

УДК 534.78

**ВЫДЕЛЕНИЕ МЕЛОДИИ РЕЧИ СПЕКТРАЛЬНО-ВРЕМЕННЫМ
МЕТОДОМ С КОРРЕКЦИЕЙ***В. Н. Соболев*

Описан алгоритм выделения мелодии речи, основанный на фазировании речевой волны. Алгоритм предусматривает различные способы автоматической коррекции ошибок. Приведены экспериментальные данные, показывающие зависимость процента ошибок от значения параметра алгоритма и способа коррекции.

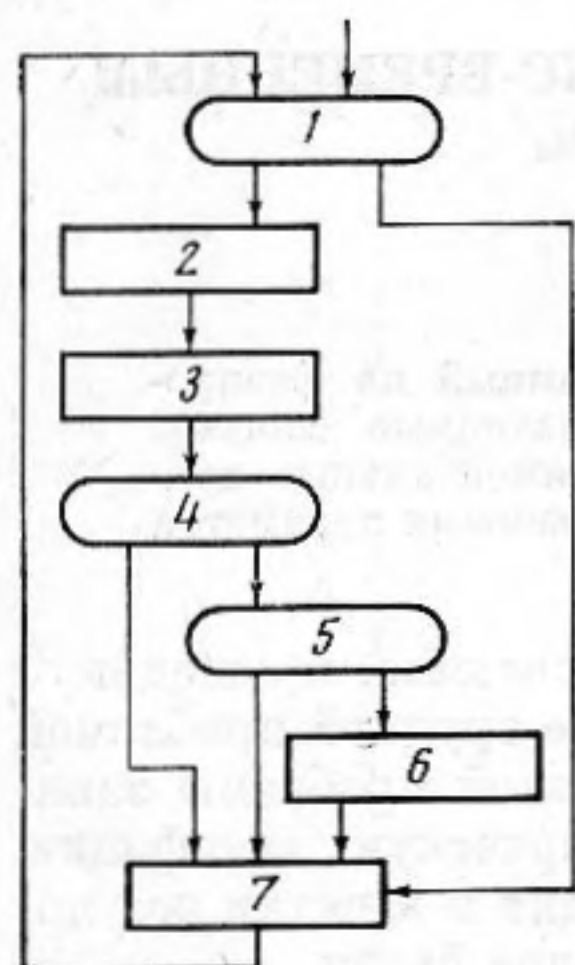
Надежное выделение основного тона из речевого сигнала, прошедшего коммутируемый телефонный канал, является наиболее трудной проблемой вокодерной техники. От успешного решения именно этой проблемы зависит в настоящее время широкое внедрение в коммерческую телефонию параметрических систем компрессии речи, позволяющих в десятки раз повысить пропускную способность существующих каналов связи.

Опыт многочисленных исследований показывает, что наибольшего успеха в повышении надежности выделения основного тона следует ожидать на пути создания многоканальных выделителей с использованием принципа адаптации [1]. Адаптация заключается в изменении параметров устройства в соответствии с накапливаемым опытом, что в ряде случаев приводит к значительному повышению надежности выделения (см., например, работы [2, 3]). Однако непрерывная адаптация имеет существенный недостаток, заключающийся в потенциальной возможности перехода алгоритма в режим слежения за гармоникой или субгармоникой основного тона.

Нами предпринята попытка замены непрерывной адаптации дискретной во времени. Результаты выделения основного тона непрерывно анализируются, регистрируются моменты резких отклонений длительности найденного периода основного тона от предшествующих значений. Такие резкие отклонения, как правило, соответствуют сбоям в работе выделителя; лишь для этих моментов производится коррекция результатов измерения путем повторного анализа формы этих участков сфазированной речевой волны. В процессе повторного анализа учитывается априорная информация о значении измеряемого периода основного тона. Эта информация в виде сведений об ожидаемом значении периода черпается из значений периода основного тона для ряда предшествующих моментов времени. Таким образом, механизм адаптации включается лишь в случае получения подозрительных значений, что снижает вероятность перехода на гармоники и субгармоники основной частоты. Исследование описанного способа адаптации проводилось путем цифрового моделирования [4]; поэтому дальнейшее изложение мы будем вести с учетом специфики алгоритмизации для ЭВМ.

На фиг. 1 представлена общая схема алгоритма, пригодного для обработки последовательности участков речевого сигнала (доз речевой информации). Блок 1, который можно назвать дискриминатором тон — шум, определяет, относится ли данный участок к звонкому (вокализованному)

звуку или к глухому звуку и речевой паузе. В случае отнесения участка к вокализованному звуку управление передается блоку 2, который формирует из речевой волны функцию времени, более удобную для последующего определения периода основного тона, чем сама речевая волна. В качестве такой функции можно использовать автокорреляционную [5] «сдвиговую» [2], «кепстральную» [6] функции или сфазированную речевую волну. Блок 3 производит поиск экстремумов сформированной функции; абсциссы экстремумов несут информацию о периоде основного тона. В простейшем случае в качестве длительности периодов основного тона, усредненной на данной дозе, принимается абсцисса главного экстремума, расположенного в диапазоне возможных значений периода первой гармоники речевого сигнала. Далее по длительностям периодов основного тона в предшествующих дозах блок 4 предпринимает попытку предсказания периода на данной дозе. Если такое прогнозирование оказывается невозможным, управление передается блоку 7. В противном случае управление получает блок 5, который устанавливает необходимость выполнения коррекции периода, найденного блоком 3. Если найденное для данной дозы значение периода сильно отличается от предшествующих значений (коррекция необходима), управление передается блоку 6. Он производит повторный (локальный) анализ сформированной блоком 2 функции и выбирает тот экстремум, абсцисса которого имеет наименьшее отличие от предсказанного блоком 4 значения длительности. Эта абсцисса и принимается за истинное значение периода основного тона. Если сравниваемые блоком 5 значения близки, то коррекция не производится и управление передается блоку 7, осуществляющему смену доз речевой информации.



Фиг. 1. Блок-схема алгоритма

Если сравниваемые блоком 5 значения близки, то коррекция не производится и управление передается блоку 7, осуществляющему смену доз речевой информации.

Перейдем к более детальному рассмотрению отдельных частей описанного алгоритма. Известны различные алгоритмы работы дискриминатора тон — шум [1]. Мы применили с целью упрощения блока 1 сравнение усредненной (на j -й дозе) энергии речевой волны $\sum_j |f(n\Delta t)|$ с задан-

ным энергетическим порогом \mathcal{D} , выбираемым экспериментально. Если усредненная энергия превышала порог, регистрировалось наличие вокализованного звука.

Формирование функции, удобной для выделения основного тона, в блоке 2 производилось путем фазирования дискретизированной речевой волны $f(n\Delta t)$ по формуле

$$\Phi_j(l) = \sum_{k=K}^K c_{jk} \cdot \cos\left(2\pi k \frac{l}{N}\right),$$

где

$$c_{jk} = \sqrt{a_{jk}^2 + b_{jk}^2},$$

$$a_{jk} = \sum_{n=(j-1)v}^{(j-1)v+N} f(n\Delta t) \cdot \cos\left[2\pi k \frac{n - (j-1)v}{N}\right],$$

$$b_{jk} = \sum_{n=(j-1)v}^{(j-1)v+N} f(n\Delta t) \cdot \sin\left[2\pi k \frac{n - (j-1)v}{N}\right],$$

$$K = \text{entier}(F_{\max} N \Delta t),$$

$$x = \begin{cases} 1 & \text{для сигнала на выходе динамического} \\ & \text{микрофона,} \\ \text{entier}(F_{\min} N \Delta t) & \text{для сигнала, прошедшего через модель} \\ & \text{телефонного тракта*}, \end{cases}$$

$$l=0, 1, 2, \dots, N/2, \quad j=1, 2, 3, \dots$$

Алгоритм испытывался при следующих значениях параметров: $F_{\min}=300$ гц, $F_{\max}=1, 2$ и 3 кгц, $\Delta t=111$ мкс, $N=256$, $v=200$. Три последних значения определяют интервал анализа $T=N\Delta t=28,4$ мс и шаг выборки доз $\theta=v\Delta t=22,2$ мс.

На фиг. 3 представлены типичные примеры спектров $s(f)$ и фазированной речи $\Phi(t)$ для мужского (фиг. 3, а) и женского (фиг. 3, б) голосов. Из фигур видно, что в графиках функции $\Phi(t)$ присутствуют явно выраженные максимумы, соответствующие периоду основного тона T_0 . Задача выделения основного тона сводится к нахождению абсцисс этих максимумов, нередко маскируемых побочными максимумами. Эффект маскировки может приводить к сбоям в работе алгоритма выделения.

Фазирование речевой волны способствует увеличению пикфактора и облегчает выделение максимумов. Формирование функции $\Phi(t)$ по приведенной выше формуле выгодно отличается от формирования автокорреляционной функции отсутствием квадрирования спектра и выравниванием интенсивности речевого сигнала на нестационарных участках. Фазирование речи не сопровождается логарифмированием спектра, что обуславливает преимущества функции $\Phi(t)$ по сравнению с «кепстральной» функцией [6] при последующем анализе формы ее графика.

Измерение периода основного тона для j -й дозы в блоке 3 заключается в определении $\arg \text{Max} [\Phi_j(l)]$, расположенного в диапазоне $20 \leq l < N/2$. Этот диапазон значений дискретного аргумента соответствует при $\Delta t=111$ мкс диапазону вероятных длительностей периода основного тона 2,2–14,2 мс.

При выборе величины периода разложения T следует учитывать, что функция $\Phi(t)$ периодична и на отрезках $[\pm(p-1)T, \pm pT]$ зеркально-симметрична относительно их середин, т. е. относительно точек $\pm p \frac{T}{2}$

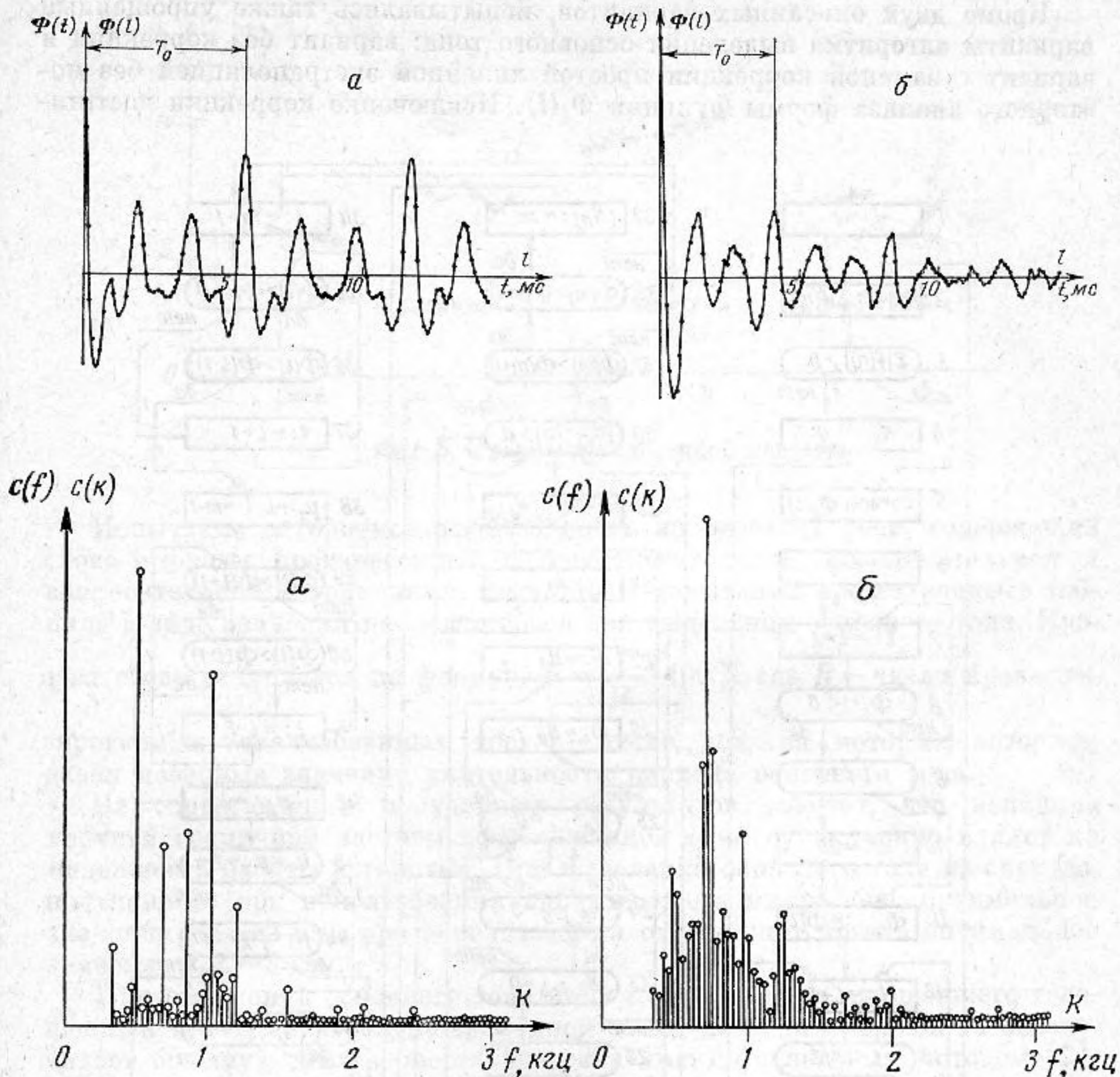
(здесь $p=1, 2, 3, \dots$). Поэтому во избежание нежелательной интерференции всплеска функции $\Phi(t)$, порожденного первой гармоникой речевого сигнала, и зеркального отражения этого всплеска, необходимо соблюдать условие $T \geq 2T_{\max}$. Из соображений наименьшего усреднения измеряемых периодов целесообразно выбирать период разложения $T=2T_{\max}$ (здесь T_{\max} — максимально возможный период основного тона речи).

По ряду причин** ординаты максимумов, соответствующих периодам основного тона, могут становиться меньше ординат других (побочных) максимумов. Возникающие вследствие этого сбои должны быть по возможности устранены в процессе коррекции, осуществляемой блоком 6, опирающимся в своей работе на прогноз, составленный блоком 4. Успех корректирования во многом зависит от точности прогноза. Мы исследовали различные алгоритмы предсказания; ниже рассматриваются лишь два наиболее простых и эффективных алгоритма.

Оба алгоритма включаются с некоторой задержкой после длительной паузы в мелодии речи и работают непрерывно до следующей длительной

* Имитация телефонного тракта заключалась в пропуске речевого сигнала с угольного микрофона через фильтр верхних частот с частотой среза 350 гц и обеспечении отношения сигнала к шуму квантования 36 дб.

** Нестрогая периодичность речевого сигнала: фазирование речевой волны не по гармоникам основного тона, а по гармоникам разложения s_k и т. д.



Фиг. 3. Сфазированная речь и спектры

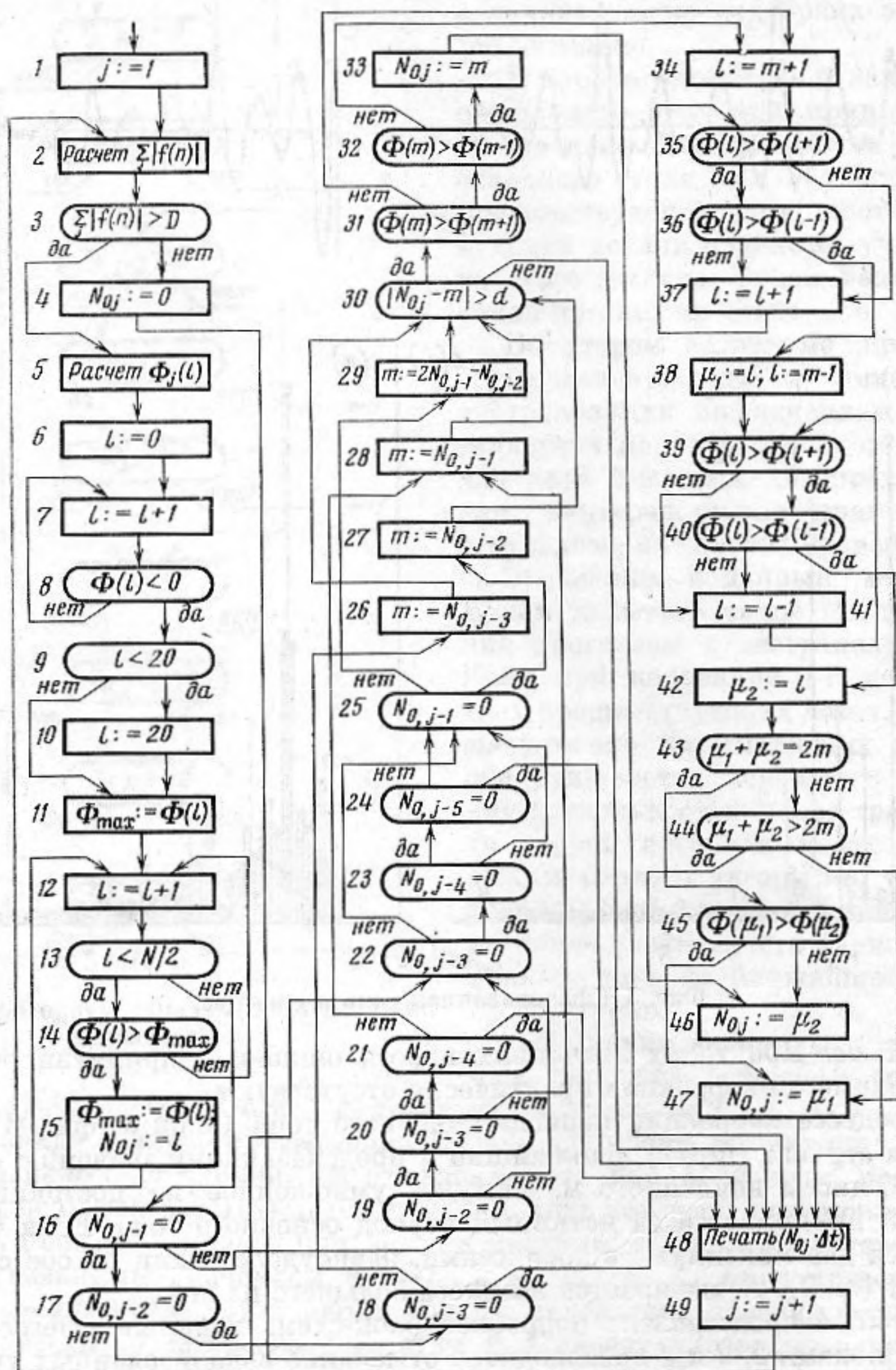
сов — 1,1 мс. При таких значениях порога ошибки в принятии решения о необходимости коррекции практически отсутствуют.

В процессе коррекции периода основного тона (блок 6, фиг. 1) определяется $\arg \max [\Phi_j(l)]$, ближайший к предсказанному значению m . Значение абсциссы найденного максимума, умноженное на постоянную отсчета Δt , принимается за истинный период основного тона. Если обнаруживаются два максимума с абсциссами, равноудаленными по обе стороны от точки $(m, 0)$, то выбирается абсцисса большего из них.

На фиг. 4 представлена подробная блок-схема испытывавшегося алгоритма. Блоками 2, 3 и 4 производится отыскание вокализованных участков речевого колебания и присвоение значению периода основного тона величины, равной нулю, в случае интонационной паузы. Блок 5 производит расчет функции $\Phi_j(l)$. Блоки 6–15 осуществляют первичный анализ функции $\Phi_j(l)$ и определяют значение периода основного тона. Блоки 16–25 производят анализ предыстории и совместно с блоками 26–29 выбирают значение m , используемое при последующей коррекции периода основного тона, выполняемой блоками 31–47. Необходимость проведения коррекции устанавливается блоком 30. Блок 48 печатает окончательное значение периода основного тона для j -й дозы. Блок 49 осуществляет смену доз речевой информации.

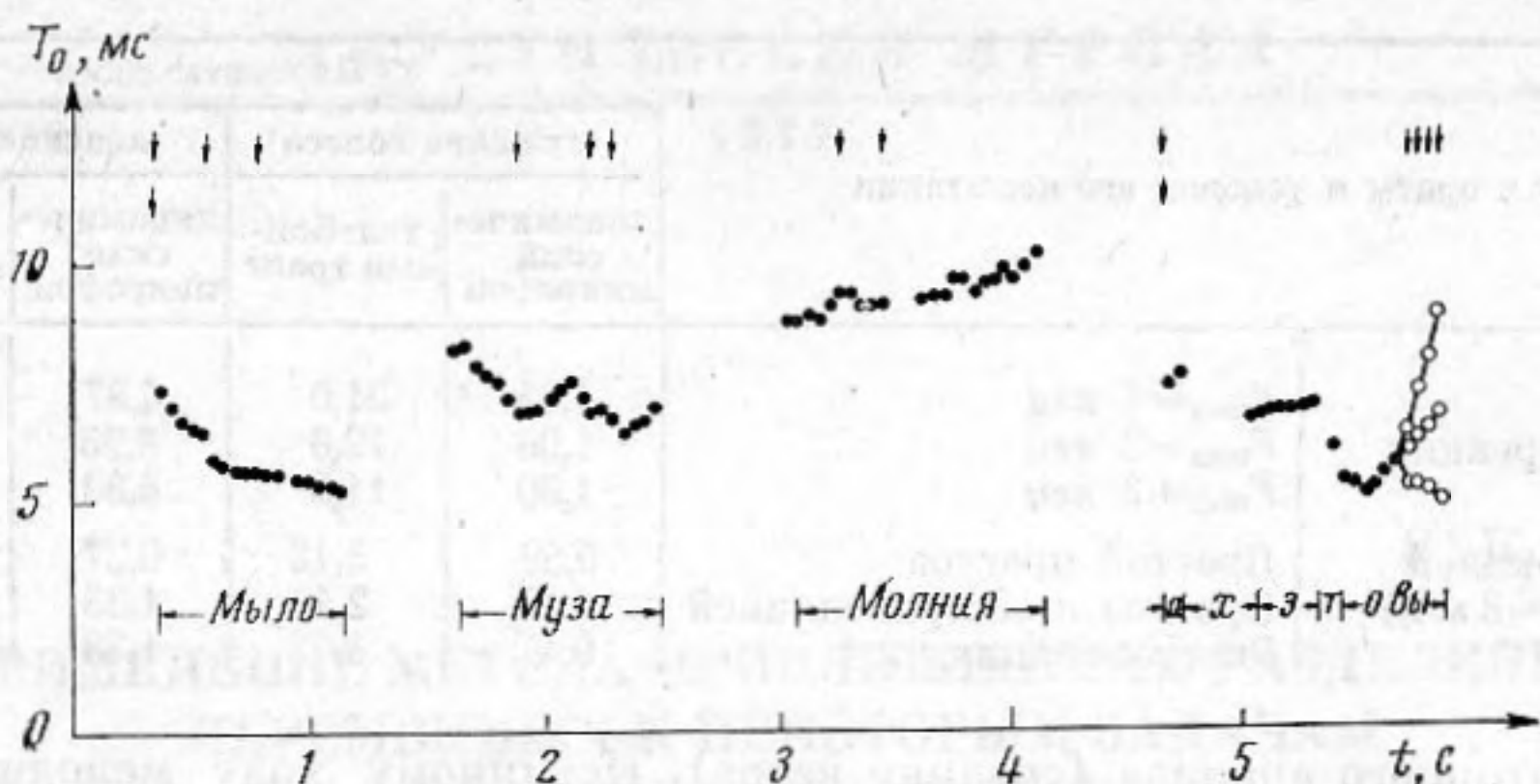
Приведенная блок-схема предусматривает прогноз с экстраполяцией. Для перехода к алгоритму простого прогноза необходимо заменить содержимое блока 29 оператором $m := N_{0, j-1}$.

Кроме двух описанных вариантов, испытывались также упрощенные варианты алгоритма выделения основного тона: вариант без коррекции и вариант с заменой коррекции простой линейной экстраполяцией без повторного анализа формы функции $\Phi_j(l)$. Исключение коррекции достига-



Фиг. 4. Цифровая модель выделителя основного тона

лось передачей управления с правого выхода блока 13 непосредственно на вход блока 48. Замена коррекции экстраполяцией достигалась включением единственного блока, содержащего оператор $N_{0j} := m$, вместо серии блоков 31–47. Тогда при необходимости коррекции и отсутствии длительной паузы в обозримой предыстории окончательное значение длительности периода основного тона в большинстве случаев определялось как результат линейной экстраполяции по длительности периодов на двух предшествующих дозах.



Фиг. 5. Фрагмент выделенной мелодии

Испытание алгоритмов производилось на отрезках речи, содержащих слова и фразы, произнесенные с повествовательной, восклицательной и вопросительной интонациями. Результаты испытаний представлены в таблице в виде значений процента сбоев при выделении основного тона. Процент сбоев исчислялся по формуле $P = \frac{r}{R} \cdot 100\%$, где R — число проанализированных вокализованных доз, r — число доз, на которых алгоритм давал неверные значения длительности периода основного тона.

Из сопоставления полученных результатов следует, что величина верхней граничной частоты сфазированной речи существенно влияет на надежность работы алгоритма. При выделении основного тона из сигнала, полученного при использовании динамического микрофона, оптимальное значение $F_{\max} = 2$ кГц; при использовании телефонного тракта оптимальное значение $F_{\max} = 3$ кГц.

При выделении основного тона из речевого сигнала, прошедшего телефонный тракт, спектрально-временной метод даже без коррекции обеспечивает большую достоверность по сравнению со сдвиговым методом [2], дающим на мужских голосах 12,2%, а на женских 9% сбоев. На женских голосах спектрально-временной метод оказывается эффективнее автокорреляционного метода [5], дающего 6,8% сбоев.

При выделении основного тона спектрально-временным методом на женских голосах более эффективной оказалась коррекция по результатам простого прогноза, а на мужских голосах — коррекция по результатам прогноза с экстраполяцией. Линейная экстраполяция без повторного локального анализа формы сфазированного речевого сигнала в большинстве случаев равноценна коррекции по результатам прогноза с экстраполяцией, однако уступает последней в точности измерения периода основного тона.

При использовании телефонного тракта спектрально-временной метод с коррекцией по результатам простого прогноза обеспечивает на женских голосах в 3,5 раза меньший процент сбоев, чем сдвиговый метод с непрерывной адаптацией [3], который дает 3,8% сбоев. На мужских голосах оба метода равноценны.

На фиг. 5 представлен фрагмент выделенной мелодии. Верхний ряд вертикальных стрелок соответствует моментам сбоев алгоритма без коррекции. Нижний ряд стрелок указывает моменты сбоев алгоритмов с коррекцией. Измеренные значения периода основного тона для доз, на которых произошли сбои, на фигуре не показаны, за исключением конца реплики «ах, это вы!». Здесь отдельно показаны результаты выделения основного тона алгоритмом с коррекцией по результатам простого прогноза (нижняя ветвь), при использовании коррекции по результатам прогноза с экстраполяцией (верхняя ветвь) и при использовании экстраполяции

Алгоритм и условия его испытания		Проценты сбоев			
		мужские голоса		женские голоса	
		динамиче- ский микрофон	телефон- ный тракт	динамиче- ский микрофон	телефон- ный тракт
Без коррекции	$F_{\max}=1$ кгц	1,65	21,0	4,87	10,2
	$F_{\max}=2$ кгц	1,06	12,6	3,93	5,21
	$F_{\max}=3$ кгц	1,30	11,8	4,33	4,56
С коррекцией ($F_{\max}=3$ кгц)	Простой прогноз	0,59	3,12	0,27	1,09
	Прогноз с экстраполяцией	0,47	2,46	1,23	1,30
	Экстраполяция	0,47	3,12	1,23	1,30

без повторного анализа (средняя ветвь). Истинному ходу мелодии соответствует верхняя ветвь.

Следует учитывать различия в характере ошибок при работе описанных алгоритмов. Алгоритм с коррекцией по результатам простого прогноза склонен к ошибкам при большой скорости изменения частоты основного тона. В процессе коррекции по результатам прогноза с экстраполяцией сбои, как правило, возникают при изменении знака производной в интограмме.

Алгоритм с экстраполяцией без повторного анализа формы функции $\Phi_j(l)$ часто дает незначительные ошибки, которые следует расценивать не как грубые сбои, а как погрешности в измерении периода основного тона. При коррекции серии ошибок, идущих подряд (на участках монотонного изменения частоты основного тона), эти погрешности накапливаются, значительно искажая форму графика изменения основного тона. С момента превышения разумно допустимого значения погрешности такие ошибки следует относить к сбоям.

По-видимому, имеются резервы повышения надежности выделения основного тона спектрально-временным методом. Во-первых, можно сузить границы первичного анализа функции $\Phi_j(l)$, осуществив настройку на мужские или женские голоса. Во-вторых, целесообразно повысить разрешающую способность спектрального анализа, увеличив интервал T и введя колоколообразную весовую функцию для компенсации увеличения времени усреднения. Это приблизит условия формирования функции $\Phi_j(l)$ к желаемым условиям фазирования гармоник основного тона. Некоторый эффект может дать, наконец, переход от линейной к более сложной и точной экстраполяции.

ЛИТЕРАТУРА

1. Вокодерная телефония (Методы и проблемы). Под ред. А. А. Пирогова. М., «Связь», 1974, 534.
2. В. Н. Соболев, С. П. Баронин. Исследование сдвигового метода выделения основного тона речи. Электросвязь, 1968, 12, 30–36.
3. В. Н. Соболев. Исследование адаптивного алгоритма выделения мелодии речи. Тр. учебных ин-тов связи, 1972, вып. 59, 3–9.
4. Р. Д. Лейтес, В. Н. Соболев. Цифровое моделирование систем синтетической телефонии. М., «Связь», 1969, 118.
5. В. Н. Соболев. Экспериментальное исследование корреляционного метода выделения основного тона речи. Акуст. ж., 1968, 14, 3, 441–448.
6. A. Noll. Short-Time Spectrum and «Cepstrum» Techniques for Vocal—Pitch Detection. J. Acoust. Soc. America, 1964, 36, 2, 292–302.

Всесоюзный заочный
электротехнический институт связи

Поступила
16 декабря 1974 г.
После окончательного исправления
19 апреля 1976 г.