УДК 519.688

Олейник Андрей Леонидович

МОДИФИКАЦИЯ ВРЕМЕННЫХ И ЧАСТОТНЫХ ХАРАКТЕРИСТИК РЕЧЕВОГО СИГНАЛА

Аннотация

В данной статье предлагаются алгоритмы модификации двух основных временных и частотных характеристик речевого сигнала: темпа и частоты основного тона. Для модификации частоты основного тона используется подход, основанный на дискретном вейвлет-преобразовании. Акцент делается на применении предлагаемых алгоритмов в системах синтеза речи. Также приводится краткий обзор алгоритмов и математических моделей, разработанных для решения данной задачи. Приводятся результаты экспериментальных исследований, перечисляются основные достоинства и недостатки предлагаемых в данной статье алгоритмов, намечается путь устранения недостатков.

Ключевые слова: модификация частоты основного тона, модификация темпа, синтез речи, дискретное вейвлет-преобразование.

ВВЕДЕНИЕ

Алгоритмы модификации речевого сигнала находят применение во многих областях: модификация естественной речи, преобразование речи одного человека в речь другого человека, коррекция дефектов речи, системы синтеза «текст в речь». Предлагаемые в данной статье алгоритмы ориентированы преимущественно на применение в системах компиляционного синтеза речи, где необходима коррекция *частоты основного тона* и *темпа* речевого сигнала.

Понятие частоты основного тона не применимо по отношению к произвольному участку речевого сигнала. Человеческая речь состоит из структурных элементов – *аллофонов*, которые делятся на два класса: *вокализованные* и *невокализованные*. Вокализованные аллофоны обладают квазипериодической структурой, то есть в них можно выделить гармоническую составляющую. Иначе говоря, они состоят из «периодов» – сходных друг с другом отрезков сигнала, длительность которых и характеризует частоту основного тона[1]. Невокализованные аллофоны, напротив, обладают стохастической, шумовой природой. Поэтому понятие частоты основного тона применимо *только* по отношению к вокализованным аллофонам.

Темп, в свою очередь, представляет собой скорость произнесения аллофонов.

Именно темп и частота основного тона определяют интонацию, с которой произносится та или иная фраза. Интонация, в свою очередь, зависит от содержания текста, который подлежит переводу в речь. При этом *тембр*, то есть индивидуальные особенности голоса диктора, должен оставаться неизменным.

Таким образом, алгоритмы модификации временных и частотных характеристик

[1] В дальнейшем будем называть данные участки периодами, опуская кавычки.

играют важнейшую роль в системах синтеза речи. Во многом именно они определяют то, насколько естественно будет звучать сгенерированный речевой сигнал.

На данный момент разработано достаточно большое количество алгоритмов и подходов к модификации временных и частотных характеристик речевого сигнала.

Рассмотрим существующие алгоритмы модификации темпа речи. Основной подход заключается в том, чтобы дублировать отдельные участки сигнала при уменьшении темпа и удалять их при увеличении темпа. Однако дублируемые или удаляемые отрезки сигнала могут быть выбраны по-разному, поэтому различные способы выбора этих участков порождают различные алгоритмы, существенно разнящиеся по своим характеристикам. Так, предложенный в [1] алгоритм Time-Domain Harmonic Scaling (TDHS) при выборе удаляемых или дублируемых участков опирается на оценку локальной частоты основного тона сигнала. Рассмотренные в [2] алгоритмы Synchronized Overlap-Add (SOLA) и его модификация Synchronized Overlap-Add, Fixed Synthesis (SOLAFS) разделяют сигнал на сегменты фиксированной длины, а после их дублирования или удаления сигнал восстанавливается из результирующих сегментов. Для обеспечения плавности перехода между сегментами они попадают в результирующий сигнал с некоторым перекрытием, в пределах которого они усредняются с некоторым весом. Длина каждого перекрытия определяется взаимной корреляцией взвешиваемых участков.

Другой подход к модификации темпа речи заключается в анализе частотных характеристик сигнала посредством дискретного преобразования Фурье (ДПФ).

Кроме того, ДПФ может быть использовано и для модификации частоты основного тона сигнала. Подобный подход нашёл своё отражение в алгоритме Spectrum Interpolation (SPECINT), описанном в [3]. Он сводится к вычислению ДПФ сигнала, интерполяции его мнимой и вещественной частей для получения новых узловых значений и вычислению обратного ДПФ. Кроме того, в [3] представлены алгоритмы Time-Domain Pitch Synchronized Overlap-Add (TD-PSOLA) и Linear-Predictive Pitch Synchronized Overlap-Add (LP-PSOLA). Первый алгоритм разбивает исходный сигнал на сегменты, содержащие по два периода, с пересечением в один период, домножает их на весовую функцию и изменяет длины периодов путём относительного смещения центров сегментов относительно друг друга. Данный алгоритм даёт приемлемые результаты лишь при незначительных изменениях частоты основного тона (+/–10 %), однако отличается исключительным быстродействием. Алгоритм LP-PSOLA комбинирует SPECINT и TD-PSOLA. Используется модель линейного предсказания (LP-модель, Linear Prediction), позволяющая представить сигнал в виде двух составляющих: LP-коэффициентов (коэффициентов линейного предсказания) и сигнала ошибки. Сигнал ошибки модифицируется при помощи алгоритма TD-PSOLA, а LP-коэффициенты модифицируются способом, сходным с алгоритмом SPECINT. Алгоритм LP-PSOLA позволяет получать качественные результаты, однако он является достаточно ресурсоёмким. Кроме того, в силу своей сложности алгоритм LP-PSOLA также порождает тембральные артефакты при восстановлении сигнала из модифицированных LP-коэффициентов и шумовой составляющей.

Ещё один алгоритм модификации темпа и частоты основного тона (ЧОТ) речевого сигнала рассматривается в [4]. Данный алгоритм основан на SOLAFS и дискретном вейвлет-преобразовании, которое позволяет независимо манипулировать составляющими сигнала из различных полос частот.

Также существует достаточно большое количество различных математических моделей, описывающих речевой сигнал. Они также открывают широкие возможности для разработки алгоритмов модификации речи. Одна из них описывается в [5] и носит название параметрической модели «гармоники+шум» (Parametric Harmonic+Noise Model). При таком подходе сигнал раскладывается в сумму двух составляющих: синусоидальную и шумовую. Синусоидальная составляющая отражает квазипериодичес-

кие свойства сигнала. Для описания шумовой составляющей используются стохастические модели. Преимущество подобного подхода в том, что он позволяет использовать различные алгоритмы при модификации двух составляющих сигнала, которые сами по себе не могут быть применены к исходному сигналу. Это потенциально позволяет разрабатывать высококачественные алгоритмы модификации сигнала. Кроме того, следует заметить, что в вокализованных аллофонах преобладает синусоидальная составляющая, в то время как в невокализованных – шумовая. Модель «гармоники + шум» автоматически учитывает данный факт. С другой стороны, подобный подход обладает своими недостатками. Так, алгоритмы разложения исходного сигнала могут пропускать часть гармонической составляющей в шумовую и наоборот, что может явиться причиной возникновения искажений результирующего сигнала.

Весьма обширное описание различных подходов к модификации речевого сигнала приводится в [6]. Здесь упор делается на использовании машинного обучения для разработки алгоритмов модификации речи, хорошо приспособленных не только для применения в системах синтеза речи, но и для преобразования речи одного человека в речь другого человека (в том числе из соображений безопасности), коррекции дефектов речи и так далее. Подробно рассматриваются подходы с использованием смешанных гауссовских моделей (GMM, Gaussian Mixture Model), а также скрытых марковских моделей (HMM, Hidden Markov Model).

Из всего вышесказанного видно, что существует достаточно большое количество алгоритмов модификации характеристик речевого сигнала, а также математических моделей, на которых такие алгоритмы основаны. Однако даже при таком разнообразии достаточно гибких и мощных подходов к анализу и модификации речевого сигнала, проблема построения алгоритмов, генерирующих неотличимый от естественного речевой сигнал по-прежнему актуальна. Одной из причин этого, как отмечается в [6], является субъективность человеческого восприятия. Это обстоятельство создаёт значительные затруднения при построении адекватных математических моделей, описывающих речевой сигнал.

Подход, предлагаемый в данной статье, разрабатывается в предположении, что структура входного речевого сигнала проанализирована и описана. В таком случае алгоритмы модификации частоты основного тона и темпа работают не с речевым сигналом в целом, а его структурными элементами: аллофонами и отдельными периодами. Это должно минимизировать вносимые в результирующий сигнал искажения, а также позволяет сохранить его структуру. Модификация темпа в таком случае осуществляется естественным образом: в отличие от SOLAFS участки для дублирования или удаления выбираются не через фиксированные отрезки сигнала, а с учётом имеющегося описания сигнала. Для модификации частоты основного тона используется дискретное вейвлет-преобразование, что позволяет модифицировать составляющие сигнала с различными частотами независимо друг от друга. Данный способ в некоторой степени является развитием идеи, предложенной в [4]. Таким образом, предлагаемые в данной статье алгоритмы должны дать результаты, как минимум сопоставимые по качеству с результатами других имеющихся на сегодняшний день алгоритмов, а также обладают большим потенциалом для дальнейшей доработки и развития.

ОПИСАНИЕ АЛГОРИТМОВ

1. ОБЗОР АЛГОРИТМОВ

Предлагаемые алгоритмы осуществляют обработку оцифрованного сигнала, представленного последовательностью дискретных отсчётов. Как было сказано выше, модификации подлежат темп и частота основного тона речевого сигнала. Их преобразование осуществляется последовательно по одному аллофону. Если аллофон вокализованный, то вычисляется его итоговая длина (с учётом модификации темпа), модифицируется частота основного тона, а затем его

длина приводится к ранее вычисленной. Если же аллофон невокализованный, то модифицируется только его длина. При этом алгоритмы модификации темпа для вокализованных и невокализованных аллофонов различаются.

Следует отметить, что при модификации частоты основного тона вокализованного аллофона его длина изменяется обратно пропорционально изменению частоты основного тона за счёт модификации длин периодов. Поэтому к устранению данного эффекта привлекается алгоритм модификации темпа.

Отметим также, что алгоритм модификации темпа вокализованных звуков оперирует периодами, в силу чего длительности вокализованных аллофонов модифицируются с некоторой погрешностью. Данная погрешность компенсируется за счёт следующих аллофонов, в силу чего не накапливается.

На параметры модификации темпа и частоты основного тона накладываются некоторые ограничения. Частота основного тона должна изменяться *максимум в два раза* (то есть на +/– одну октаву). Как отмечается в [3], среднестатистический диктор способен изменять частоту основного тона своего голоса примерно в два раза, поэтому такое ограничение вполне разумно. Аналогичное ограничение задаётся для модификации темпа речи: он не должен изменяться более, чем в два раза. С учётом влияния изменения частоты основного тона на длительность сигнала алгоритм модификации темпа речи должен изменять его *максимум в четыре раза*.

Перечисленные выше ограничения не делают модификацию речевого сигнала при значениях параметров, выходящих за пределы указанных границ, невозможной, однако в таких случаях в результирующем сигнале неизбежно возникнут артефакты.

Также для корректной работы алгоритмов помимо входного речевого сигнала необходимо его описание, которое включает в себя следующие данные:

1. Номера отсчётов, находящихся на границах аллофонов.

2. Для вокализованных аллофонов – номера отсчётов, находящихся на границах периодов.

Несложно видеть, что для произвольного звукового сигнала подобное описание получить невозможно. Поэтому использование предлагаемых алгоритмов для модификации какого-либо иного (не речевого) сигнала невозможно даже при внесении в их структуру значительных изменений.

При этом следует обратить внимание на то, что аллофоническая структура речи присуща большинству существующих естественных языков, так как она обоснована в большей степени особенностями строения человеческого речевого тракта и органов слуха, нежели особенностями той или иной культуры. Поэтому предлагаемые алгоритмы способны работать с речевыми сигналами на различных языках.

2. МОДИФИКАЦИЯ ЧАСТОТЫ ОСНОВНОГО ТОНА

Частота основного тона вокализованных аллофонов модифицируется по периодам, каждый из которых раскладывается на высокочастотную и низкочастотную составляющие при помощи дискретного вейвлет-преобразования. Затем низкочастотная составляющая растягивается/сжимается в соответствии с заданным коэффициентом, а высокочастотная составляющая определённым образом «подгоняется» под низкочастотную. Сумма полученных составляющих даёт преобразованный период.

2.1. Дискретное вейвлет-преобразование

Дискретное вейвлет-преобразование заключается[1] в разложении сигнала на поддиапазоны при помощи *квадратурных зеркальных фильтров* h и g. Для того, чтобы h и g были квадратурными зеркальными фильтрами, их Фурье-преобразования $H(\omega)$ и $G(\omega)$ должны удовлетворять следующим условиям:

$$\begin{aligned} &|H(\omega)|^2 + |G(\omega)|^2 \equiv 2 \\ &\overline{H(\omega)}\, H(\omega+\pi) + \overline{G(\omega)}\, G(\omega+\pi) \equiv 0. \end{aligned} \qquad (2.1)$$

[1] В данном контексте.

Если при этом известен фильтр h такой, что

$$|H(\omega)|^2 + |H(\omega+\pi)|^2 \equiv 2, \qquad (2.2)$$

то парный к нему фильтр g получится по формуле $G(\omega) = -e^{-i\omega}\overline{H(\omega+\pi)}$ или, что то же самое, $g_k = (-1)^k h_{1-k}$.

В данном алгоритме используются фильтры Добеши 10-го порядка.

Прямое вейвлет-преобразование заключается в свёртке сигнала x с фильтрами h и g с последующим прореживанием. Будут получены последовательности a_1 и d_1:

$$\begin{aligned} a_{1r} &= \sum_s h_s x_{2r+s}, \\ d_{1r} &= \sum_s g_s x_{2r+s}. \end{aligned} \qquad (2.3)$$

Последовательность a_1 содержит информацию о составляющей сигнала x с частотами $0 \div \pi/2$, а d_1 – о составляющей с частотами $\pi/2 \div \pi$. Здесь π – частота Найквиста для сигнала x[1].

Выполнив обратное вейвлет-преобразование a_1 и d_1, получим al_1 и dl_1:

$$\begin{aligned} al_{1r} &= \sum_s h_{r-2s} a_{1s}, \\ dl_{1r} &= \sum_s g_{r-2s} d_{1s}. \end{aligned} \qquad (2.4)$$

В силу того, что h и g удовлетворяют условиям (2.1), $al_1 + dl_1 = x$. При этом al_1 – низкочастотная составляющая сигнала, а dl_1 – высокочастотная.

Отметим, что вейвлет-преобразование может быть применено и к a_1, в результате чего будут получены al_2 и dl_2 и т. д. Таким образом, исходный сигнал можно разложить в сумму составляющих:

$$x = al_n + \sum_{i=1}^{n} dl_i. \qquad (2.5)$$

Очевидно, что al_n содержит составляющую сигнала x с частотами $0 \div \pi \cdot 2^{-n}$.

Так как при представлении сигнала и фильтра в ЭВМ может хранится конечное число коэффициентов, то реализация (2.3) и (2.4) требует дополнительных пояснений. Далее фильтр, применяемый к сигналу, обозначим как h, хотя это в равной степени относится как к фильтру низких частот h, так и к фильтру высоких частот g.

Прямое вейвлет-преобразование

Фильтр h представлен последовательностью из $m + M + 1$ чисел[2] (см. рис. 1).

Аналогично, исходный сигнал x представлен $p + P + 1$ числами (рис. 2).

Дискретное вейвлет-преобразование схематично изображено на рис. 3.

Введём обозначения:

$$N = \left[\frac{P+m}{2}\right], \; n = \left[\frac{M+p}{2}\right],$$

где квадратные скобки означают целую часть.

Перепишем теперь (2.3) с учётом сдвигов индексов:

$$a_{r+n} = \sum_s h_{s+m} x_{2r+s+p},$$

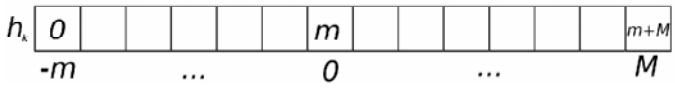

Рис. 1. Фильтр

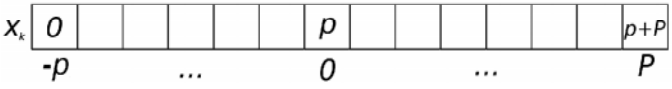

Рис. 2. Исходный сигнал

[1] Теоретическое обоснование даётся в рамках концепции ортогонального многомасштабного анализа.

[2] Числа, находящиеся под изображениями, обозначают «настоящие» индексы, а числа, написанные внутри, – индексы, используемые в ЭВМ. Далее подразумеваются именно они.

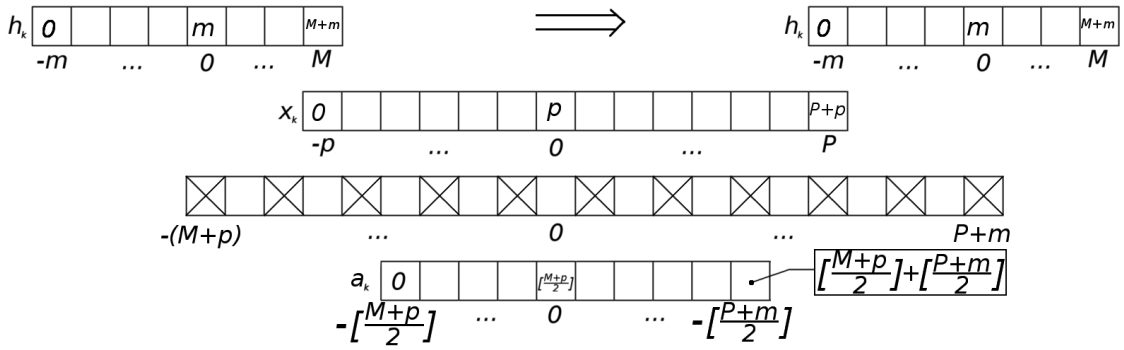

Рис. 3. Прямое вейвлет-преобразование

$s = -m, ..., M$, $r = -n, ..., N$

Сделаем замены $s + m \to s$ и $r + n \to n$:

$$a_r = \sum_s h_s x_{2(r-n)+s+p-m},$$

$s = 0, ..., M + m$, $r = 0, ..., N + n$

Учтём, что индекс при x должен быть в пределах $0 \ldots P + p$:

$$2(r-n) + s + p - m = 0, ..., P + p \Rightarrow$$
$$\Rightarrow s = m - p - 2(r-n), ..., P + m - 2(r-n).$$

В итоге имеем:

Вход: $\{h, m, M\}$, $\{x, p, P\}$

Выход: $\{a, n, N\}$

$$N = \left[\frac{P+m}{2}\right], \; n = \left[\frac{M+p}{2}\right],$$

$$a_r = \sum_s h_s x_{2(r-n)+s+p-m}, \qquad (2.6)$$

$$s = \max(0, m - p - 2(r-n)), ...,$$
$$\min(M + m, P + m - 2(r-n)),$$

$r = 0, ..., N + n$.

Обратное вейвлет-преобразование

Здесь помимо фильтра h и последовательности a должны быть заданы p и P ($p, P \geq 0$), где $-p, ..., P$ – индексы выходного сигнала. Если p, P велики, то часть выходной последовательности al будет просто заполнена нулями. Необходимость во введении p, P обусловлена тем, что при прямом вейвлет-преобразовании сигнал прореживается вдвое. Поэтому невозможно определить, имел ли он чётную длину или нет.

Перепишем (2.4) с учётом сдвигов индексов:

$$al_{r+p} = \sum_s a_{s+n} h_{r-2s+m},$$

$s = -n, ..., N$, $r = -p, ..., P$

Сделаем замены $s + n \to s$ и $r + p \to n$:

$$al_r = \sum_s a_s h_{r-p-2(s-n)+m}$$

$s = 0, ..., N + n$, $r = 0, ..., P + p$.

Учтём, что индекс при h должен быть в пределах $0 \ldots M + m$:

$$r - p - 2(s-n) + m = 0, ..., M + m \Rightarrow$$
$$\Rightarrow s = \left\lceil \frac{r - M - p}{2} \right\rceil + n, ..., \left\lfloor \frac{r + m - p}{2} \right\rfloor + n.$$

Получили:

Вход: $\{h, m, M\}$, $\{a, n, N\}$, $\{p, P\}$

Выход: al

$$al_r = \sum_s a_s h_{r-p-2(s-n)+m}$$

$$s = \max\left(0, \left\lceil \frac{r - M - p}{2} \right\rceil + n\right), ..., \qquad (2.7)$$
$$\min\left(N + n, \left\lfloor \frac{r + m - p}{2} \right\rfloor + n\right),$$

$r = 0, ..., P + p$.

2.2. МОДИФИКАЦИЯ АЛЛОФОНА

Модификация частоты основного тона вокализованного аллофона заключается в выделении из него периодов и их модификации с последующим «склеиванием». При этом в «текущий» период включается начальный «отмеченный» отсчёт, но не вклю-

чается последний. Исключение составляет последний период, в который включаются оба отсчёта. «Новая» длина каждого периода определяется как целая часть от произведения текущей длины периода и заданного коэффициента модификации. На рис. 4 показана модификация частоты основного тона аллофона с коэффициентом $10/7$ (длина умножается на 0,7). Подписанные отсчёты считаются «помеченными».

Модификация отдельного периода

На данном этапе на вход поступает один период x_T, содержащий N отсчётов, и новое количество отсчётов N', соответствующее длине преобразованного периода x'_T. Так как известно, что x_T содержит ровно один период аллофона длительностью T, то надо выделить составляющую сигнала с частотой $\rho = \frac{1}{T}$. Пусть $\pi = \frac{N}{2T}$ – частота Найквиста. Тогда с помощью дискретного вейвлет-преобразования можно выделить низкочастотную составляющую al_n с максимальной частотой $2^{-n}\pi \approx \rho$. Подставив соответствующие значения, получим:

$$\frac{1}{T} \approx 2^{-n}\frac{N}{2T} \Rightarrow 2^{n+1} \approx N \Rightarrow n \approx \log_2 N - 1.$$

Таким образом, положим

$$n = \lfloor \log_2 N \rfloor - 1. \qquad (2.8)$$

При этом, очевидно, необходимо требовать $N \geq 2$.

Обозначим al_n как AL, а $\sum_{i=1}^{n} dl_i$ – как DL. Тогда (2.5) запишется как:

$$x_T = AL + DL.$$

Очевидно, что в нашем случае DL получится как $DL = x - AL$. Это позволит ускорить работу алгоритма за счёт отказа от вычисления dl_i.

Далее низкочастотная и высокочастотная составляющие AL и DL определённым образом модифицируются (см. ниже) и преобразуются в AL' и DL'. Модифицированный период x'_T получится по формуле:

$$x'_T = AL' + DL'.$$

Модификация низкочастотной составляющей

Модификация низкочастотной составляющей AL осуществляется посредством её растяжения (сжатия) до длины в N' отсчётов. Для этого может быть использован любой алгоритм интерполяции. Опишем этот процесс подробнее.

Пусть известны значения сигнала в точках $t = 0, 1, ..., N - 1$. Необходимо сжать (растянуть) сигнал так, чтобы были известны его значения в точках $t = 0, 1, ..., N' - 1$. С помощью интерполяционного алгоритма несложно получить значения сигнала в точках $t_1 = t'\frac{N-1}{N'-1} = 0, \frac{N-1}{N'-1}, ..., N-1$. Полученные значения и дают искомую составляющую AL'.

Модификация высокочастотной составляющей

Высокочастотная составляющая не должна искажаться. Поэтому DL дублируется таким образом, чтобы «крайние» части DL' были близки к таковым у DL. Пересекаю-

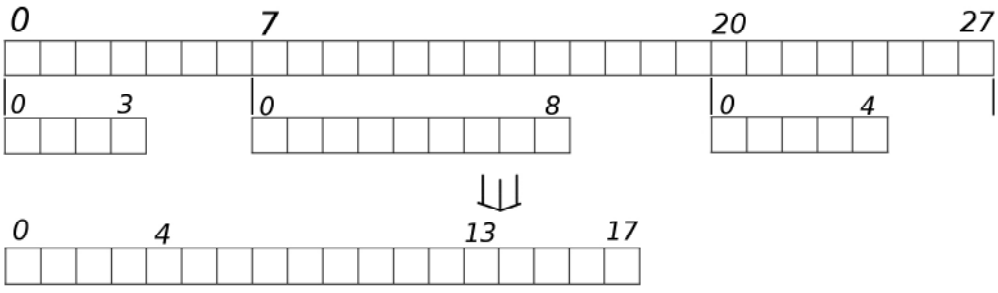

Рис. 4. Модификация частоты основного тона с коэффициентом $10/7$

щиеся части суммируются с весовой функцией. В качестве весовой функции в данном алгоритме используется половина окна Ханна:

$$\omega(k) = 0{,}5\left(1 - \cos\frac{\pi k}{K-1}\right),\ k = 0, ..., K-1,$$

где K – длина половины окна Ханна.

На рис. 5 приведено схематичное изображение модификации высокочастотной составляющей.

Если $N' \le N$, то область пересечения заполняет всю длину DL'. Отметим также, что если N' близко к $2N$ или, более того, $N' \ge N$, то в сигнале возникнут заметные помехи. В этом случае надо было бы повторять DL более двух раз.

3. МОДИФИКАЦИЯ ТЕМПА РЕЧИ

Алгоритмы удлиннения/укорочения аллофонов учитывают тот факт, что любой аллофон обладает некоторой «неоднородностью»: его частотные характеристики непостоянны во времени. Поэтому предпочтительными для модификации являются участки с наиболее стабильными частотными характеристиками. Для вокализованных и невокализованных звуков критерии «стабильности» различаются.

Во всех случаях модификация темпа речи осуществляется в два этапа:

1. Поиск наиболее «стабильных» участков.

2. Собственно модификация найденных «стабильных» участков.

Опишем алгоритмы модификации вокализованных и невокализованных звуков по отдельности.

3.1. Модификация вокализованных звуков

Оценка стабильности

Стабильность оценивается исходя из частоты основного тона сигнала. Так как известны метки границ периодов, то можно определить длину каждого из периодов, а следовательно и изменение длины периодов для каждой пары смежных периодов (иначе говоря, для каждого *стыка* двух периодов). Чем меньше абсолютное значение этого изменения, тем более стабильна частота основного тона. Положим, что исходный аллофон содержит N периодов. Символом D обозначим вектор модулей изменений длин периодов, содержащий $N-1$ чисел. При этом i-й стык считается стабильным, если

$$D_i < \alpha \frac{\sum_{j=1}^{N-1} D_i}{(N-1)}. \qquad (3.1)$$

Здесь множитель $\alpha > 0$ определяет «жёсткость» критерия стабильности. При удлиннении «нестабильные» стыки модификации не подлежат. «Стабильные» же стыки сортируются по возрастанию D_i, давая вектор D'.

Увеличение длины аллофона

При увеличении длины аллофона необходимо из двух смежных периодов с длинами l_1 и l_k получить $k \ge 2$ периодов с длинами $l_1, ..., l_k$. Два крайних периода – это исходные два периода. Периоды с номерами $2, ..., k-1$ синтезируются из 1-го и k-го периодов способом, аналогичным модификации высокочастотной составляющей при

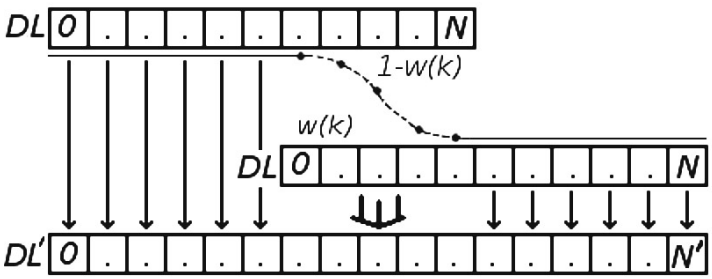

Рис. 5. Модификация высокочастотной составляющей

изменении частоты основного тона. Их длины вычисляются по формуле (3.2).

$$l_j = l_1 + \left\lfloor \frac{l_k - l_1}{k-1} \right\rfloor, j = 2, ..., k-1. \qquad (3.2)$$

Перед модификацией для каждого «стабильного» стыка производится расчёт количества добавляемых периодов (то есть k). Для этого последовательно рассчитывается новая длина аллофона при добавлении периодов с номерами из D'. Если номера из вектора D' исчерпаны, то добавляется ещё по одному периоду. Добавление периодов продолжается, пока рассчитанная длина аллофона не станет равной требуемой длине или не превзойдёт её.

Уменьшение длины аллофона

Для укорочения аллофона используется операция «схлопывания» соседних аллофонов в один, то есть взвешивания их с половиной окна Ханна точно таким же способом, каким это делается при модификации высокочастотной составляющей для изменения частоты основного тона сигнала. При этом длина «синтетического» периода принимается равной среднему арифметическому длин «схлопываемых» периодов.

Данная операция последовательно проводится для стыков с номерами из D_i (не D'_i), пока длина результирующего сигнала превосходит требуемую.

3.2. Модификация невокализованных звуков

Оценка стабильности

Для невокализованных звуков используется иной способ оценки стабильности. Если в вокализованных аллофонах ищутся наиболее стабильные стыки периодов, то в случае невокализованных аллофонов ищется последовательность отсчётов заданной длины с наименьшим среднеквадратическим отклонением (*СКО*). Будем называть такой участок *стационарным* и обозначим символом S. В данном случае наибольшую трудность представляет подбор длины модифицируемого участка. В случае удлинения и укорочения это делается по-разному. Однако так как искажения (пусть и незначительные) на краях аллофона крайне нежелательны, то первые и последние 256 отсчётов аллофона не подвергаются модификации.

Увеличение длины аллофона

В данном случае длительность стационарного участка S определяется следующим образом: задаётся начальное значение в 256 отсчётов и делится на два, пока оно больше четверти длины аллофона или $\dim S \geq l - 2 \cdot 256 = l - 512$, где l – длина аллофона.

Синтез нового удлинённого сегмента производится следующим образом:

1. Генерируется нормальный белый шум, то есть вектор нормально распределённых случайных величин W. Его длина равна

$$\dim W = dl + 2\Delta + 2\dim S .$$

Здесь $\Delta = 64$ – величина перекрытия при вставке синтезированного участка в аллофон, а $dl = l_1 - l$, где l_1 – требуемая длина модифицированного аллофона.

2. Выполняется линейная свёртка исходного сегмента S и шума W:

$$S_1 = S * W .$$

В силу свойств линейной свёртки первые и последние $\dim S$ отсчётов сигнала S_1 будут искажены. Поэтому они отбрасываются, после чего получается сигнал S'. Так как нормальный белый шум W имеет равномерное распределение в частотной области, сигналы S_1 и S' будут обладать теми же частотными характеристиками, что и исходный сегмент S, с точностью до умножения на константу.

3. Производится нормировка S':

$$S^* = \frac{\max S}{\max S'} S' .$$

4. Нормированный сигнал S^* вставляется в модифицируемый аллофон с перекрытием на $\Delta = 64$ отсчётов с каждой стороны. В зонах перекрытия он взвешивается с исходным аллофоном с половиной окна Ханна, что позволяет обеспечить «гладкость» сигнала.

В данном случае многократное дублирование S привело бы к возникновению периодической составляющей в синтезированном сегменте, что было бы причиной артефактов при уменьшении темпа речи.

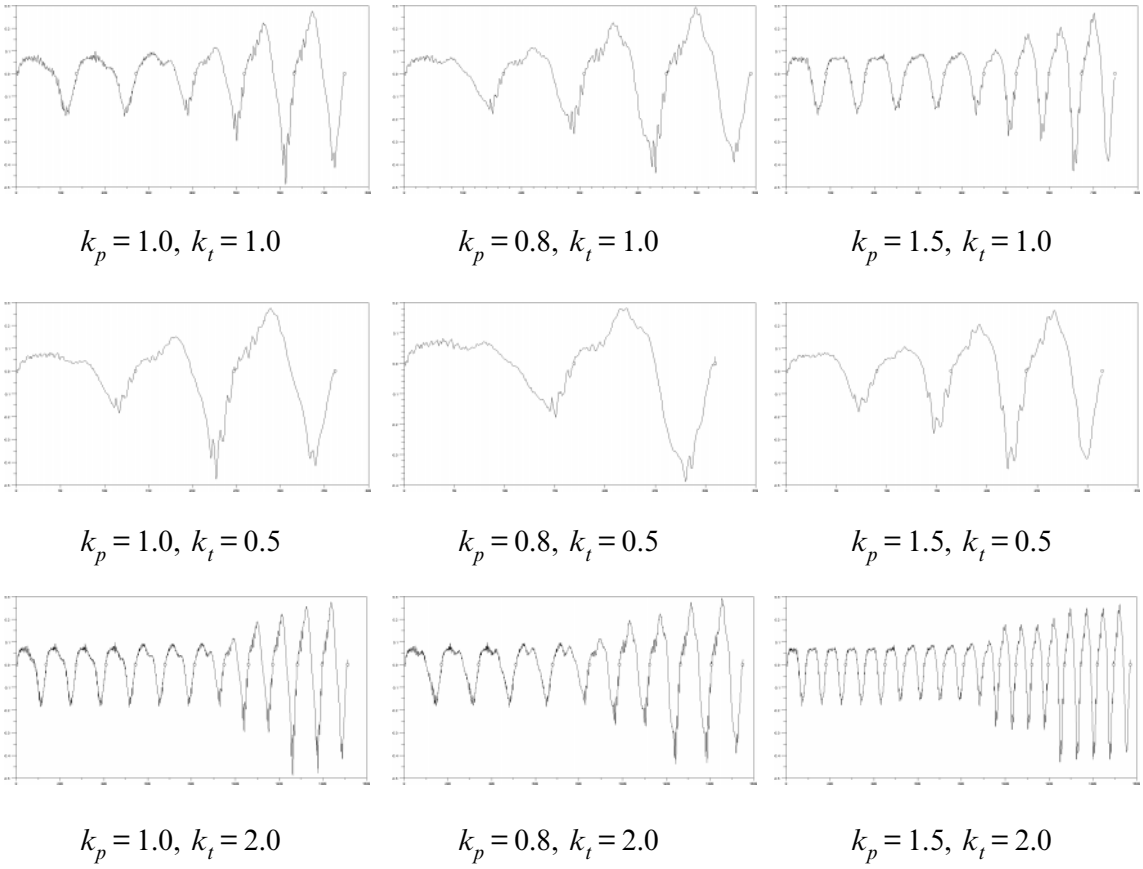

Рис. 6. Модификация невокализованного аллофоа

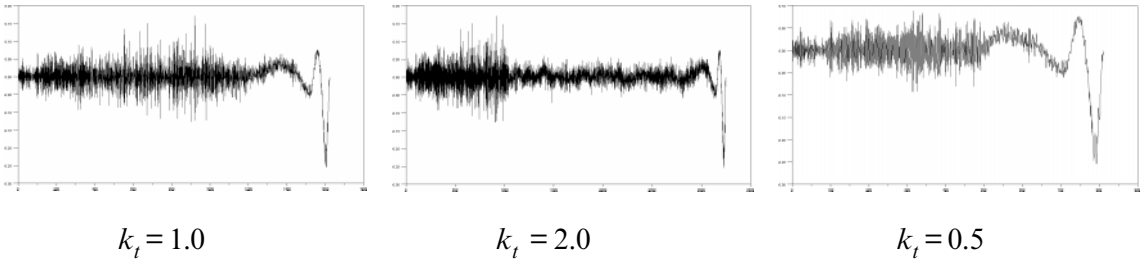

Рис. 7. Модификация невокализованного аллофона

Уменьшение длины аллофона

При укорочении аллофона ищется стационарный участок S длительностью в $dl + \Delta = dl + 64$ отсчётов. Затем выбранный участок взвешивается сам с собой способом, описанным выше. Длина нового сегмента S^* равна $\Delta = 64$ отсчётам. Затем он помещается в то место в аллофоне, откуда был извлечён исходный сегмент S. Такая операция позволяет удалить из аллофона dl отсчётов без потери «гладкости» сигнала, приводя его к требуемой длине.

Отметим, что все использованные значения длин перекрытий, «припусков» и начального значения стационарного участка подобраны эвристически и могут быть изменены.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

Модификация тестового сигнала проводилась совместно для темпа и частоты основного тона сигнала. Изменение частоты основного тона проводилось с коэффициентами 1.0, 0.8 и 1.5, а модификация темпа осуществлялась с коэффициентами 1.0, 0.5 и 2.0 (указаны с учётом изменения частоты основного тона).

Для наглядности приводятся графики только для одного вокализованного и невокализованного аллофона. Коэффициенты модификации по частоте основного тона и по длине обозначены как k_p и k_t соответственно.

Результаты модификации частоты основного тона и темпа вокализованного аллофона приведены на рис. 6. Аллофон, у которого $k_p = k_t = 1.0$, выбран из исходного сигнала.

Результат модификации длины невокализованного аллофона приведён на рис. 7.

Из приведённых выше рисунков видно, что предложенные алгоритмы дают достаточно качественные результаты. При модификации частоты основного тона соответствующим образом меняются число периодов и расстояния между пиками. При модификации длительности сохраняется форма сигнала, не наблюдается значительных артефактов, хотя можно отметить некоторое нарушение динамики амплитуд периодов сигнала при слишком сильном удлинении аллофона, что практически не сказывается на восприятии сигнала человеком. Однако при коэффициентах модификации частоты основного тона, близких к 0.5, при прослушивании наблюдаются заметные искажения сигнала.

При модификации длительности невокализованных аллофонов сохраняются их частотные и стохастические свойства. Следует отметить, что при удлинении был выбран сегмент с сильно выраженной синусоидальной составляющей, что говорит о том, что критерий выбора стационарного участка нуждается в доработке. Однако данное явление не сильно влияет на звучание результирующего сигнала.

Таким образом, разработанные алгоритмы позволяют получать достаточно качественный модифицированный сигнал без нарушения естественности речи и индивидуальных характеристик голоса (тембра). Несмотря на наличие артефактов при уменьшении частоты основного тона, можно отметить более качественную работу алгоритмов по сравнению с имеющимися аналогами.

ЗАКЛЮЧЕНИЕ

В данной статье были предложены алгоритмы модификации темпа и частоты основного тона речевого сигнала. Они отличаются высоким качеством модификации с сохранением естественности речи, чем выгодно выделяются на фоне многих имеющихся на сегодняшний день аналогов. Кроме того, предложенные алгоритмы отличаются сравнительно высоким быстродействием, так как при их работе не выполняются такие ресурсоёмкие операции, как дискретное преобразование Фурье.

Искажения при сильном уменьшении частоты основного тона связаны с особенностями частотных характеристик вокализованных звуков и человеческого слуха. Возможным способом устранения данного недостатка является использование дискретного вейвлет-преобразования для разложения сигнала не на две, а на большее количество составляющих.

При этом для работы алгоритмов необходимы дополнительные данные в виде разметки по аллофонам и периодам. Поэтому наиболее подходящей сферой применения для данных алгоритмов являются системы синтеза речи, в которых подобная информация доступна изначально. Для применения алгоритмов в других сферах (как, например, модификация естественной речи) необходимы дополнительные средства получения разметки, разработка и реализация которых представляет собой отдельную задачу.

Автор выражает искреннюю благодарность своему научному руководителю Рыбину Сергею Витальевичу, под чьим чутким и внимательным руководством писалась данная статья, а также Чистикову Павлу Геннадьевичу, чьи критические замечания и советы оказали большое влияние на данную статью и помогли значительно улучшить её качество.

Кроме того, автор выражает благодарность ООО «Центр Речевых Технологий» и его экспертам, любезно предоставившим образцы исходных сигналов с разметкой.

Литература

1. *Malah D.* Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals // IEEE Transactions on Acoustics, Speech and Signal Processing. 1979. Vol. 27, № 2. P. 121–133.

2. *Hejna D., Musicus B.R.* The SOLAFS time-scale modification algorithm // BBN Technical Reports. 1991.

3. *Чистиков П.Г., Рыбин С.В.* Проблемы естественности речевого сигнала в системах синтеза // Компьютерные инструмкнты в образовании, 2011. № 1.

4. *Faycal Y., Mesaoud B., Lotfi B.* Prosody modification of standard arabic speech using combining synchronous overlap and add with fixed-synthesis algorithm and multi level discrete wavelet transform // Journal of Computer Science, 2010. P. 392–405.

5. *Bailly G.* A parametric harmonic+noise model // Improvements in Speech Synthesis. John Wiley & Sons, 2002. P. 22–38.

6. Speech Enhancement, Modeling and Recognition-Algorithms and Applications / Ed. by S. Ramakrishnan. InTech, 2012. P. 69–94.

Abstract

This paper presents speech signal pitch and duration modification algorithms. Approach to pitch modification is based on discrete wavelet transform. Emphasis is put on application of proposed algorithms in speech synthesis systems. In addition, brief review of algorithms and mathematical models developed to solve the problem of speech modification is stated. Results of experimental research are presented, main advantages and disadvantages of proposed algorithms are stated and the way of their improvement is outlined.

Keywords: pitch modification, rhythm modification, speech synthesis, discrete wavelet transform.

Олейник Андрей Леонидович,
бакалавр прикладной математики и информатики, студент 1 курса магистратуры СПбГЭТУ «ЛЭТИ», кафедра МО ЭВМ ФКТИ,
olen19@yandex.ru