифицированный сигнал методом SPECINT с $\beta[n] = 2$,
d) модифицированный сигнал методом TD-PSOLA с $\beta[n] = 2$,
e) модифицированный сигнал методом LP-PSOLA с $\beta[n] = 0.5$

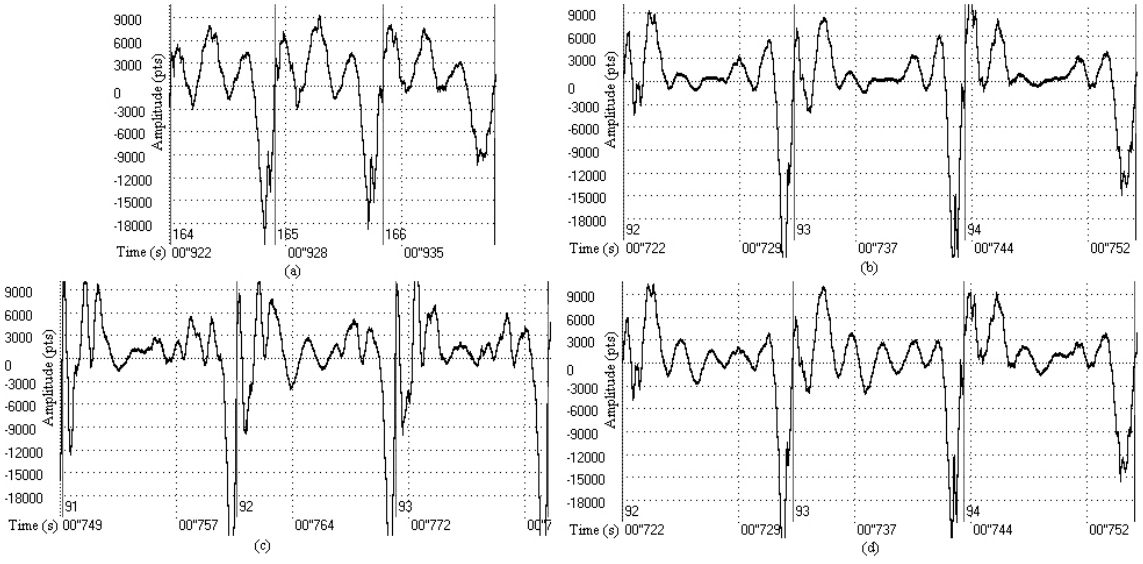

Рис. 8. Фрагменты сигналов в примере 3:
a) исходный сигнал,
b) модифицированный сигнал методом TD-PSOLA,
c) модифицированный сигнал методом SPECINT,
d) модифицированный сигнал методом LP-PSOLA

ного сигнала, для решения этих задач, и описаны основные недостатки, возникающие при использовании каждого из них.

Показано, что наилучшие результаты для получения желаемой частоты основного тона достигаются при использовании алгоритма TD-PSOLA для LP функции ошибки. Преимущество этого метода заключается в сохранении индивидуальности каждого гортанного импульса. Неформальные субъективные испытания показали хорошие результаты модификации частоты основного тона в диапазоне $0{,}5 \le \beta[n] \le 2$. Однако, поскольку в сигнале ошибки частично остаётся формантная структура, метод LP-PSOLA не является совершенным и зависит от порядка LP фильтра.

Авторы выражают благодарность сотруднику ООО «Центр речевых технологий» (Санкт-Петербург) Таланову А.О. за ценные рекомендации и комментарии в процессе исследований. Данная работа финансировалась ООО «Центр речевых технологий». Все права на интеллектуальную собственность принадлежат спонсору.

Литература

1. *Громова В.И., Васильева Г.А.* Энциклопедия безопасности, Россия, 1998.

2. *Корольков В.А., Главатских И.А., Таланов А.О.* Синтез естественной русской речи при помощи метода Unit selection // Тр. XXXVI межд. филолог. конф. «Формальные методы анализа русской речи». Россия, 2008.

3. *Moulines E., Verhelst W.* Time-domain and frequency-domain techniques for prosodic modification of speech in Speech Coding and Synthesis. Netherland, 1995. P. 519–555.

4. *Taylor P.* Text-to-Speech Synthesis/ United Kingdom^ Cambridge University Press, 2008. P. 426-433.

5. *Главатских И.А., Чистиков П.Г.* Метод модификации физических параметров речевого сигнала на основе периодосинхронного Фурье-анализа // Тр. XXXVII межд. филолог. конф. «Формальные методы анализа русской речи». Россия, 2009.

6. *Rafael C. D. de Paiva, Luiz W. P. Biscainho and Sergio L. Netto*.On the application of RLS adaptive filtering for voice pitch modification // in proceedings of the 10th International Conference on Digital Audio Effects. France, 2007.

7. *R. C. D. de Paiva, L. W. P. Biscainho, and S. L. Netto*. A sequential system for voice pitch modification // in proceedings of the 5th AES-Brazil Conference. Brazil, 2007.

8. *S. Kadambe and G. F. Boudreaux-Bartels*. Application of the wavelet transform for pitch detection of speech signals // IEEE Transactions on Information Theory, 1992.Vol. 38, № 2. P. 917–924.

9. *C. Ma, Y. Kamp, and L. F. Willems*. A Frobenius norm approach to glottal closure detection from the speech signal // IEEE Transactions on Speech and Audio Processing, 1994. Vol. 2, № 2. P. 258–265.

Abstract

The paper deals with the approach to speech synthesis based on speech elements concatenation. This approach is the most popular and widely used in the latest systems to generate natural speech. We describe the problems of realizing these methods and present a solution. We present three pitch modification methods: TD-PSOLA, SPECINT and LP-PSOLA. We examine the positive and negative aspects of these methods and choose LP-PSOLA as the most effective of them on the basis of experiments.

Keywords: pitch modification, speech synthesis, text-to-speech.

Чистиков Павел Геннадьевич, магистрант 2 курса кафедры математического обеспечения и применения ЭВМ СПбГЭТУ «ЛЭТИ», pgchistikov@gmail.com,

Рыбин Сергей Витальевич, кандидат физико-математических наук, доцент кафедры ВМ-2 СПбГЭТУ «ЛЭТИ», rsvvm2leti@gmail.com.

© Наши авторы, 2011.
Our authors, 2011.