УДК 004.934;004.912

ОБРАБОТКА СИГНАЛА В ЧАСТОТНОЙ ОБЛАСТИ ПРИ РАСПОЗНАВАНИИ РЕЧИ[1]

А. С. Колоколов

Институт проблем управления им. В. А. Трапезникова, г. Москва

Рассмотрены способы обработки речевого сигнала в частотной области, обеспечивающие описание сигнала, устойчивое к частотным искажениям и аддитивным шумам. В их основу положены преобразования логарифмического спектра, реализованные на базе полосовой фильтрации спектральной огибающей. В этих преобразованиях в общих чертах учтены эффект латерального торможения и ответы фазических нейронов в слуховом анализаторе.

ВВЕДЕНИЕ

Распознавание речи представляет собой многоуровневый процесс декодирования речевого сообщения, начинающийся с распознавания фонем. В результате непрерывный речевой сигнал $s(t)$ представляется цепочкой дискретных элементов — фонем, принадлежащих конечному алфавиту, размер которого не превышает нескольких десятков элементов. Этим достигается резкое сокращение описания сигнала в сравнении с его описанием на уровне акустической волны, представляемой изменением звукового давления в функции времени. Затем по полученной цепочке фонем последовательно декодируются слова, предложения и, наконец, смысл высказывания.

Процессу распознавания фонем предшествует предварительная обработка речевого сигнала в целях сокращения описания речевого сигнала и последующего его представления набором информативных признаков. Однако до сих пор система информативных признаков, обеспечивающих распознавание речевого сигнала на фонетическом уровне, исследователями не найдена. Тем не менее, совокупность имеющихся данных, полученных при исследовании речеобразования и восприятия речи, а также слухового анализатора, позволяет сделать вывод, что передача информации в речевом сигнале осуществляется изменениями его кратковременного амплитудного спектра $S(f, t)$, отражающими способ и место образования звука в процессе артикуляции [1—4]. Это обусловливает большой интерес к проблеме спектрального анализа речи у исследователей речевого сигнала и разработчиков систем распознавания речи.

В настоящей работе рассматриваются вопросы получения спектрального описания речевого сигнала и применения специальных преобразований спектра, сокращающих избыточность и повышающих устойчивость спектрального описания.

1. СПЕКТРАЛЬНЫЙ АНАЛИЗ РЕЧЕВОГО СИГНАЛА

Поскольку речевой сигнал представляет собой изменяющийся во времени процесс, то его спектральное описание основывается на концепции кратковременного анализа [5]. Для этого речевой сигнал $s(t)$ разбивается на равные перекрывающиеся отрезки, называемые фреймами или кадрами, в пределах которых свойства сигнала мало изменяются и его можно считать квазистационарным. Обычно длительность фрейма выбирается равной 10—30 мс, а его формирование осуществляется умножением сигнала $s(t)$ на окно $w(t - n\Delta T)$, где $n = 0, 1, 2, \ldots$ — индекс, определяющий номер фрейма, ΔT — интервал между соседними фреймами, составляющий 5—10 мс, обеспечивающий необходимую детальность спектрального описания во времени. Далее для каждого фрейма выполняется спектральный анализ, в результате чего находится последовательность амплитудных спектров $S(f, n)$, где f — частота, n — номер фрейма. Последовательность спектров $S(f, n)$, представляющих речевой сигнал, обычно называется динамической спектрограммой или видимой речью. Найденный спектр $S(f, n)$ отличается от текущего спектра $S(f, t)$ тем, что представляет последний в дискретные моменты времени $n\Delta T$.

Для получения спектров $S(f, n)$ обычно используются разные модификации дискретного преобразования Фурье, применяемые для спектрального анализа с линейной частотной шкалой. В этом случае спектр находится на ряде дискретных, равноотстоящих частот. В последнее время наблюдается значительный интерес к спектральным анализаторам, выполненным на основе гребенки полосовых фильтров, в общих чертах учитывающих особенности частотного анализа звука в слуховой системе. Их особенность состоит в использовании нелинейной частотной шкалы Барков или мелов и срав-

[1] Работа доложена на Научных чтениях памяти профессора А. М. Петровского, Москва, Ин-т проблем управления, 17 марта 2005 г.

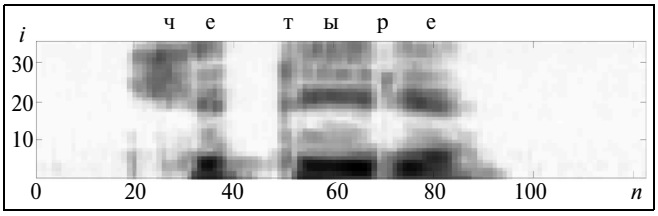

Рис. 1. Спектрограмма слова «четыре»

нительно широкополосных, низкодобротных фильтров с полосами пропускания, выбранными в соответствии с зависимостью критической полосы слуха от частоты [6, 7]. Этим обеспечивается сравнительно низкое разрешение анализатора по частоте, однако достаточное для выделения характерных резонансов речевого тракта, и хорошее разрешение по времени вследствие низкой добротности анализирующих фильтров.

В практике спектрального анализа речи часто используются логарифмическая шкала интенсивности. Ее применение обосновано тем, что кодирование интенсивности в рецепторах подчиняется закону Вебера—Фехнера, согласно которому минимально заметный прирост ΔI внешнего воздействия на рецептор пропорционален воздействию I, т. е. $\Delta I \sim \Delta\beta I$, где $\Delta\beta$ — приращение ответа рецептора. Отсюда $\beta \sim \lg I$ и, следовательно, ответ рецептора пропорционален логарифму внешнего воздействия.

В настоящей работе для получения спектрограмм речевых сигналов использовалась гребенка из 35-ти цифровых полосовых фильтров, центральные частоты которых f_{0i}, $i = 1, 2, ..., 35$, располагались равномерно по шкале Барков с шагом 0,57 Барка, начиная с 1,95 Барка, что соответствовало диапазону частот от 200 до 8660 Гц. Частотные характеристики слуховых фильтров аппроксимировались полосовыми фильтрами Баттерворта четвертого порядка с наклонами частотной характеристики 12 дБ/окт и шириной полосы пропускания 1,5 Барка. Последовательные спектры находились с интервалом $\Delta T = 8$ мс и таким образом формировалась последовательность спектров $S(i, n)$, где i — номер фильтра.

На рис. 1 приведена спектрограмма $S(i, n)$ слова «четыре», на которой интенсивность спектра выражена в логарифмической шкале и передается уровнем черного. Из спектрограммы видно, что речевой сигнал состоит из ряда квазистационарных сегментов, представляющих отдельные фонемы. При изолированном прослушивании таких сегментов они обычно воспринимаются с определенным фонетическим качеством. Это свидетельствует о том, что в выделяемых сегментах присутствует необходимая фонетическая информация.

Поскольку в кратковременном спектре речевого сигнала содержится информация для его распознавания, может показаться привлекательной распространенная гипотеза о том, что восприятие речи основано на сравнении спектра речи с набором эталонных спектров [8]. Однако в результате частотных искажений вид спектра речевого сигнала может существенно изменяться, что имеет место при использовании микрофонов с различными частотными характеристиками, дифференцировании сигнала, присутствии реверберации и т. д. Тем не менее, во всех перечисленных случаях, несмотря на значительные частотные искажения, разборчивость сигнала не изменяется, а его восприятие сопровождается лишь определенными изменениями тембра звучания, что свидетельствует об устойчивости восприятия речи к частотным искажениям. Таким образом, более интересной представляется гипотеза о том, что информативные признаки, определяющие фонетическое качество звука, связаны с неоднородностями его спектра по частоте и времени [4], представляющими локальные особенности спектра.

Далее рассматриваются способы обработки кратковременного спектра, основанные на локальных свойствах спектра и обеспечивающие выделение информативных признаков речевого сигнала, устойчивых к частотным искажениям.

2. ВЫДЕЛЕНИЕ ЛОКАЛЬНЫХ НЕОДНОРОДНОСТЕЙ СПЕКТРА ПО ЧАСТОТЕ

Выделение спектральных неоднородностей по частоте в слуховом анализаторе связывается с эффектом латерального торможения. Данный эффект обычно объясняется локальной обработкой спектра в слуховом анализаторе, которую можно представить сверткой кратковременного амплитудного спектра $S(f, t)$ с весовой функцией $\varphi(t)$, описывающей распределение возбуждающих и тормозных связей, и последующим нелинейным преобразованием. В результате формируется преобразованный спектр

$$S_1(f, t) = Q(S(f, t) \otimes \varphi(f)), \qquad (1)$$

где $\otimes$ — операция свертки, $Q(x) = \begin{cases} x, x \geq 0 \\ 0, x < 0. \end{cases}$

Уравнение (1) описывает однородную нейронную сеть, обычно используемую для моделирования эффекта латерального торможения. Как известно, такая сеть при соответствующем подборе функции $\varphi(t)$ подчеркивает максимумы и резкие перепады в спектре $S(f, t)$ [9]. Для этих целей используется функция $\varphi(f) = (-\delta(f - \Delta f) + 2\delta(f) - \delta(f + \Delta f)) \otimes w_1(f)$, где $\Delta(f)$ — дельта функция, $w_1(f)$ — сглаживающее окно. В этом случае свертка $S(f, t) \oplus \varphi(f)$ является отрицательной сглаженной второй разностью спектра $-\tilde{\Delta}_f^2 S(f, t) = (-S(f - \Delta f, t) + 2S(f, t) - S(f - \Delta f)) \otimes w_1(f)$, приближенно представляющей отрицательную вторую производную $-\partial^2 S(f, t)/\partial f^2$. Типичный вид функции $\varphi(f)$ приведен на рис. 2, *а*. Она имеет центральный положительный лепесток, описывающий распределение возбуждающих связей, и два боковых отрицательных лепестка, характеризующих распределение тормозных связей. На рис. 2, *б* представлен модуль преобразования Фурье функции $\varphi(f)$, показывающий, что она является импульсной характеристикой полосового фильтра. Поэтому свертку в уравнении (1) можно трактовать как процесс полосовой фильтрации спектра $S(f, t)$.

Результат такой фильтрации оказывается более интересным [10], если вместо амплитудного спектра $S(f, t)$ использовать логарифмический спектр $F(f, t) = \lg S(f, t)$. В этом случае уравнение (1) принимает вид

$$F_1(f, t) = Q(F(f, t) \otimes \varphi(f)). \qquad (2)$$

Заметим, что согласно линейной модели речеобразования речевой сигнал в частотной области может быть представлен в виде произведения $S(f, t) = H(f, t)E(f, t)W(f)$, где $H(f, t)$ — частотная характерис-

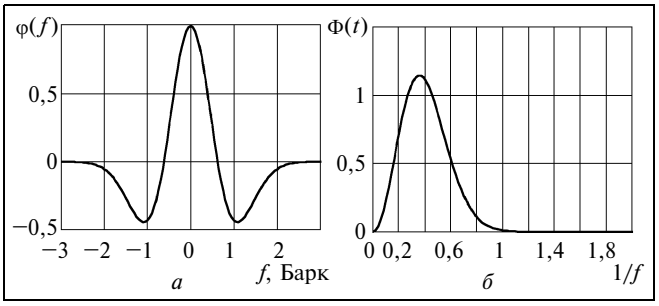

Рис. 2. Типичный вид весовой функции $\varphi(f)$ — (*а*), модуль преобразования Фурье функции $\varphi(f)$ — (*б*)

тика речевого тракта, $E(f, t)$ — спектр шумового или голосового источника, $W(f)$ — характеристика фильтра, описывающего частотные искажения речевого сигнала. После логарифмирования произведение переходит в сумму $F(f, t) = \lg S(f, t) = \lg H(f, t) + \lg E(f, t) + \lg W(f)$. При этом составляющие $F(f, t)$ с разной скоростью изменяются с частотой f и могут быть разделены с помощью линейной фильтрации. Составляющая $W(f)$, связанная с частотными искажениями речи в акустической среде или канале связи, обычно сравнительно медленно изменяется с частотой. В случае шумового источника спектр $E(f, t)$ медленно убывает с частотой со скоростью −6...−12 дБ/окт. Для голосового источника спектр имеет более сложный вид $E(f, t) = I(f, t)G(f, t)$, где $I(f, t)$ — спектр почти периодической последовательности Δ-функций, $G(f, t)$ — спектр импульса голосового источника. Спектр $I(f, t)$ близок к последовательности гармоник с равной амплитудой и в силу этого быстро изменяется с частотой. Спектр $G(f, t)$, как и в случае шумового источника, медленно убывает с частотой со скоростью −6...−12 дБ/окт. Скорость изменения составляющей $H(f, t)$, определяемая резонансами речевого тракта, попадает в область средних скоростей изменения с частотой относительно всех частотных составляющих, рассмотренных ранее. Поэтому, производя полосовую фильтрацию логарифмического спектра $F(f, t)$, можно в обработанном спектре $F_1(f, t)$ значительно ослабить составляющие, связанные с частотными искажениями и источником, обусловливающие вариации спектра речевого сигнала, и сохранить пики $H(f, t)$, связанные с резонансами речевого тракта. Тем самым оказывается возможным сделать более стабильным сравнение речевого сигнала со спектральными эталонами при распознавании.

Процесс полосовой фильтрации $F(f, t)$ завершается выполнением нелинейного преобразования $Q(x)$. С его помощью в обработанном спектре $F_1(f, t)$ сохраняются фрагменты $F(f, t)$, связанные с максимумами (формантами) $H(f, t)$, где отношение сигнал/шум велико. Отрицательные значения $F(f, t) \otimes \varphi(f)$, соответствующие минимумам или нулям спектра $F(f, t)$, в значительной степени зависят от уровня аддитивного шума. Поэтому их исключение с помощью нелинейного преобразования $Q(x)$ позволяет обеспечить дополнительную стабилизацию $F_1(f, t)$ при наличии фоновых широкополосных шумов со спектральной плотностью, сравнительно медленно изменяющейся с частотой.

Таким образом, предложенная обработка логарифмического спектра, сочетающая линейную фильтрацию логарифмического спектра с последующей нелинейной обработкой, позволяет ослабить вариации спектра, вызванные частотными искажениями сигнала, изменениями формы импульса голосового источника и аддитивными шумами.

Заметим, что рассмотренное ранее преобразование спектра (2) является модификацией обработки спектра, основанной на полосовой лифтрации кепстра [11]. Кепстр $C(\psi, t)$ [12] представляет собой косинус-преобразование Фурье логарифма амплитудного спектра $C(\psi, t) = \int_0^\infty F(f, t)\cos(2\pi f\psi)df$, а его лифтрация состоит в умножении кепстра на окно $\theta(\psi)$. После лифтрации кепстра производится обратный переход в частотную область с помощью косинус-преобразования Фурье. Таким образом находится обработанный спектр

$$F(f, t) = \int_0^\infty C(\psi, t)\theta(\psi)(2\pi f\psi)d\psi,$$

являющийся результатом лифтрации кепстра. Как известно, умножение Фурье-образа на окно равнозначно линейной фильтрации оригинала этого образа. Умножение $C(\psi, t)$ на окно, сохраняющее область малых значений ψ, принадлежащих интервалу $0 \leqslant \psi \leqslant \psi_1$, приводит к сглаживанию спектра, тогда как применение окна, выделяющего область средних значений ψ, соответствующих интервалу $\psi_1 \leqslant \psi \leqslant \psi_2$, эквивалентно полосовой фильтрации спектра, подчеркивающей его неоднородности.

Хотя свертка логарифмического спектра $F(f, t)$ с весовой функцией $\varphi(f)$ и умножение кепстра $C(\psi, t)$ на окно $\theta(\psi)$ эквивалентны фильтрации спектра $F(f, t)$, использование для фильтрации свертки имеет ряд преимуществ:

— сокращаются требуемые вычислительные ресурсы, так как вычисление прямого и обратного преобразований Фурье заменяется вычислением короткой свертки;

— в отличие от лифтрации кепстра, фильтрация в частотной области с помощью свертки не ограничивается одними симметричными весовыми функциями $\varphi(f)$.

Поэтому с помощью фильтрации возможно подчеркивание в спектре не только его максимумов, но и других его локальных неоднородностей, например, резких спектральных наклонов, замечаемых при прослушивании звуков. Наконец, использование нелинейного преобразования после фильтрации спектра позволяет несколько повысить устойчивость к шумам.

Эффективность преобразования (2) была проверена на реальных фрагментах речевого сигнала, которые искажались с помощью дифференцирования. Спектральный анализ сигналов проводился с помощью рассмотренной ранее модели слухового частотного анализатора, реализованной с помощью гребенки из 35-ти цифровых полосовых фильтров. С ее помощью для отдельных фреймов речевых сигналов находились логарифмические спектры $F(i)$, где i — номер фильтра. Для фильтрации спектра использовалась симметричная весовая функция $\varphi(i) = -0{,}25\delta_k(i - 2) + 0{,}5\delta_k(i) + 0{,}25\delta_k(i + 2)$, где $\Delta_k(i)$ — функция Кронекера, $i = ...-2, -1, 0, 1, 2, ...$, и вычисление свертки сводилось к суммированию взвешенных спектральных отсчетов.

На рис. 3, *а* приведены спектры $F(i)$ и $F_1(i)$ для гласного «ы» в слове «четыре». На рис. 3, *б* даны спектры для продифференцированного сигнала гласного. Легко видеть, что дифференцирование приводит к существенно-

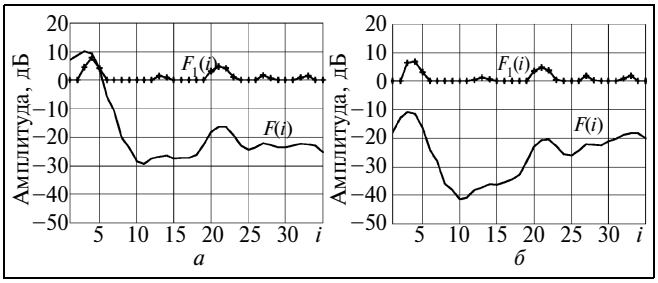

Рис. 3. Логарифмический амплитудный спектр $F(i)$ для гласного «ы» в слове «четыре» и результат его обработки $F_1(i)$ (отмечен крестиками) (*а*), те же зависимости после дифференцирования сигнала (*б*)

му искажению спектра $F(i)$, однако это различие практически отсутствует у обработанных спектров $F_1(i)$.

Полученные результаты показывают, что для проведения сравнения речевых образцов с эталонами использование обработанного логарифмического спектра $F_1(f, t)$ имеет явное преимущество перед использованием спектра $F(f, t)$, обеспечивая устойчивость к частотным искажениям речевого сигнала.

3. ВЫДЕЛЕНИЕ ЛОКАЛЬНЫХ НЕОДНОРОДНОСТЕЙ СПЕКТРА ВО ВРЕМЕНИ

Выделение слуховым анализатором спектральных неоднородностей сигнала во времени подтверждается наличием в слуховой системе фазических нейронов, избирательно реагирующих на начало и конец акустического стимула. Такие реакции нейронов обычно объясняются обработкой спектра $S(f, t)$ временным окном $\varphi_1(t)$, являющимся импульсной характеристикой дифференцирующего фильтра, вычисляющего сглаженную разность первого порядка. В результате получается обработанный спектр

$$S_2(f, t) = S(f, t) \otimes \varphi_1(t) \approx \partial S(f, t)/\partial t.$$

Типичный вид функции $\varphi_1(t)$ приведен на рис. 4, *а*. На рис. 4, *б* показан модуль преобразования Фурье $\varphi_1(t)$, свидетельствующий о том, что $\varphi_1(t)$ является импульсной характеристикой полосового фильтра. Учитывая сказанное, можно заключить, что рассмотренная ранее обработка спектра сводится к нахождению скорости изменения огибающей амплитудного спектра по времени $\partial S(f, t)/\partial t$ и представляет собой разновидность полосовой фильтрации временной огибающей амплитудного спектра $S(f, t)$, реализующей нахождение сглаженной разности спектра по времени $\Delta_t S(f, t) = (S(f, t + \Delta t) - S(f, t - \Delta t)) \otimes w_2(t)$, где $w_2(t)$ — сглаживающее окно.

Результат рассмотренной обработки получится более интересным, если ее применить к логарифмическому спектру $F(f, t)$ и найти скорость изменения во времени огибающей логарифмического спектра

$$F_2(f, t) = F(f, t) \otimes \varphi_1(t). \qquad (3)$$

В этом случае огибающая $F_2(f, t)$ становится независимой от частотных искажений и интенсивности сигнала $s(t)$. Действительно, при наличии частотных искажений спектр сигнала представляет собой произведение $S(f, t)\ W(f)$, где $W(f)$ — частотная характеристика, определяющая частотные искажения сигнала. Поэтому $F(f, t) = \log S(f, t) + \log W(f)$ и функция $F_2(f, t)$ оказывается не зависящей от характеристики $W(f)$. Кроме того, при обработке логарифмического спектра окном $\varphi_1(t)$ происходит удаление фоновой стационарной шумовой компоненты, присутствующей в речевом сигнале.

Для демонстрации особенностей преобразования (3) был применен рассмотренный ранее спектральный анализатор, реализованный на основе гребенки из 35-ти полосовых фильтров. С его помощью получались логарифмические спектры $F(i, n)$, где $i = 1, 2, ..., 35$ — номер фильтра анализатора, n — номер спектра. Спектры находились в моменты $n\Delta T$, где $n = 0, 1, 2, ...$, $\Delta T = 8$ мс, чем обеспечивалась необходимая точность воспроизведения логарифмических огибающих в частотных каналах анализатора. Для обработки спектра $F(i, n)$ использовалась весовая функция

$$\varphi_1(m) = \frac{1}{\sum\limits_{m=-5} m} \begin{cases} -1, & m = -5, ..., -2, -1 \\ 1, & m = 1, 2, ..., 5 \\ 0, & \text{в остальных случаях.} \end{cases}$$

Особенности преобразования (3) на примере слова «шесть» поясняются на рис. 5. В верхней части рисунка приведены логарифмический спектр $F(i, n)$ и обработанный спектр $F_2(i, n)$, значения которых переданы уровнем черного. Под ними представлены спектральные огибающие $F(4, n)$ и $F(34, n)$, полученные в низкочастотном и высокочастотном каналах анализатора, а также результаты обработки огибающих $F_2(4, n)$ и $F_2(34, n)$. Из рис. 5 видно, что спектр $F(i, n)$ имеет хорошо выраженную сегментную структуру, отражающую фонемный состав слова. В обработанном спектре $F_2(i, n)$ начала сегментов представлены вертикально ориентированными темными областями, а концы сегментов — вертикально ориентированными светлыми областями. Пики функций $F_2(4, n)$ и $F_2(34, n)$ отмечают границы квазистационарных сегментов огибающих $F(4, n)$ и $F(34, n)$, представляющих отдельные фонемы, причем положительные пики отмечают начала сегментов, а следующие за ними отрицательные пики — их концы. Таким образом, с помощью преобразования (3) в речевом потоке выделяются границы фонемных сегментов и становится возможным измерение длительности акустических событий, являющейся важной характеристикой способа образования согласных звуков и ударности гласных.

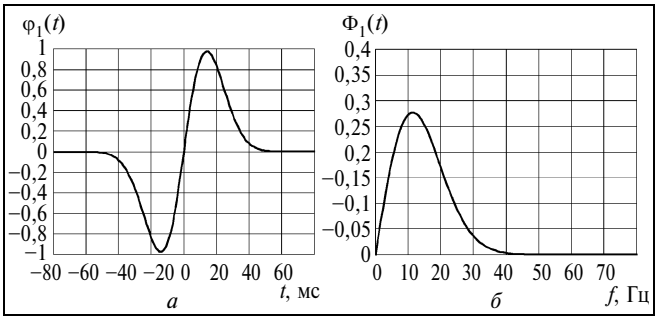

Рис. 4. Типичный вид весовой функции $\varphi_1(f)$ — (*а*), модуль преобразования Фурье $\varphi_1(f)$ — (*б*)

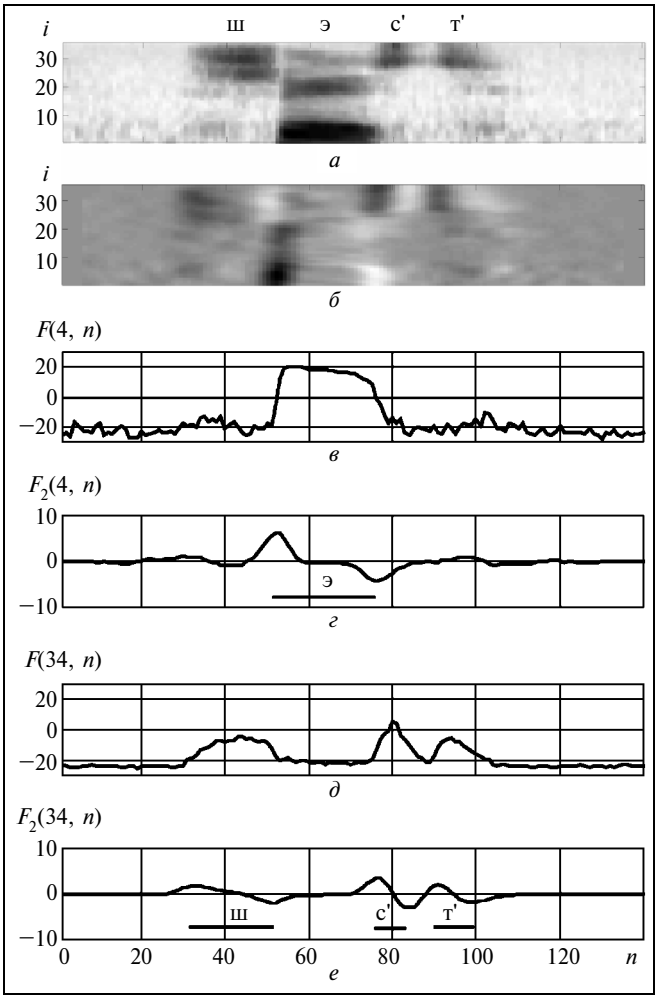

Рис. 5. Обработка логарифмического спектра $F(i, m)$ окном $\varphi_1(t)$ для слова «шесть»:

а — спектрограмма $F(i, n)$; *б* — результат обработки $F_2(i, n)$; *в* — спектральная огибающая $F(4, n)$; *г* — обработанная огибающая $F_2(4, n)$; *д* — спектральная огибающая $F(34, n)$; *е* — обработанная огибающая $F(34, n)$

Поскольку фонемы занимают различные частотные области, то информация об их границах может присутствовать в разных компонентах функции $F_2(i, n)$. Из рис. 5 видно, что информация о границах гласного «э» присутствует в функции $F_2(4, n)$, тогда как границы согласных «ш», «с′» и «т′» (знак «′» обозначает мягкие согласные) отчетливо выражены в функции $F_2(34, n)$. В связи с этим для описания границ речевых звуков более удобно использовать специальную сегментирующую функцию, обобщающую информацию о границах квазистационарных речевых сегментов в различных составляющих $F_2(i, n)$. Для этого можно использовать сегментирующую функцию вида

$$G(n) = \frac{1}{35}\left[\sum_{i=1}^{i=35} F_2^2(i, n)\right]^{1/2}. \qquad (4)$$

На рис. 6, *а* и *б* приведены примеры спектра $F_2(i, n)$ и функции $G(n)$, полученные для случая произнесения слова «шесть». Можно видеть, что функция $G(n)$ имеет характерные положительные пики, положение которых отмечает границы звуков. Недостаток сегментирующей функций $G(n)$ состоит в том, что в ней, в отличие от спектра $F_2(i, n)$, теряется информация о начале и конце артикуляции звука, что не позволяет выделять такие объекты речевого сигнала как смычка, возникающая в результате произнесения взрывного согласного, или провал в огибающей сигнала при артикуляции сонанта «р».

Чтобы обойти это ограничение, можно модифицировать функцию сегментации (4). Для этого ее следует разделить на положительную и отрицательную составляющие, вычислив две функции

$$G_+(n) = \begin{cases} \frac{1}{35}\left[\sum_{i=1}^{i=35} F_2^2(i, n)\right]^{1/2} & \text{при } F_2(i, n) > 0 \\ 0, \text{ иначе} \end{cases}$$

и

$$G_-(n) = \begin{cases} -\frac{1}{35}\left[\sum_{i=1}^{i=35} F_2^2(i, n)\right]^{1/2} & \text{при } F_2(i, n) < 0 \\ 0, \text{ иначе.} \end{cases}$$

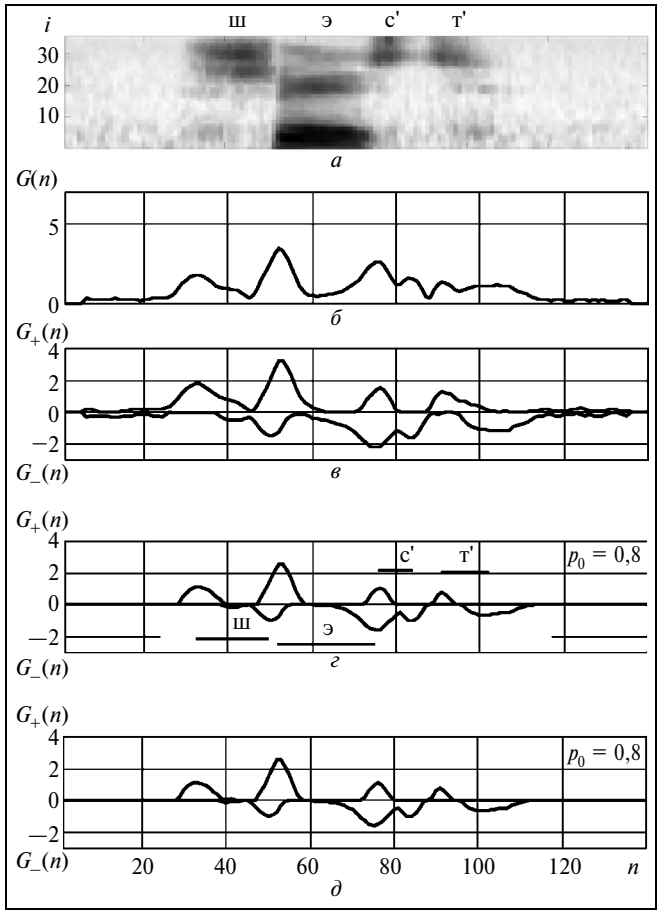

Рис. 6. Сегментация слова «шесть»:

а — спектрограмма $F(i, n)$; *б* — функция $G(n)$; *в* — функции $G_+(n)$ и $G_-(n)$; *г* — функции $G_+(n)$ и $G_-(n)$, модифицированные введением порога $p_0 = 0{,}8$; *д* — функция $G_+(n)$ и $G_-(n)$ для порога $p_0 = 0{,}8$ в случае дифференцирования сигнала

На рис. 6, *в* показаны функции $G_+(n)$ и $G_-(n)$, полученные при произнесении слова «шесть». Из этого примера видно, что начало артикуляции каждого звука отмечается положи тельным пиком в функции $G_+(n)$, а конец его артикуляции отмечается следующим за ним отрицательным пиком в функции $G_-(n)$. Кроме того, глухая смычка (короткая пауза в речевом сигнале) перед артикуляцией взрывного согласного «т′» отмечается отрицательным пиком в функции $G_-(n)$ с последующим положительным пиком в функции $G_+(n)$.

Можно улучшить выраженность пиков в функции сегментации и подавить небольшие ложные пики введением фиксированного порога p_0. Для этого при получении функций $G_+(n)$ и $G_-(n)$ следует суммировать не величины $F_2^2(i, n)$, а $(|F_2(i, n)| - p_0)^2$, для которых модуль $|F_2(i, n)| > p_0$. На рис. 6, *г* приведен вид функций $G_+(n)$ и $G_-(n)$ в случае введения порога $p_0 = 0{,}8$. Видно, что введение порога заметно улучшает выраженность пиков в $G_+(n)$ и $G_-(n)$. На рис. 6, *д* показаны те же функции $G_+(n)$ и $G_-(n)$ при $p_0 = 0{,}8$ в случае дифференцирования сигнала. Можно видеть, что они практически идентичны функциям $G_+(n)$ и $G_-(n)$ на рис. 6, *г*.

Отметим, что все рассмотренные ранее функции сегментации, полученные на основе обработки логарифмического спектра (3), практически не изменялись при наличии частотных искажений речевого сигнала, вызываемых акустикой помещения, регулировкой тембра, сменой микрофона и т. п. Это значит, что с их помощью возможно устойчивое членение речевого потока, предшествующее его фонетической разметке, в присутствии частотных искажений сигнала.

ЗАКЛЮЧЕНИЕ

В настоящей работе рассмотрены два преобразования логарифмического кратковременного спектра речевого сигнала, основанные на полосовой фильтрации, которые учитывают особенности анализа звука в слуховой системе. Их применение в процессе обработки речи снижает вариации спектра, обусловленные характеристиками среды и речевым источником. Эффективность рассмотренных преобразований подтверждается примерами их применения при обработке реальных речевых сигналов.

Первое преобразование выделяет локальные неоднородности логарифмического спектра по частоте и позволяет получить частотное описание речевого сигнала, устойчивое к частотным искажениям и широкополосным аддитивным шумам. Таким образом, в результате применения преобразования становится возможным более надежное сравнение с эталонными описаниями речевых фрагментов при распознавании речи. В основе преобразования лежит разновидность полосовой фильтрации спектра по частоте, используемая для нахождения второй сглаженной разности спектра, и последующая нелинейная обработка, совместно имитирующие эффект латерального торможения в слуховом анализаторе, подчеркивающий максимумы и резкие срезы спектра.

Второе преобразование выделяет локальные неоднородности спектра по времени и может быть применено для сегментации речевого сигнала. Оно представляет собой вариант полосовой фильтрации спектра, реализующей нахождение сглаженной первой разности логарифмической огибающей спектра по времени. В результате преобразования отмечаются моменты начала и конца артикуляторных событий. Подобное преобразование спектра имеет место в слуховом анализаторе, о чем свидетельствует присутствие так называемых тонических нейронов, избирательно реагирующих на начало и конец акустического стимула. На основе преобразования предложена устойчивая к частотным искажениям процедура сегментации речевого сигнала, позволяющая выделять в непрерывном речевом сигнале последовательные квазистационарные сегменты, необходимые для его фонетической разметки.

Возможные области применения рассмотренных преобразований — распознавание речи, идентификация и верификация диктора, и др.

ЛИТЕРАТУРА

1. *Фант Г.* Акустическая теория речеобразования. — М.: Наука, 1964.
2. *Фланаган Дж.* Анализ, синтез и восприятие речи. — М.: Связь, 1968.
3. *Stevens K. N.* Acoustic correlates of some phonetic categories // J. Acoust. Soc. Amer. — 1980. — Vol. 68. — P. 836—842.
4. *Физиология* речи. Восприятие речи человеком / Л. А. Чистович, А. В. Венцов, М. П. Гранстрем и др. // В серии «Руководство по физиологии». — Л.: Наука, 1976.
5. *Picone J. W.* Signal modeling techniques in speech recognition // Proc. IEEE. 1993. — Vol. 81. — N 9. — P. 1215—1247.
6. *Zwicker E., Terhardt E.* Analytical expressions for critical — band rate and critical bandwidth as a function of frequency // J. Acoust. Soc. Amer. — 1980. — Vol. 68. — N 5. P. 1523— 1525.
7. *Traunmυller H.* Analytical expressions for the tonotopic sensory scale // Ibid. — 1990. — Vol. 88. — N 1. — P. 97—100.
8. *Варшавский Л. А., Чистович Л. А.* Средние спектры русских гласных фонем // Проблемы физиологической акустики. — 1959. — Т. IV. — С. 181—186.
9. *Любинский И. А., Позин Н. В., Яхно В. П.* Анализ моделей однородного нейронного слоя с латеральными связями // Автоматика и телемеханика. — 1967. — № 10. — С. 168—181.
10. *Колоколов А. С.* Предварительная обработка сигнала для распознавания речи // Автоматика и телемеханика. — 2002. — № 3. — С. 190—198.
11. *Juang B. H., Rabiner L. R., Wilpon J. G.* On the use of bandpass liftering in speech recognition // IEEE Trans. on Acoust., Speech, and Signal Proc. — 1987. — Vol. 35. — N 7. — P. 947—954.
12. *Чайлдерс Д. Дж., Скиннер Д. П., Кемерейт Р. Ч.* Кепстр и его применение при обработке данных // ТИИЭР. — 1977. — Т. 5. — № 10. — С. 5—23.
13. *Колоколов А. С.* Предварительная обработка и сегментация речевого сигнала в частотной области для распознавания речи // Автоматика и телемеханика. — 2003. — № 6. — С. 152—162.

☎ *(495) 334-88-91*

E-mail: kolokolo@ipu.rssi.ru ❑