ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ЭКСПЕРТНЫЕ СИСТЕМЫ
(Вычислительные системы)
1996 год Выпуск 157

УДК 621.31:534.4

АЛГОРИТМ ОЦЕНИВАНИЯ ТРАЕКТОРИИ ЧАСТОТЫ ОСНОВНОГО ТОНА [1]

А.В.Кельманов, С.А.Хамидуллин

Введение

Как известно, устной речи свойственно чередование звонких (вокализованных, голосовых или тональных) и глухих (невокализованных, неголосовых или нетональных) звуков. Образование вокализованных звуков речи обусловлено колебаниями голосовых связок. Длительность одного цикла колебаний голосовых связок называют периодом основного тона, а обратную величину — частотой основного тона. При генерации невокализованных звуков речи голосовые связки практически неподвижны, так что с достаточной для большинства приложений точностью можно считать частоту их колебаний равной нулю, а период колебаний — бесконечно большим.

Под траекторией частоты (или периода) основного тона обычно понимают зависимость этой частоты (или периода) от времени. Указанную зависимость часто также называют мелодическим контуром.

[1]Работа выполнена в рамках проекта №94-01-00169-а, поддержанного Российским фондом фундаментальных исследований

Оценивание или выделение траектории частоты основного тона является одной из важнейших задач в области автоматической обработки речевых сигналов. Контур частоты основного тона необходим для идентификации и верификации диктора по голосу и полезен при определении границ фраз в слитной устной речи. При передаче сжатой кодированной речи по каналам связи от точности выделения траектории основного тона зависит разборчивость и натуральность речи, восстановленной на приемном конце тракта связи. Мелодический контур фразы необходим в системах обучения иностранным языкам для контроля правильности произношения. Траектория частоты основного тона связана с эмоциональным и физическим состоянием человека, акцентом и т.д.

В работе изложен алгоритм оценивания траектории частоты основного тона по речевому сигналу, представленному в виде дискретной последовательности отсчетов непрерывной функции, которая описывает изменение электрического сигнала во времени на выходе микрофона или микрофонного усилителя.

1. Сущность проблемы и особенности решения

Почти все известные алгоритмы оценивания траектории частоты колебаний голосовых связок основаны на подходе, допускающем аппроксимацию речевого сигнала локально-стационарной моделью. Такой же подход принят в данной работе. В рамках этого подхода речевой сигнал представляется в виде последовательности участков локальной стационарности. При этом считается, что каждый участок локальной стационарности сигнала представим в виде процесса, наблюдаемого на выходе некоторой гипотетической линейной динамической системы. Предполагается, что изменение параметров системы и характеристик процесса возможно только при переходе от одного участка локальной стационарности к другому. Платой за простоту локально-стационарной модели является множество мешающих параметров, которые затрудняют точное и надежное оценивание контура

частоты основного тона. Тем не менее, аппроксимация речевого сигнала более сложной моделью (применительно к задаче выделения мелодического контура) большинством авторов признана нецелесообразной из-за ощутимого увеличения трудоемкости алгоритмов, несоизмеримого с ожидаемым выигрышем по точности и надежности.

Фактически к мешающим факторам относятся те объективно существующие явления, сопровождающие процесс речеобразования, которые не учитываются или слабо учитываются принятой локально-стационарной моделью сигнала. Среди них, в частности, можно указать отличие сигнала, формируемого голосовыми связками, от идеальной периодической последовательности. С одной стороны, отличие от периодичности проявляется в том, что два последовательных по времени импульса голосовых связок могут иметь весьма слабое сходство. С другой стороны, временной интервал между импульсами не является константой, а изменяется в некотором диапазоне, ширина которого зависит от диктора. Другим фактором являются методологические, формальные и вычислительные трудности определения моментов времени начала и завершения ненаблюдаемых импульсов голосовых связок. Для точного нахождения этих моментов времени необходимо решение обратной задачи, т.е. оценивание входного сигнала (импульсов голосовых связок) по наблюдаемому выходному (речевому сигналу) в условиях отсутствия значительного объема информации о передаточной характеристике речевого тракта. Следует также упомянуть о слабо изученном влиянии речевого тракта на форму импульсов голосовых связок. Это влияние особенно заметно на участках сигнала, соответствующих переходу от одного звука к другому. Наконец, мешающим фактором является отсутствие в устной речи четких границ между звуками. Непрерывному (плавному) переходу от звука к звуку соответствует непрерывное изменение характеристик обрабатываемого речевого сигнала. Для оценивания же контура частоты основного тона необходимо принять дискретное решение об участке сигнала:

вокализованный/невокализованный (звонкий/глухой, голосовой/неголосовой или тональный/нетональный).

Ограничившись перечисленным набором мешающих факторов, отметим, что дополнительные трудности оценивания траектории частоты основного тона возникают, когда обрабатываемый сигнал содержит помехи в виде аддитивных шумов и нелинейных амплитудных искажений.

Проблема выделения траектории частоты основного тона затрагивалась многими авторами. Интенсивность исследований по решению этой проблемы не ослабевает, так как через некоторое время после публикации очередного нового алгоритма обычно оказывается, что найденное решение не учитывает какие-либо вновь обнаруженные или ранее мало изученные особенности речевых сигналов. Подобная ситуация является следствием процесса постоянного уточнения свойств речевых сигналов и математических моделей речеобразования. Здесь следует заметить, что из-за отсутствия адекватной модели речеобразования построение алгоритма, как правило, сопровождается использованием различных эвристических процедур и приемов, дополняющих формализованное математическое решение. Применение эвристических приемов позволяет хотя бы частично компенсировать имеющуюся неадекватность модели и сделать алгоритм более устойчивым к мешающим факторам. Забегая вперед, отметим, что предлагаемый в данной работе алгоритм, как и другие известные алгоритмы, также опирается на некоторые эвристические решения.

Цель данной работы состоит в построении и обосновании вычислительного алгоритма, обеспечивающего меньшую, по сравнению с известными алгоритмами, погрешность оценивания траектории частоты основного тона за счет более точного учета квазипериодичности речевого сигнала на вокализованных участках речи.

Предлагаемый алгоритм относится к классу алгоритмов обработки сигнала во временной области. В отличие от известных [1–6] алгоритмов, постулирующих пери-

одичность сигнала на коротких участках (кадрах анализа) вокализованной речи, изложенный алгоритм опирается на квазипериодичность импульсов голосовых связок по времени, что позволяет повысить точность оценивания. Искомые оценки среднего периода основного тона в анализируемом кадре находятся за два прохода, каждый из которых состоит в решении дискретной экстремальной задачи. Двухпроходовый принцип обработки сигнала уменьшает число ошибок на участках сигнала, соответствующих переходам от одного звука к другому. Наиболее близким к предлагаемому (по используемому подходу и существу решения) является алгоритм, первоначально заявленный в [7] и подробно описанный в [8].

По общепринятой классификации приведенное решение задачи относится к апостериорно-последовательным методам решения подобных задач (т.е. к методам обработки сигнала в режиме скользящего кадра фиксированной длины) и опирается на принцип максимального правдоподобия. Максимизация функции правдоподобия сведена к минимизации аддитивного функционала, которая осуществляется методом динамического программирования.

Для принятой квазипериодической модели вокализованных участков речевого сигнала в работе приведены формулы пошаговой оптимизации, составляющие сущность вычислительного алгоритма, учитывающие специфику решаемой задачи и обеспечивающие получение траектории основного тона. Дана оценка трудоемкости алгоритма. Вместе с вычислительным алгоритмом в работе приведены результаты обработки речевых сигналов.

2. Постановка задачи

Пусть $x(t)$ ($x \in R$, $t \in R$, R — числовая прямая) — функция, описывающая изменение речевого сигнала во времени, а $x_n = x(n\tau_s)$, $n = 0, \pm 1, \pm 2, \ldots$, — дискретные значения (отсчеты) этой функции (сигнала), взятые через равные промежутки времени $\tau_s = 1/F_s$ (F_s — частота дискретизации сигнала). Образуем вектор $\mathbf{X} = (x_0, x_1, \ldots, x_{N-1})$,

компонентами которого являются N расположенных друг за другом отсчетов сигнала, и последовательность $\mathbf{X}_k = \left(x_0(k), x_1(k), \ldots, x_{N-1}(k)\right), k = 0, 1, 2, \ldots$, из таких векторов, где k обозначает момент времени, в который наблюдается вектор $\mathbf{X}$. Вектор $\mathbf{X}_k$, очевидно, является участком сигнала длины $N\tau_s$, взятым или выделенным в момент времени k. Индекс k у вектора $\mathbf{X}_k$ будем опускать в тех случаях, когда при чтении текста не возникает неоднозначной интерпретации записи. Каждый из выделенных в некоторый момент времени участков сигнала будем также называть анализируемым кадром.

Будем говорить, что анализируемый кадр пуст, если норма вектора $\mathbf{X}$ равна нулю, т.е. $\|\mathbf{X}\| = 0$, и не пуст, если $\|\mathbf{X}\| > 0$. Любой непустой кадр определим как участок сигнала, вычлененный из бесконечной квазипериодической (почти периодической) импульсной последовательности x_n, $n = 0, \pm 1, \pm 2, \ldots$, вида:

$$x_n = \begin{cases} u_n - n_i, & n = n_i, n_i + 1, \ldots, n_i + q - 1, \\ 0, & n = n_i + q, \ldots, n_{i+1} - 1, \end{cases} \qquad (1)$$

$$i = 0, \pm 1, \pm 2, \ldots,$$

где u_n, $n = 0, 1, \ldots, q-1$, — числовая детерминированная последовательность или эталонный импульс $\mathbf{U} = \left(u_0, u_1, \ldots, u_{q-1}\right)$ длины $q \geq 1$, обладающий свойством:

$$0 < \sum_{n=0}^{q-1} u_n^2 < \infty; \quad u_n = 0, \quad n < 0, \quad n > q-1; u_0 \neq 0, \quad u_{q-1} \neq 0, \quad (2)$$

n_i, $i = 0, \pm 1, \ldots$, — почти периодическая последовательность моментов времени начала импульсов такая, что

$$q \leq T_{min} \leq n_i - n_{i-1} \leq T_{max}. \qquad (3)$$

В неравенстве (3) через T_{min} и T_{max} обозначены соответственно минимальное и максимальное расстояния между

двумя последовательными импульсами. Величина

$$T_0 = \lim_{M\to\infty} \frac{1}{2M+1} \sum_{i=-M}^{M} (n_i - n_{i-1}) \qquad (4)$$

есть средний интервал (период) повторения импульсов в бесконечной последовательности, т.е. среднее число дискретных отсчетов сигнала между импульсами, а

$$\tau_0 = \tau_s T_0, \quad F_0 = \frac{1}{\tau_0} \quad - \qquad (5)$$

средний период и частота повторения импульсов в этой последовательности, выраженные в единицах времени и частоты.

Будем считать, что выделенный кадр $\mathbf{X} = \left(x_0, x_1, \ldots, x_{N-1}\right)$ длины N содержит неизвестное число $M > 2$ импульсов, среди которых имеется хотя бы один полный импульс, т.е. импульс фиксированной длины q, не разбитый на две части левой или правой границами кадра. Последовательности $n_1, n_2, \ldots, n_M$ поставим в соответствие начала M импульсов из выделенного кадра.

Определим средние значения числа отсчетов T_{0N} и длительности интрервала τ_{0N} между импульсами, а также среднюю частоту F_{0N} импульсов в кадре:

$$T_{0N} = \frac{1}{M-1} \sum_{i=2}^{M} (n_i - n_{i-1}), \quad \tau_{0N} = \tau_s T_{0N}, \quad F_{0N} = \frac{1}{\tau_{0N}}. \qquad (6)$$

Все непустые кадры разобъем на 2 класса: 1) вокализованные, 2) невокализованные. К вокализованным отнесем кадры, для которых при $M > 2$

$$\delta = \frac{1}{M-1} \sum_{i=2}^{M} |n_i - n_{i-1} - T_{0N}| \le \theta_1, \qquad (7)$$

т.е. те кадры, в которых средний разброс δ длины интервалов между началами соседних импульсов не больше некоторого порога θ_1. Участки сигнала, для которых

неравенство (7) не справедливо, отнесем к невокализованным. Для невокализованных участков сигнала и пустых кадров положим $T_{0N} = \infty$, $\tau_{0N} = \infty$ и $F_{0N} = 0$.

Предположим, что наблюдаемый временной ряд (сигнал) искажается аддитивной помехой, т.е.:

$$y_n = x_n + \varepsilon_n, \quad n = 0, \pm 1, \pm 2, \ldots, \tag{8}$$

где ε_n — гауссовская последовательность независимых, одинаково распределенных случайных величин, имеющих нулевое математическое ожидание $M\varepsilon_n = 0$ и известную дисперсию $M\varepsilon_n^2 = \sigma^2 < \infty$.

Пусть наблюдается последовательность зашумленных кадров, т.е. выборок $\mathbf{Y}_k = \left(y_0(k), y_1(k), \ldots, y_{N-1}(k)\right)$, $k = 0, 1, \ldots$, длины N. Задача состоит в том, чтобы оценить траекторию частоты основного тона как последовательность $F_{0N}(k)$ средних значений частоты колебаний импульсов голосовых связок в анализируемых кадрах.

3. Алгоритм оценивания

3.1. *Ядро алгоритма.* Для получения траектории частоты основного тона воспользуемся методом максимального правдоподобия. Из (1)–(3) и (8) следует, что функция правдоподобия для наблюдаемых выборочных данных имеет вид:

$$\mathrm{Ln}\rho\left(\mathbf{Y}; \mathbf{U}, \sigma^2, n_1, n_2, \ldots, n_{M-1}, n_M\right) =$$

$$= -\frac{N}{2}\mathrm{Ln}(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left(y_n - \sum_{i=1}^{M} u_n - n_i\right)^2. \tag{9}$$

Неизвестное число M импульсов, попавших в анализируемый кадр длины N, лежит в интервале $[M_{min}, M_{max}]$, границы которого M_{min} и M_{max}, очевидно, являются функциями от величин N, T_{min}, T_{max}, q.

Предположим что импульс $\mathbf{U}$ задан. Обозначим через $\{n_i\}$ множество возможных моментов начала i-го импульса, $i = 1, 2, \ldots, M$. Тогда, учитывая, что по условию зада-

чи σ^2 известна, для нахождения оценок моментов времени начала импульсов и числа этих импульсов в выборке достаточно максимизировать (9) на семействе декартовых множеств

$$Z = \{\{n\}^M : \{n\}^M = \{n_1\} \times \{n_2\} \times \ldots \times \{n_{M-1}\} \times \{n_M\}, \\ M \in [M_{min}, M_{max}]\}$$

всевозможных последовательностей моментов времени начала импульсов. Найденные оценки $\hat{M}$ и $\left(\hat{n}_1, \hat{n}_2, \ldots, \hat{n}_{\hat{M}-1}, \hat{n}_{\hat{M}}\right)$ после их подстановки в (6) и (7) дадут искомую траекторию.

Нетрудно заметить, что задачу получения оценок $\hat{M}$ и $\left(\hat{n}_1, \hat{n}_2, \ldots, \hat{n}_{\hat{M}-1}, \hat{n}_{\hat{M}}\right)$ можно трактовать как задачу проверки гипотез о среднем $\mathbf{X}$ случайного вектора $\mathbf{Y}$. В этой трактовке число проверяемых гипотез равно мощности множества Z. Если положить

$$S(n_1, n_2, \ldots, n_{M-1}, n_M) = \\ = \sum_{n=0}^{N-1} \left(y_n - \sum_{i=1}^{M} u_{n-n_i}\right)^2, \qquad (10)$$

где $M = M_{min}, M_{min}+1, \ldots, M_{max}$, то из (9) следует, что для принятия решения по критерию максимального правдоподобия необходимо минимизировать сумму (10) на семействе множеств Z. Таким образом, искомые оценки моментов времени начала импульсов и их число находятся по правилу:

$$\left(\hat{n}_1, \hat{n}_2, \ldots, \hat{n}_{\hat{M}-1}, \hat{n}_{\hat{M}}; \hat{M}\right) = \\ = \operatorname{Arg}\min_{Z} S(n_1, n_2, \ldots, n_{M-1}, n_M). \qquad (11)$$

Обозначив

$$A(n_i) = \sum_{n=0}^{N-1} u_{n-n_i}[u_{n-n_i} - 2y_n], \quad i = 1, \ldots, M, \qquad (12)$$

сумму (10) можно представить в виде:

$$S(n_1, n_2, \ldots, n_{M-1}, n_M) = \sum_{n=0}^{N-1} y_n^2 + \sum_{i=1}^{M} A(n_i). \qquad (13)$$

Поэтому задача получения оценок сводится к минимизации аддитивного функционала

$$D(n_1, n_2, \ldots, n_{M-1}, n_M) = \sum_{i=1}^{M} A(n_i) \qquad (14)$$

на семействе Z всевозможных последовательностей моментов времени с ограничениями на целочисленные переменные n_i в виде неравенств:

$$n_1 \geq 1-q, n_M \leq N-1, T_{min} \leq n_i - n_{i-1} \leq T_{max}, i = 2, \ldots, M, \qquad (15)$$

которые вытекают из (3).

Следуя [9], для решения экстремальной задачи (14) при ограничениях (15), воспользуемся методом динамического программирования. Для каждого $n \in [-q+1, N+T_{min}-1]$ определим множество:

$$\Gamma(n) = \{m: \ max[-q+1, n-T_{max}] \leq m \leq n - T_{min}\}$$

и, в соответствии с принципом оптимальности, организуем многошаговый процесс минимизации в виде:

$$\left.\begin{aligned} D_n &= \min_{m \in \Gamma(n)} \{D_m + A(m)\}, \\ I(n) &= \arg \min_{m \in \Gamma(n)} \{D_m + A(m)\}, \\ n &= -q + T_{min}, \ldots, N + T_{min} - 1, \end{aligned}\right\} \qquad (16)$$

$$D_N = D_{min} = \min_{N \leq n \leq N + T_{min} - 1} D_n, \qquad (17)$$

где через D_n и $I(n)$ обозначены минимальное значение функционала и указатель минимума на n-м шаге. Начальными условиями для вычислений по формулам (16) и (17) являются нулевые значения:

$$D_n = 0, \ I(n) = 0; \ n = -q+1, \ldots, T_{min} - q - 1, \qquad (18)$$

а также значения $A(n), n = -q+1, \ldots, N$, подсчитанные по формуле (12) перед началом оптимизации.

Число импульсов и их расположение в выборке определяется рекуррентными вычислениями в обратном порядке по указателю $I(n)$ оптимального пути:

$$\left.\begin{aligned} m_0 &= \arg \min_{N \le n \le N + T_{min} - 1} D_n, \\ m_i &= I(m_{i-1}), \qquad i = 1, 2, \ldots \end{aligned}\right\} \tag{19}$$

Вычисления заканчиваются при таком шаге $i = r$, что $I(m_r) = 0$. В результате получаем последовательность $m_r, m_{r-1}, \ldots, m_1$ такую, что $(\hat{n}_1, \hat{n}_2, \ldots, \hat{n}_{\hat{M}-1}, \hat{n}_{\hat{M}}) = (m_r, m_{r-1}, \ldots, m_1)$. Величина r дает оценку $\hat{M}$ числа импульсов (включая неполные), попавших в выборку. При этом если $\hat{n}_1 \in \{-q+1, \ldots, -1\}$, то в начале выборки имеется неполный импульс. Если $\hat{n}_{\hat{M}} \in \{N-q+1, \ldots, N-1\}$, то неполный импульс имеется в конце выборки.

Таким образом, если импульс U задан, то получение оценок моментов начала импульсов в кадре анализа обеспечивают вычисления по формулам (12) и (16)-(19). На практике при обработке речевых сигналов импульс U всегда неизвестен. Чтобы воспользоваться описанной выше процедурой, требуется определить недостающие априорные данные. Для нахождения этих данных поступим следующим образом.

3.2. *Задание длины кадра, формы и длины импульса.* Установим длину анализируемого кадра исходя из необходимости соблюдения двух разумных, но противоречивых требований. С одной стороны, кадр должен перекрывать по крайней мере три импульса (два из которых могут быть неполными). Это требование вытекает из формул (6) и (7). При меньшем числе импульсов в кадре разброс δ, вычисляемый по формуле (7), будет всегда равен нулю. В этом случае классифицировать кадр на вокализованный и невокализованный будет невозможно и задача анализа квазипериодичности теряет смысл. С другой стороны, анализируемый кадр должен иметь возможно меньшую

длину, чтобы внутри кадра с достаточной для практики точностью можно было считать форму импульса неизменной. Минимальная длина кадра, покрывающего не менее трех импульсов, очевидно, равна $3T_{max} - q + 1$.

Если решение о квазипериодичности сигнала принимать по нескольким последовательным перекрывающимся кадрам (например, по двум — текущему и предыдущему), то длину анализируемого кадра можно уменьшить. Для этого необходимо, чтобы каждый кадр перекрывал не менее двух импульсов. При этом величина разброса δ может быть вычислена по оценкам моментов начала импульсов в соседних кадрах. Минимальная длина кадра, покрывающего не менее двух импульсов, равна $2T_{max} - q + 1$.

Из-за отсутствия информации о форме и длине импульса анализ квазипериодичности сигнала становится невозможным при попадании в кадр длины $2T_{max} - q + 1$ только двух неполных импульсов. В связи с этим, длину анализируемого кадра следует увеличить так, чтобы среди импульсов, попавших в кадр, был хотя бы один полный. Наименьшая длина кадра, покрывающего хотя бы один полный импульс, равна $T_{max} + q - 1$. Таким образом, в случае принятия решения по двум или более последовательным перекрывающимся кадрам оптимальная длина N анализируемого кадра равна $\max(2T_{max} - q + 1,\ T_{max} + q - 1)$. В приведенное выражение входит неизвестная длительность q импульса. Уточним выражение для размера кадра после решения вопроса о задании длины и формы импульса.

Механизм образования вокализованных участков речи состоит в возбуждении акустической системы, объединяющей ротовую и носовую полости, импульсами голосовых связок. После воздействия на акустическую систему очередного импульса в ней возникают свободные затухающие колебания. Наблюдаемый речевой сигнал (осциллограмма) для каждого цикла колебаний голосовых связок содержит два характерных участка. На первом начальном участке происходит нарастание амплиту-

ды сигнала, на втором — ее спад.

Два соседних отрезка речевого сигнала, соответствующие двум последовательным циклам колебаний голосовых связок, обычно наиболее похожи на тех участках, где амплитуда и энергия сигнала достигают максимальных значений. При наличии помех эти участки имеют наибольшее отношение сигнал/помеха. Меньшее сходство наблюдается на тех участках сигнала, соответствующих началу и окончанию цикла колебаний голосовых связок, где амплитуда и энергия сигнала невелики. На этих участках отношение сигнал/помеха заметно ниже, чем на участках с большей амплитудой и энергией. В связи с этим, для алгоритмического анализа квазипериодичности в качестве импульса наиболее подходят такие отрезки речевого сигнала, не превосходящие по длине минимальную длительность одного цикла колебаний голосовых связок, на которых сигнал имеет наибольшую амплитуду и энергию. Подчеркнем, что для выделения траектории основного тона не требуется определение истинных моментов времени начала импульсов голосовых связок.

Подходящим критерием для выделения в каждом кадре отрезков сигнала с наибольшей амплитудой и энергией является максимум суммы

$$V_n^2 = \sum_{j=n}^{n+q-1} y_j^2, \quad n = 0, \ldots, N-q, \tag{20}$$

квадратов нескольких последовательных отсчетов сигнала. Вопрос лишь в том, какова длина q отрезка сигнала, на котором следует оценивать сумму квадратов (20), т.е. какова длина отрезка сигнала, который можно взять в качестве импульса.

В соответствии с ограничениями (3) длина q импульса не может превышать величины T_{min}. Поэтому и длина отрезка сигнала, задаваемого как импульс, не может быть больше T_{min}. С другой стороны, не следует задавать длину отрезка сигнала, выбираемого в качестве им-

пульса, меньше, чем $]T_{min}/2[+1$. В противном случае, может возникнуть ситуация, когда истинный импульс будет содержать два или более идентичных участка, один из которых был задан в качестве импульса. В этой ситуации ограничения (3) окажутся нарушенными (так как два, следующих друг за другом идентичных отрезка сигнала окажутся расположенными на расстоянии, меньшем T_{min}), что, в конечном итоге, приведет к невозможности однозначного принятия решения об оцениваемых моментах времени. Таким образом, ширина N анализируемого кадра и длина q задаваемого импульса (при получении оценок по двум соседним кадрам) связаны соотношениями:

$$\left.\begin{aligned} &]\frac{T_{min}}{2}[+1 \le q \le T_{min}, \\ &N \ge \max\left(T_{max}+q-1,\ 2T_{max}-q+1\right). \end{aligned}\right\} \qquad (21)$$

Выберем q и N в соответствии с (21) и для задания формы эталонного импульса в каждом анализируемом кадре положим

$$u_j \equiv y_{n^*+j}, \quad j=0,\ldots,q-1, \qquad (22)$$

где

$$n^* = \arg \max_{n \in [0,N-q]} V_n^2. \qquad (23)$$

Теперь все недостающие величины определены и обработка сигнала может осуществляться при помощи процедуры, описанной в п.3.1.

Здесь следует заметить, что теоретически вместо задания эталонного импульса по приведенному правилу следовало бы перебрать всевозможные положения эталонного импульса в кадре в комбинации со всевозможными длинами этого импульса, для каждой из комбинаций найти квазипериодическую последовательность и по минимальному значению функционала (14), подсчитанному для всех комбинаций, указать оптимальное решение, включающее наилучшую форму импульса и его длину.

Однако подобный подход практически нереализуем из-за высокой трудоемкости.

Как отмечено выше, при оценивании траектории основного тона неадекватность локально-стационарной модели сигнала, которая проявляется в слабом учете возможных изменений формы импульса внутри одного анализируемого кадра, и недостающую априорную информацию об импульсах голосовых связок приходится компенсировать дополнительными мерами, повышающими точность алгоритма. Некоторые из подобных мер и эвристических приемов представлены ниже.

3.3. *Повторное оценивание с измененной формой импульса.* Задание импульса по правилу (22), (23) может оказаться неудачным в тех случаях, когда анализируемый кадр содержит импульсы заметно отличающиеся по форме. Подобные кадры имеют место, например, при переходе от одного звука к другому. Выбранный в качестве импульса отрезок сигнала может оказаться нетипичным. В этой ситуации целесообразен повторный анализ квазипериодичности с измененной формой импульса.

Предположим, что в результате первой прогонки алгоритма в анализируемом кадре обнаружено два или более полных импульса. Тогда для повторной прогонки можно использовать один из обнаруженных импульсов. Если в кадре всего два полных импульса, один из которых задан, а другой обнаружен, то для второго прохода используем обнаруженный импульс. Если число импульсов в кадре больше двух, то для повторного оценивания используем тот из обнаруженных импульсов, который наиболее близок по времени к середине анализируемого кадра. Если кадр содержит только один полный (заданный) импульс, а обнаруженные импульсы оказываются неполными, повторную прогонку не проводим.

Результатом повторного анализа квазипериодичности кадра являются новые оценки моментов возникновения импульсов и значение разброса δ. Из двух найденных наборов оценок выберем тот, которому соответствует наименьшая величина разброса δ. Если после второго прохо-

да оказывается, что в кадре имеется лишь один полный импульс, а при первом проходе число полных импульсов было больше единицы, то в качестве оценок используем результаты первого прохода.

3.4. *Обработка неполных импульсов.* Неполные импульсы, обнаруженные на границах кадра, имеют меньшую, чем полные, длительность и энергию. Поэтому на границах кадра точность оценок моментов времени возникновения импульсов оказывается ниже, чем внутри кадра [9] и, как следствие, неполные импульсы вносят более весомую, чем полные, погрешность в оценку среднего интервала между импульсами. В этой связи, при подсчете среднего интервала между импульсами представляется целесообразным отбрасывать неполные импульсы, если их длительность или энергия оказываются меньше некоторых заданных порогов. Из-за отсутствия необходимой априорной информации об импульсах теоретическое определение этих порогов оказывается проблематичным. Поэтому значения этих порогов устанавливаются эмпирическим путем. Удовлетворительные результаты были получены при отбрасывании тех неполных импульсов, у которых длительность меньше четверти длины q полного импульса или энергия (сумма квадратов отсчетов) меньше третьей части энергии V_{n*}^2 полного импульса.

3.5. *Дополнительный критерий.* Классификация анализируемого кадра на голосовой и неголосовой по критерию (7) невозможна, когда число M импульсов, обнаруженных в кадре, равно двум. Для кадров, содержащих два импульса, разброс δ всегда равен нулю независимо от того, какой в действительности участок сигнала анализируется. В исходной постановке задачи случай $M = 2$ исключен. Однако при обработке сигнала в кадре минимальной длины N, равной $\max\left(2T_{max} - q + 1,\ T_{max} + q - 1\right)$, этот случай допустим.

Чтобы обработка сигнала с минимальной шириной кадра была возможна, для кадров, содержащих всего два импульса, следует использовать какой-либо дополни-

тельный критерий, позволяющий разделять голосовые и неголосовые кадры. Если в соответствии с этим критерием анализируемый кадр относится к голосовым, то искомые оценки находятся по формулам (6).

При $M = 2$ для классификации кадров на вокализованные и невокализованные воспользуемся статистикой

$$\eta = \frac{\sum\limits_{n=0}^{N-1} \left(y_n - \hat{x}_n \right)^2}{\sum\limits_{n=0}^{N-1} y_n^2}, \tag{24}$$

где через

$$\hat{x}_n = \begin{cases} u_{n-\hat{n}_i}, & n = \hat{n}_i, \hat{n}_i + 1, \ldots, \hat{n}_i + q - 1, \\ 0, & n = \hat{n}_i + q, \ldots, \hat{n}_{i+1} - 1, \end{cases} \tag{25}$$

$$i = 1, 2, \ldots, \hat{M}, \quad n = 0, 1, \ldots, N - 1,$$

обозначена оценка квазипериодического сигнала, построенная алгоритмом. Возможность применения статистики (24) установлена в результате численного моделирования и обработки реальных речевых сигналов. Решение вокализованный/невокализованный принимается в результате сравнения η с порогами, величины которых зависят от вида импульсов, обнаруженных в кадре. Если $\eta > \theta_2$ и оба импульса полные, то анализируемый отрезок сигнала относится к неголосовым. Если же из двух найденных импульсов один неполный, то кадр относится к неголосовым, когда $\eta > \theta_3$. При этом пороги связаны неравенством: $\theta_2 > \theta_3$.

3.6. *Коррекция оценок.* Для уменьшения ошибок оценивания траектории частоты основного тона обычно используется известное свойство процесса речеобразования, состоящее в том, что "переключение" звуков речи — вокализованный/невокализованный — не может происходить слишком часто. Поэтому, если в окружении нескольких соседних вокализованных кадров, обнаруженных алгоритмом, оказывается один невокализованный, то реше-

ние о последнем можно скорректировать в пользу вокализованного, и наоборот. Процедуры подобной коррекции описаны во многих работах. В предлагаемом алгоритме используется коррекция первичных оценок по четырем последовательным кадрам [5,10]. Скорректированная оценка частоты основного тона вычисляется как среднее значение оценок, найденных в двух предшествующих и одном последующем кадрах.

3.7. *Общая схема вычислений.* Для каждого анализируемого кадра проверяется гипотеза о наличии/отсутствии полезного сигнала. Считается, что кадр пуст, если сумма квадратов отсчетов сигнала из этого кадра меньше заданного порога. По существуб принятие решения об отсутствии полезного сигнала базируется на статистическом критерии χ^2 для проверки гипотезы о том, что анализируемый кадр содержит последовательность независимых одинаково (нормально) распределенных случайных величин с нулевым средним и известной дисперсией. Все пустые кадры отождествляются с невокализованными.

Для непустых кадров алгоритмическим путем по правилу (22), (23) задается эталонный импульс и проводится анализ квазипериодичности. Для этого используются процедуры и методы обработки сигнала, описанные в п.3.1.–3.7.

3.8. *Трудоемкость алгоритма.* Опираясь на результаты работы [9] (теорема 1), трудоемкость ядра алгоритма можно оценить как величину $O[(N+q)(q+T_{max}-T_{min})]$. Такие же затраты по времени требуются при повторной прогонке. Нетрудно заметить, что трудоемкость дополнительных вычислений растет линейно с увеличением ширины анализируемого кадра. Поэтому трудоемкость обработки одного кадра есть величина $O[(N+q)(q+T_{max}-T_{min})]$ и, таким образом, затраты по времени для выделения траектории частоты основного тона по сигналу, содержащему K кадров, оцениваются величиной $O[K(N+q)(q+T_{max}-T_{min})]$.

4. Экспериментальные результаты

Описанный алгоритм реализован программно и опробован при оценивании траектории частоты основного тона реальных и синтезированных речевых сигналов. Обработке подвергались сигналы, введенные в компьютер через динамический микрофон и восьмиразрядный аналого-цифровой преобразователь при частоте дискретизации

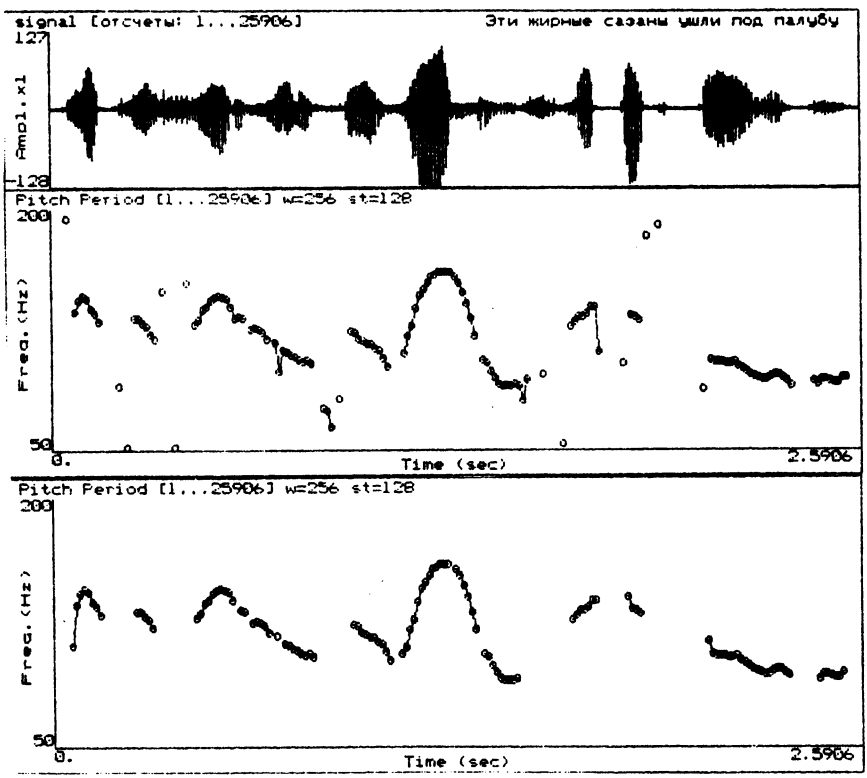

Рис.1

10 кГц. Ширина N анализируемого кадра равнялась 256 отсчетам, а расстояние между двумя последовательными кадрами — 128 отсчетам. Значения параметров, управляющих работой алгоритма, были следующими: $T_{min} = 80$, $T_{max} = 160$, $q = T_{min}$, $\theta_1 = 3$, $\theta_2 = 0.8$, $\theta_3 = 0.5$.

В качестве примера на рис.1 представлены три графика. Верхний график - осциллограмма анализируемого речевого сигнала, соответствующего фразе "Эти жирные сазаны ушли под палубу". Ниже расположен график оценок частоты основного тона, полученный посредством обработки сигнала процедурой анализа квазипериодичности (ядро алгоритма) без использования дополнительных средств регуляризации траектории. Самый нижний график иллюстрирует результат алгоритмического анализа того же сигнала с применением повторной прогонки, обработки неполных импульсов, дополнительного критерия классификации голосовой/неголосовой и коррекции оценок. Разрывы в траекториях, приведенных на графиках, соответствуют тем участкам сигнала, которые были классифицированы как неголосовые.

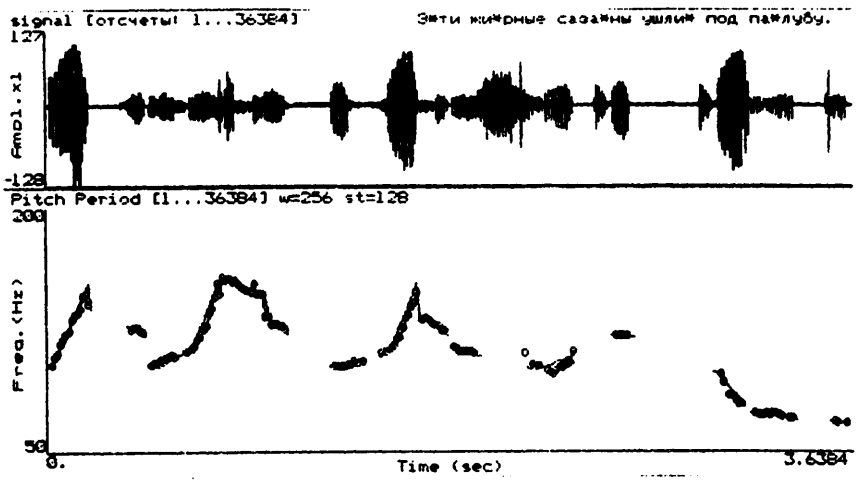

Рис.2

На рис.2 представлены результаты компьютерного анализа синтезированного речевого сигнала. Для синтезированного сигнала контур частоты основного тона задается при генерации устной речи. Поэтому возможно сопоставление истинной и оцененной траекторий. Верхний график является осциллограммой сигнала, синтезированного по тексту "Эти жирные сазаны ушли под палубу". На нижнем графике совмещены сгенерированная синтезатором траектория частоты основного тона и ее оценка, найденная с помощью описанного алгоритма. График иллюстрирует хорошее совпадение заданной и найденной траектории.

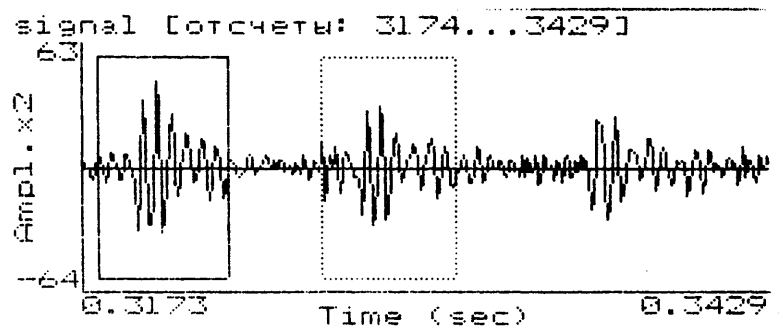

Рис.3

На рис.3–5 приведены примеры обработки отдельных анализируемых кадров. На этих рисунках представлены осциллограммы участка сигнала, попавшего в кадр. Прямоугольные рамки на каждом из рисунков соответствуют найденным с помощью алгоритма импульсам. Рис. 3 иллюстрирует результаты первой прогонки для кадра, вычлененного из вокализованного сигнала, соответствующего фонеме /ы/ (перед фонемой /р/) в слове "жирные". Левый импульс на этом рисунке — это найденный "эталонный" импульс. Правый — это импульс, обнаруженный в результате анализа квазипериодичности. Еще один импульс, который обнаруживается визуально,

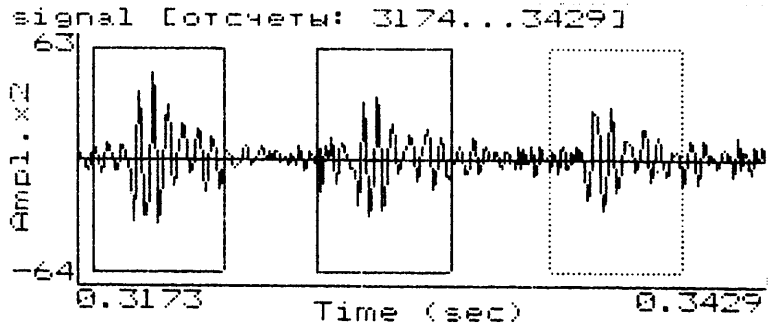

Рис.4

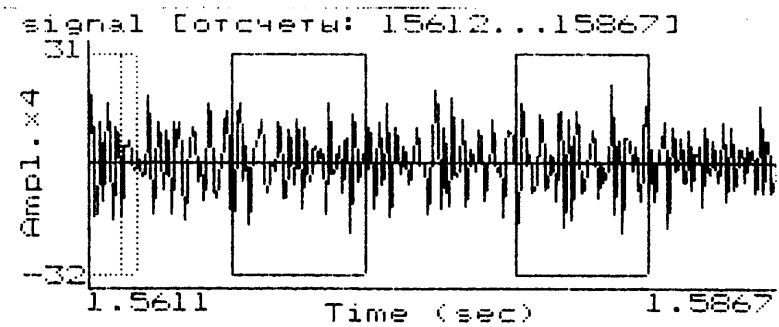

Рис.5

при первой прогонке алгоритмом ошибочно пропущен. На рис.4 представлены результаты повторной обработки того же кадра. В качестве нового "эталонного" импульса алгоритмом выбран найденный в результате первой прогонки правый импульс (см. рис.3). После повторной прогонки обнаруживаются все три имульса (см. рис.4) и, таким образом, ошибка оказывается исправленной.

Рис. 5 иллюстрирует обработку невокализованного участка сигнала, соответствующего фонеме /ш/ в слове "ушли". Результаты обеих прогонок на этом рисунке совмещены. При первой прогонке в качестве "эталонного" был выбран правый импульс, а обнаружены — левый полный импульс и неполный импульс в начале кадра. При второй прогонке в качестве "эталонного" служил обнаруженный при первой прогонке полный импульс. В результате повторной прогонки обнаружен один полный и один неполный импульсы. При этом расположение в кадре полного импульса, обнаруженного в результате второй прогонки, совпало с "эталонным" импульсом для первой прогонки. Положения неполных импульсов для двух прогонок отличаются (на рисунке это отличие помечено пунктирной линией). Анализируемый кадр был классифицирован как неголосовой, так как величина разброса δ оказалась равной 20, т.е. выше заданного порога.

Заключение

В работе предложен новый алгоритм оценивания траектории частоты основного тона по речевому сигналу, представленному в виде числовой последовательности. Оценки частоты основного тона, составляющие траекторию, находятся как средние значения интервала между импульсами голосовых связок в скользящем по речевому сигналу кадре фиксированной длины. Расстояние между импульсами голосового источника в анализируемом кадре находится за два прохода, каждый из которых состоит в решении дискретной экстремальной задачи.

Результаты экспериментов показали, что двухпроходовый принцип обработки сигнала вместе с более точной обработкой участков сигнала, соответствующих границам кадра и содержащих неполные импульсы, позволяет повысить точность оценивания.

Алгоритм имеет невысокую трудоемкость, что дает возможность получения искомых контуров частоты основного тона на современных персональных компьюте-

рах без ощутимых для пользователя задержек, хотя и не в реальное время.

Простота программирования и настройки алгоритма позволяют надеяться на его применение в разнообразных системах автоматической обработки речевых сигналов.

Литература

1. МАРКЕЛ Дж.Д., ГРЭЙ А.Х. Линейное предсказание речи // Пер. с англ. под ред. Ю.Н.Прохорова, В.С.Звездина. — М.: Радио и связь. — 1980. — 308 с.

2. РАБИНЕР Л.Р., ШАФЕР Р.В. Цифровая обработка речевых сигналов // Пер. с англ. под ред. М.В.Назарова, Ю.Н.Прохорова. — М.: Радио и связь. — 1981. — 495 с.

3. RABINER L.R., CHENG M.J., ROSENBERG A.E., McGONEGAL A.M. A Comparative Performance Study of Several Pitch Detection Algorithms // IEEE Trans. Acous. Speech, Signal Processing. – 1976. – Vol. ASSP-24, №5, Oct. — P. 399–418.

4. WISE J.D., CAPRIO J.R., PARKS T.W. Maximum Likelihood Pitch Estimation // IEEE Trans. Acous. Speech, Signal Processing. — 1976. — Vol. ASSP-24, №5, Oct. — P. 418–424.

5. КЕЛЬМАНОВ А.В. Алгоритм выделения основного тона по разностной функции ряда остаточных ошибок модели авторегрессии // Методы обнаружения закономерностей с помощью ЭВМ. — Новосибирск, 1981. — Вып. 91: Вычислительные системы. — С. 113–124.

6. DOLOGLOU I., CARAYANNIS G. Pith Detection Based on Zero–Phase Filtering //Speech Communication. — Vol. 8, №4, Dec. 1989. — P. 309–319.

7. ЛЮДОВИК Е.К. Определение периода основного тона с помощью динамического программирования // Тез. докл. 9-й Всесоюз. школы–семинара "Автоматическое распознавание слуховых образов" (АРСО– 9). — Минск, 1976. — С.48.

8. ВИНЦЮК Т.К. Анализ, распознавание и интерпретация речевых сигналов // Наукова думка, Киев. — 1987. — 262 с.

9. КЕЛЬМАНОВ А.В., КУТНЕНКО О.А. Алгоритм распознавания квазипериодической последовательности импульсов и обнаружения моментов их начала в гауссовском шуме // Настоящий сборник. — С. 137–182.

10. КЕЛЬМАНОВ А.В. Алгоритм классификации тон/шум, основанный на критерии адекватности модели авторегрессии // Методы обработки информации. — Новосибирск, 1978. — Вып. 74: Вычислительные системы. — С. 129–148.

Поступила в редакцию
12 ноября 1996 года