СБОРНИК **НАУЧНЫХ ТРУДОВ НГТУ. – 2019. – № 3–4 (96).** – 139–164

УДК 004.934.2 DOI: 10.17212/2307-6879-2019-3-4-139-164

ОБЗОР МЕТОДОВ ИЗВЛЕЧЕНИЯ АКУСТИЧЕСКИХ ПРИЗНАКОВ РЕЧИ В ЗАДАЧЕ РАСПОЗНАВАНИЯ ДИКТОРА*

А.В. СУДЬЕНКОВА

630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, магистрант кафедры защиты информации. E-mail: sudenkova.2018@stud.nstu.ru

Биометрические технологии являются перспективным направлением в области информационной безопасности. Голосовая биометрия на сегодняшний день широко распространена, и работы над повышением качества голосовых систем не теряют своей актуальности. Выбор метода извлечения речевых признаков – один из ключевых этапов проектирования голосовых автоматических систем. В статье рассматриваются акустические параметры, обусловленные физиологическими свойствами речевого тракта человека: частота основного тона, огибающая спектра, форманты и антиформанты. Тема статьи касается методов их извлечения. Большую часть составляют различные варианты кепстрального анализа, поскольку именно они наиболее часто встречаются в современных разработках как в виде использования популярных мел-частотных кепстральных коэффициентов, так и в новых модификациях. Также внимание уделяется алгоритмам линейного предсказания, спектрального центроида и вейвлет-анализа. Параметризация речевых характеристик входит в распознавание речи, эмоций, языка, гендера. Хотя в статье перечислены основные подходы извлечения акустических признаков речи с целью распознавания диктора, материал может быть полезен и в вышеперечисленных задачах обработки речевых сигналов

Ключевые слова: распознавание диктора, анализ речи, кепстральные коэффициенты, линейное предсказание, перцептивное линейное предсказание, спектральный центроид, вейвлет-анализ

ВВЕДЕНИЕ

Биометрические технологии активно внедряются в жизнь общества. Об этом свидетельствует существующий и прогнозируемый рост рынка биометрии как на мировом, так и на отечественном уровне [1].

Распознавание по голосу благодаря широкой доступности оборудования, возможности удаленной идентификации, простому процессу обучения и ис-

* Статья получена 26 сентября 2019 г.

пользования для потребителя является популярной биометрикой, применяемой в области информационной безопасности. Повышение качества распознавания личности в различных условиях и противодействие спуфинговым атакам остаются актуальными проблемами речевой обработки сигналов. Значимый компонент автоматических систем голосовой биометрии – извлечение информативных параметров речевого сигнала. Целью настоящей статьи является предоставление обзора существующих методов извлечения индивидуальных характеристик речи в задачах определения диктора по голосу для разработчиков и исследователей, интересующихся цифровой обработкой речевой информации.

1. РЕЧЕВЫЕ ХАРАКТЕРИСТИКИ

Индивидуальность голоса обеспечивается сочетанием поведенческих и физиологических признаков. К поведенческим относят семантику, дикцию, произношение, ритм, интонации и др. Они обусловлены социальными факторами и могут быть довольно изменчивыми в зависимости от ситуации. Более надежными являются анатомические особенности речевого тракта, поэтому для работы автоматического распознавания наиболее адаптированы алгоритмы измерения акустических характеристик.

Акустическая теория речи рассматривает речевую волну как результат работы источника звука и фильтров. Подробное изложение о физиологических процессах речеобразования и моделях речевого тракта можно найти в книгах [2–4]. Здесь же кратко приведены только те параметры, которые участвуют в автоматическом распознавании дикторов.

Характерные черты голоса конкретного человека в цифровой обработке сигналов получают через спектральный анализ речевой волны.

Частота первой гармоники спектра является *частотой основного тона* (основной частотой голоса). Частота основного тона F_0 – обратная величина длительности T_0 одного цикла работы голосовых связок: $F_0 = 1/T_0$. Основная частота определяет высоту голоса – ощущение, связанное с воздействием тона на слуховую систему человека.

Индивидуальность данного параметра объясняется тем, что длительность T_0 зависит от массы и упругости голосовых связок, а также от перепада давления над и под связками. Поэтому пол и возраст диктора оказывают влияние на значения основной частоты.

Каждый человек имеет свой диапазон изменений частоты основного тона. Как правило, для взрослого он составляет от полутора до двух октав. В задаче распознавания личности по голосу необходимо определять базовую основную

частоту, т. е. привычный и удобный для идентифицируемого человека режим работы голосовых связок.

Важную роль в определении индивидуальных голосовых особенностей играют и остальные гармоники, называемые обертонами. Частота конкретного обертона выражается как $n \cdot F_0$, где n – порядковый номер обертона в спектре. В совокупности *огибающая спектра* (линия, соединяющая вершины амплитуд обертонов) отражает регистр, тембр, основную частоту и громкость речи. Ее форма определяется размерами и конфигурациями полостей рта, гортани и носа, взаимным расположением зубов, языка и губ. Спектральная огибающая показывает относительный вклад гармоник в общую энергию речевого сигнала. При анализе речи могут учитываться наклон и скорость спада спектральной огибающей.

Акустические резонансы в голосовом тракте создают пики в огибающей спектра звука. Такие пики называются *формантами*. По *частотам формант* можно анализировать положение артикуляционных органов, что активно используется в фонетическом анализе сказанного. Однако частоты формант зависят не только от воспроизводимых фонем, но и от говорящего: графики спектра воспроизведения одного и того же звука от двух дикторов имеют отличия. Вариативность формантных частот для разных фонем и контекста весьма широк, но для определенного человека в обозначенном контексте фонематическому различию между звуками соответствуют свои различия в спектральной картине.

Кроме частоты, форманты характеризуются шириной. *Ширина (полоса) форманты* ограничивает диапазон частот по обе стороны от частоты форманты, усиление которых составляет не менее 70,7 % от максимального резонансного усиления на частоте форманты и служит мерой частотной избирательности речевого тракта при резонансе. Иногда измерения формант дополняется нахождением *антиформант* – глубоких минимумов спектра сигнала, возникающих при произнесении некоторых звуков речи.

2. ЛИНЕЙНОЕ ПРЕДСКАЗАНИЕ

Линейное предсказание уже продолжительное время остается одним из основных подходов к задачам цифровой обработки речи. Оно может использоваться для оценки периода основного тона, формант и других основных параметров речи. Принцип метода линейного предсказания состоит в том, что участок речевого сигнала можно аппроксимировать линейной комбинацией предыдущих участков сигнала. Предполагается, что речь создается возбуждением линейного изменяющегося во времени фильтра (речевого тракта) слу-

чайным шумом для невокализованных речевых сегментов или последовательностью импульсов для голосовой речи. Упрощенный процесс речеобразования (рис. 1) описывается линейной системой с переменными параметрами и передаточной функцией:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}},$$

где G – коэффициент усиления, a_k – коэффициенты предсказания, p – порядок линейного предсказания [5].

Рис. 1. Структурная схема основной модели речеобразования

Из концепции линейного предсказания зависимость n-го отсчета речевого сигнала $s(n)$ от сигнала возбуждения $u(n)$ для схемы на рис. 1 выражается в виде

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n).$$

Линейный предсказатель с коэффициентами a_k представляется в виде системы с сигналом на выходе:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k).$$

Суть вычислений заключается в нахождении *линейных коэффициентов предсказания* (linear prediction coefficients (codes) – LPC) a_k по речевому сигналу с минимизацией погрешности предсказания. Погрешность предсказания $e(n)$ определяется как

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k).$$

Существует три базовых алгоритма расчета коэффициентов линейного предсказания: ковариационный, автокорреляционный и лестничный (рис. 2). Их подробное описание можно найти в [5], а краткое изложение с сравнительным анализом – в статье [6].

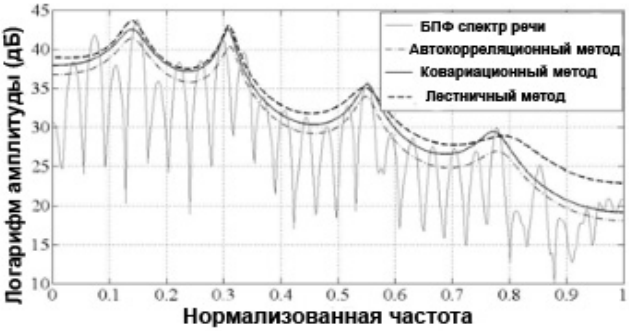

Рис. 2. Сравнение эффективности алгоритмов LPC

3. КЕПСТРАЛЬНЫЙ АНАЛИЗ

Доминирующим алгоритмом обработки речевых сигналов в автоматических системах является нахождение кепстральных коэффициентов. Кепстром называется спектр логарифма спектра временной волны [7], который определяется как

$$c[n] = F^{-1}\{\log | F\{x(n)\} |\},$$

где F и F^{-1} – прямое и обратное дискретное преобразование Фурье (ДПФ).

Целесообразность использования кепстрального анализа в задачах идентификации диктора состоит в том, что кепстр описывает огибающую спектра сигнала в сжатом виде.

Первым этапом речевой сигнал проходит предобработку фильтром, усиливающим высокие частоты спектра, которые обычно уменьшаются в процессе воспроизведения речи. Формула фильтра:

$$x_p(t) = x(t) - ax(t-1),$$

где значение a лежит в интервале [0.95, 0.98]. Такая предобработка не является обязательным компонентом получения кепстра, однако используется во многих системах с кепстральным анализом. Далее сигнал делится на одинаковые последовательные перекрывающиеся временные участки – фреймы. Для ослабления искажений сигнала применяется сглаживающая оконная функция (например, окно Хемминга). Длина окна обычно составляет 20 или 30 миллисекунд, а перекрытие – 10 миллисекунд. Через преобразование Фурье для каждого окна находится спектр, который перемножается со спектром принятого набора фильтров (filterbank) для получения среднего значения в конкретной полосе частот. Затем берется логарифм от полученной огибающей спектра, при необходимости представления амлитуды в децибелах результат логарифма умножается на 20. Последний шаг для получения кепстральных коэффициентов – дискретное косинусное преобразование [8].

Наиболее широкое распространение в цифровой обработке речевых сигналов получили мел-частотные кепстральные коэффициенты – MFCC (mel-frequency cepstral coefficients). Согласно систематическому обзору [9] среди научных публикаций по распознаванию диктора за 2011–2016 годы работы с применением методов MFCC составили 97 %.

Мел-частотный анализ представляет частоты речи с позиции психоакустического параметра слуха – высоты тона. Высота тона определяет, насколько высоким или низким кажется тон слушателю. Нелинейную связь между частотой звука и его высотой отображает мел-частотная шкала (рис. 3). Принято, что высота звука частотой 1000 Гц при уровне 40 дБ равна 1000 мел [10].

Перевод частоты из герц в мел осуществляется по формуле

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right),$$

где f – частота в герцах, Mel – частота в мелах.

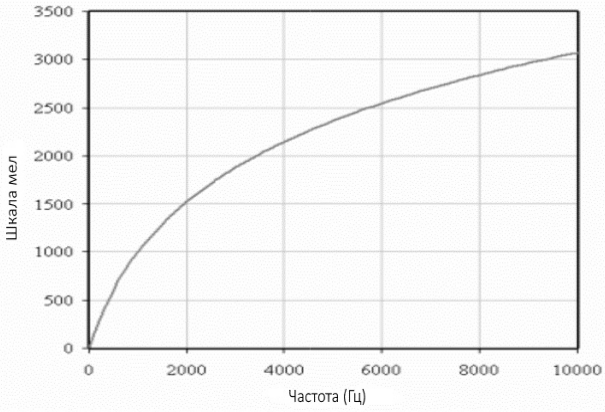

Рис. 3. Мел-частотная шкала

Гребенка фильтров для MFCC – набор треугольных окон в мел-шкале. Поскольку высота тонов, начиная с частоты больше 1 кГц, возрастает гораздо медленнее, высокочастотные фильтры гребенки имеют большую полосу пропускания, чем фильтры низких частот. Идея о том, что при такой фильтрации может быть упущена значимая информация, воплощена в комплементарном к MFCC методе – обратных мел-частотных кепстральных коэффициентах (IMFCC – inverted mel-frequency cepstral coefficients) [11]. Противоположно MFCC в гребенке фильтров для IMFCC низкие частоты имеют большую полосу пропускания.

Также набор треугольных фильтров используется в нахождении линейно-частотных кепстральных коэффициентов (LFCC – linear frequency cepstral coefficients), но фильтры расположены равномерно по линейной полосе частот. Актуальность данного метода объясняется тем, что по теории речеобразования строение речевого тракта, и в частности его длина, отображается в высокочастотной области спектра, которой мало внимания уделяется в MFCC [12].

Также существуют кепстральные коэффициенты прямоугольного набора фильтров (RFCC – rectangular filterbank cepstral coefficients) [13]. Они были предложены для улучшения распознавания речи в условиях шума с учетом эффекта Ломбарда и вдохновлены перцептивным линейным предсказанием, о котором будет сказано далее. Заранее следует отметить, что отличие RFCC от PLP состоит в том, что обработка критических полос слуха производится с помощью однородных непересекающихся прямоугольных фильтров, распределенных по линейной частотной шкале.

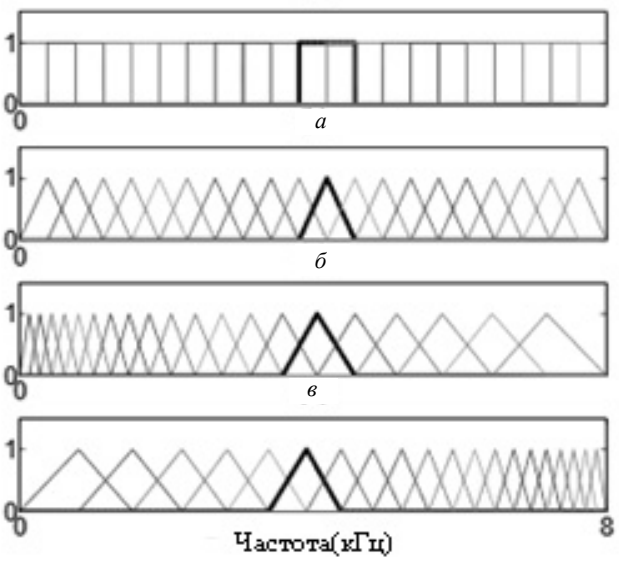

Рис. 4. Наборы фильтров для расчета:
а – RFCC, *б* – LFCC, *в* – MFCC, *г* – IMFCC [14]

Гамматон-частотные кепстральные коэффициенты (GFCC – gammatone frequency cepstral coefficients) получают путем использования гамматон-фильтров (рис. 5), которые считаются стандартной моделью фильтрации ушной улитки [15]. Импульсная характеристика гамматон-фильтра с центральной частотой f:

$$g(f,t) = \begin{cases} t^{a-1} e^{-2\pi f t} \cos(2\pi f t) & if \quad t \geq 0, \\ 0 & else, \end{cases}$$

где t – время, a – порядок фильтра, b – прямоугольная ширина полосы частот, которая возрастает с увеличением центральной частоты f. Алгоритм GFCC не содержит в себе логарифмических операций, поэтому не является кепстральным исходя из определения кепстра. Однако разработчики причислили свой метод к кепстральным из-за функционального сходства с MFCC.

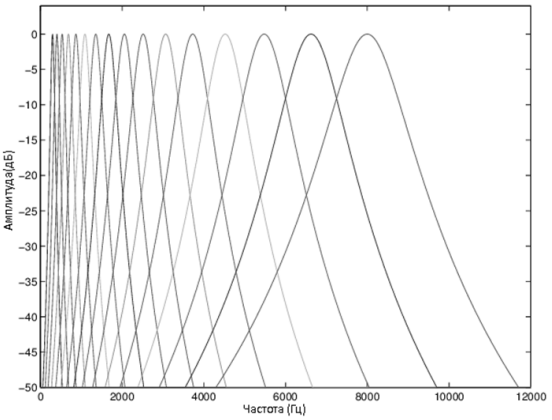

Рис. 5. Набор гамматон-фильтров [16]

Перевод коэффициентов линейного предсказания в кепстральные коэффициенты (LPCC – Linear Predictive Cepstral Coefficients) обеспечивается применением рекурсивной функции [17]:

$$c(n) = \begin{cases} 0, & n < 0, \\ \ln(A), & n = 0, \\ a_n - \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c(k) a_{n-k}, & 0 < n < p, \\ \sum_{k=n-p}^{n-1} \left(\frac{k}{n}\right) c(k) a_{n-k}, & n > p. \end{cases}$$

Другая модификация метода линейного предсказания была предложена в статье [18] исходя из того, что результаты линейного предсказания не соответствуют слуховой значимости компонентов речи. Чтобы устранить эту проблему, перцептивное линейное предсказание (perceptual linear prediction – PLP) принимает во внимание три психоакустических фактора: критические полосы слуха с маскированием, кривую равной громкости, степенное соотношение между громкостью и интенсивностью звука.

Речь делится на кратковременные участки, к которым применяется оконная функция и дискретное преобразование Фурье. Кратковременный спектр мощности $P(\omega)$ складывается из суммы квадратов действительной и мнимой

компонент спектра. Частоты спектра мощности ω (рад/с) переводятся в шкалу барк, описывающую связь между частотой и воспринимаемой высотой тона, по формуле

$$\Omega(\omega) = 6\ln\left(\frac{\omega}{1200\pi} + \left(\frac{\omega^2}{1200\pi} + 1\right)^{0.5}\right).$$

Шкала барк удобна по причине того, что увеличению частот на одну критическую полосу соответствует возрастание высоты тона на один барк (рис. 6).

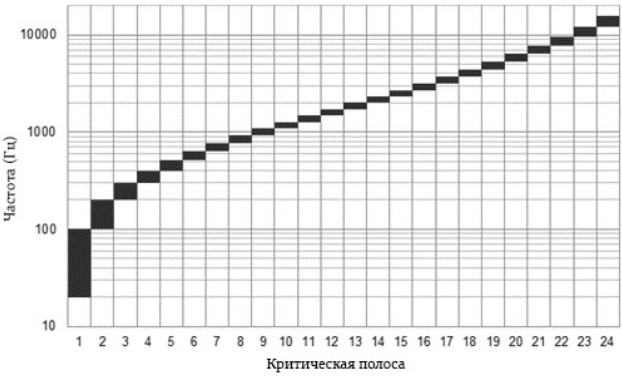

Рис. 6. Критические полосы по шкале барк [19]

Затем полученный спектр мощности перемножается со спектром мощности кривой маскирования критической полосы Ψ:

$$\Psi(\Omega) = \begin{cases} 0, & \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)}, & -1.3 \le \Omega \le -0.5, \\ 1, & -0.5 < \Omega < 0.5, \\ 10^{-1.0(\Omega-0.5)}, & 0.5 \le \Omega \le 2.5, \\ 0, & \Omega > 2.5, \end{cases}$$

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega).$$

Человеческое ухо имеет разное восприятие громкости для звука различных частот (рис. 7). Поэтому далее спектр сглаживается функцией кривой равной громкости $E(\omega)$, являющейся имитацией чувствительности уха на уровне 40 дБ:

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)].$$

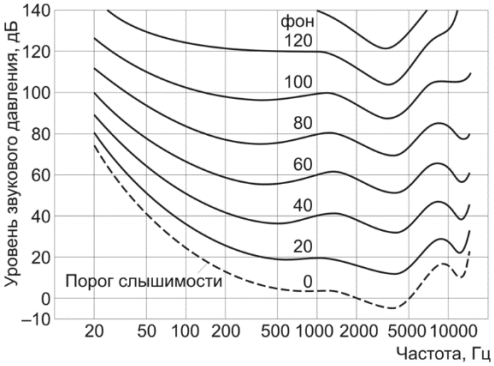

Рис. 7. Стандартные кривые равной громкости чистых тонов при прослушивании в условиях свободного звукового поля [20]

Последний шаг перед применением модели линейного предсказания – извлечение из амплитуды спектра кубического корня:

$$\Phi(\Omega) = \Xi(\Omega)^{0.33}.$$

Таким образом, учитывается закон Стивенса, устанавливающий степенную зависимость между интенсивностью физического стимула и воспринимаемой величиной ощущения, создаваемого стимулом.

Далее, как и в LPCC, идет расчет коэффициентов предсказания и их перевод в кепстральный набор.

Разработка технологии выделения признаков PNCC (power-normalized cepstral coefficients) была инициирована стремлением получить набор практических характеристик для распознавания речи, более надежных в отношении акустической изменчивости в исходной форме, без потери производительности при неискаженном речевом сигнале и со степенью вычислительной сложности, сопоставимой с MFCC и PLP [21].

Начальные стадии обработки PNCC не отличаются от соответствующих этапов нахождения кепстральных коэффициентов, частотный анализ осуществляется с использованием гамма-тон (gammatone) фильтров. Затем следует серия операций по временному анализу: интегрирование по времени для исследования среды, асимметричное шумоподавление, временное маскирование, сглаживание спектра. Эти процессы обеспечивают вычитание шума и некоторую устойчивость по отношению к реверберации.

Далее производится нормирование средней мощности для уменьшения влияния масштабирования амплитуды. Для этого находится оценка средней мощности $\mu(m)$:

$$\mu(m) = \lambda_\mu(m-1) + \frac{1-\lambda_\mu}{L}\sum_{l=0}^{L-1} T(m,\ l),$$

где m и l – индексы фрейма и канала, L – общее количество частотных каналов, $\lambda_\mu = 0.999$, $T(m,\ l)$ – спектральная функция. Нормирование мощности выводится непосредственно из оценки средней мощности путем деления на нее входной мощности:

$$U(m,\ l) = k\frac{T(m,\ l)}{\mu(m)},$$

где значение константы k произвольно.

В следующем шаге, подобно PLP, учитывается степенная зависимость между интенсивностью и громкостью, но показатель степени принят равным 1/15. Причины этого решения изложены в [21]. Завершается алгоритм дискретным косинусным преобразованием.

Относительно новым подходом для задач извлечения признаков можно назвать CQCC (constant Q cepstral coefficients) [22]. Особенностью данного метода извлечения признаков является альтернативный преобразованию Фурье переход из временной области сигнала в частотную с целью улучшения спектрального разрешения (рис. 8). Константное Q-преобразование временного сигнала $x(n)$ определяется в виде

$$X^{CQ}(k,\ n) = \sum_{j=n-\frac{N_k}{2}}^{n+\frac{N_k}{2}} x(j)a_k^*\left(j-n+\frac{N_k}{2}\right),$$

где k = 1, 2, …, K – индекс частотного интервала, N_k – вариативная длина окна, a_k^* – комплексно-сопряженное базисных функций $a_k(n)$.

Базисные функции $a_k(n)$ рассчитываются по формуле

$$a_k(n) = \frac{1}{C}\left(\frac{n}{N_k}\right)\exp\left[i\left(2\pi n\frac{f_k}{f_s} + \Phi_k\right)\right],$$

где f_k – центральная частота частотного интервала k, f_s – частота дискретизации, $w(t)$ – оконная функция, Φ_k – сдвиг фазы, C – коэффициент масштабирования.

Длины окон N_k зависят от Q-фактора:

$$N_k = \frac{f_k}{f_s}Q,$$

где $Q = f_k / (fk_{-1} - f_k)$.

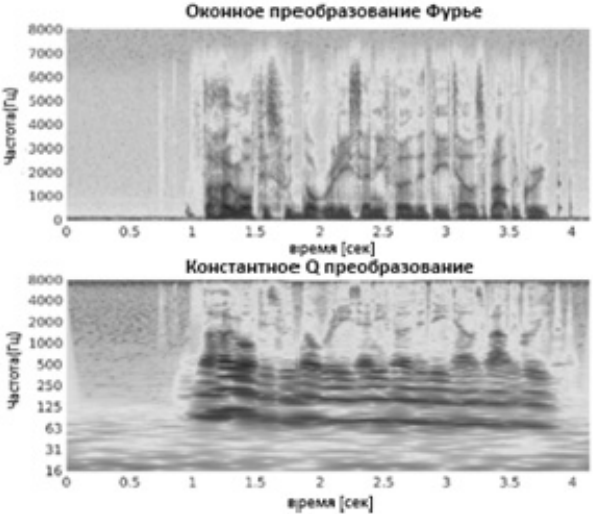

Рис. 8. Спектрограммы одинаковой фразы, полученные оконным преобразованием Фурье и константным Q-преобразованием

Перед применением косинусного преобразования к логарифму энергетического спектра, полученного константным Q-преобразованием, необходимо произвести переход из геометрического пространства в линейное.

Коэффициенты усредненной огибающей Гильберта были разработаны в качестве альтернативы MFCC для распознавания диктора в условиях шума и реверберации. Отличительной чертой данной методики является применение гамматон-фильтров и преобразования Гильберта [23].

Первым делом речевой сигнал проходит через набор гамматон-фильтров. Выходной сигнал в каждом канале фильтра представляет собой свертку речевого сигнала $s(t)$ с импульсной характеристикой в этом канале $h(t, j)$, т. е.

$$s(t, j) = s(t) \cdot h(t, j),$$

Сигнал поддиапазона $s(t, j)$, имеющий структуру амплитудно-частотной модуляции, можно выразить следующим образом:

$$s(t, j) = a(t, j)\cos[\varphi(t, j)],$$

где $a(t, j)$ и $\varphi(t, j)$ – мгновенные амплитудные и фазовые сигналы в j-м канале соответственно. Это показано на рис. 9 для образца речевого сигнала поддиапазона на центральной частоте $f_j = 1000$ Гц, где медленно изменяющаяся огибающая накладывается на частотную модуляцию. Мгновенная частота несущего сигнала является функцией скорости вибрации голосовых складок.

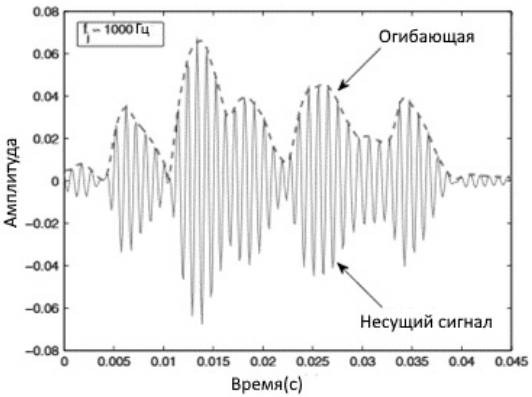

Рис. 9. Пример речевого сигнала поддиапазона на центральной частоте $f_j = 1000$ Гц

Аналитический сигнал можно представить в виде

$$s_a(t, j) = s(t, j) + i\hat{s}(t, j),$$

где $\hat{s}(t, j)$ – результат преобразования Гильберта сигнала $s(t, f)$, i – мнимая единица. Огибающая Гильберта $e_s(t, f)$ является квадратом аналитического сигнала:

$$e_s(t, j) = s^2(t, j) + \hat{s}^2(t, j),$$

Далее огибающая сглаживается с помощью фильтра низких частот:

$$e_{sn}(t, j) = (1 - \alpha)e_s(t, j) + \alpha e_s(t - 1, j),$$

где α – сглаживающий фактор, зависящий от частоты среза фильтра f_c :

$$\alpha = \exp\left(\frac{-2\pi f_c}{F_s}\right).$$

После этого сглаженная огибающая Гильберта делится на фреймы, к которым применяется оконная функция. Для оценки амплитуды в первом фрейме вычисляется среднее значение:

$$S(l, j) = \frac{1}{N}\sum_{t=0}^{N-1} \omega(t)e_{sn}(t, j),$$

где $\omega(t)$ – оконная функция, N – размер фрейма. Конечными этапами алгоритма являются использование логарифма от $S(l, j)$ и дискретное косинусное преобразование.

4. СПЕКТРАЛЬНЫЙ ЦЕНТРОИД

Частота спектрального центроида (spectral centroid frequency – SCF) представляет собой средневзвешенную частоту для данного поддиапазона, где весовые коэффициенты представляют собой нормированную энергию каждо-

го частотного компонента в этом поддиапазоне [24]. По частотам спектрального центроида можно оценить приблизительное местоположение формант, которые проявляются в виде пиков в соседнем поддиапазоне. На SCF влияют изменения основного тона и гармонической структуры. Частота спектрального центроида поддиапазона F_k для k-го поддиапазона определена следующим образом:

$$F_k = \frac{\sum_{f=l_k}^{u_k} f \,|\, S[f]\omega_k[f]\,|}{\sum_{f=l_k}^{u_k} |\, S[f]\omega_k[f]\,|},$$

где u и l – верхняя и нижняя граничная частота поддиапазона; $S[f]$ – спектр фрейма, разделенного на k поддиапазонов, ω_k – частотный отклик фильтра.

Амплитуда спектрального центроида (spectral centroid magnitude – SCM) – это средневзвешенное значение амплитуды для данного поддиапазона, где весовые коэффициенты – это частоты каждого компонента амплитуды в этом поддиапазоне, вычисленные с помощью формулы

$$M_k = \frac{\sum_{f=l_k}^{u_k} f \,|\, S[f]\omega_k[f]\,|}{\sum_{f=l_k}^{u_k} f}.$$

SCM фиксирует в приближении первого порядка распределение энергии в поддиапазоне. Благодаря весовой функции каждый из двух сигналов будет представлен различными значениями SCF и SCM. Также может быть отмечена разность крутизны весовой функции относительно ширины поддиапазона – это приводит к различным дисперсиям элементов. Средняя энергия может быть вычислена с использованием уравнения выше $f = 1$. Поскольку амплитуда спектрального центроида представляет собой амплитуду в положении частоты спектрального центроида, она также будет нести информацию, касающуюся формант, которая полезна для распознавания говорящего.

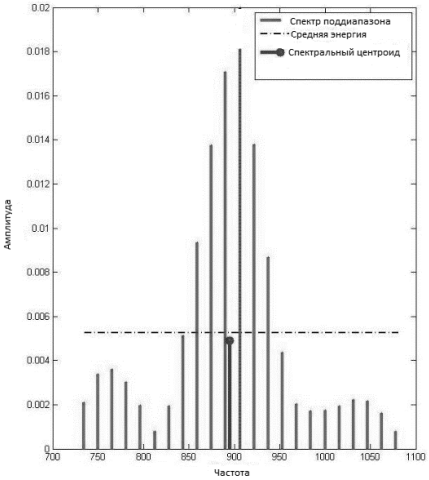

Рис. 10. Спектр поддиапазона, средняя энергия, спектральный центроид для поддиапазона с центральной частотой 906 Гц

5. ВЕЙВЛЕТ-АНАЛИЗ

Для наиболее информативного анализа сложных реальных сигналов необходима обработка как по частотным, так и по временным характеристикам, а также достоверное представление уровней детализации для обнаружения закономерностей. Этим требованиям отвечают вейвлеты – масштабируемые базисные функции преобразования определенной формы.

Идея применения вейвлетов состоит в многомасштабной обработке сигнала, т. е. в анализе сигнала в разном увеличении с разной степенью детализации. Вейвлеты являются семейством функций $\psi_{j,k}(t)$, образованным от базовой функции ψ, называемой материнским вейвлетом:

$$\psi_{j,k}(t) = 2^{j,k}\psi(2^j t - k), \qquad j,k \in Z,$$

где Z – множество целых чисел, j – коэффициент масштаба и k – коэффициент сдвига.

Дискретное вейвлет-преобразование (discrete wavelet transform – DWT) приводит к структуре бинарного дерева (рис. 11). Оно выполняет рекурсивное разложение диапазонов более низких частот, однако для распознавания диктора также требуются некоторые особенности из высокочастотных поддиапа-

зонов. Такое разложение может быть осуществлено с помощью пары фильтров нижних и верхних частот, что достигается посредством вейвлет-пакетного преобразования (wavelet packet transform – WPT).

Благодаря разложению полос и высоких и низких частот в результате WPT получается более сбалансированное бинарное дерево (рис. 12). Каждый узел W_j^p в дереве индексируется по его глубине j и числу p подпространств под ним. Ортогональные основания вейвлет-пакета определены как

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} h[n]\psi_j^p(k-2^j n),$$

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} g[n]\psi_j^p(k-2^j n),$$

где $h[n]$ – фильтр нижних частот, а $g[n]$ – фильтр верхних частот [25].

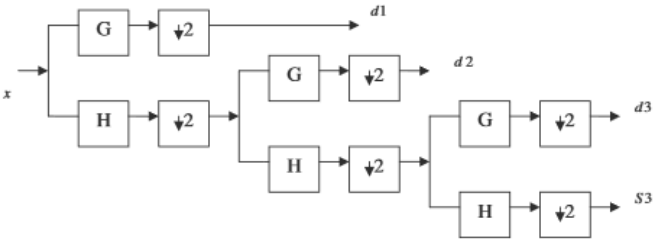

Рис. 11. Бинарное дерево дискретного вейвлет-преобразования [26]

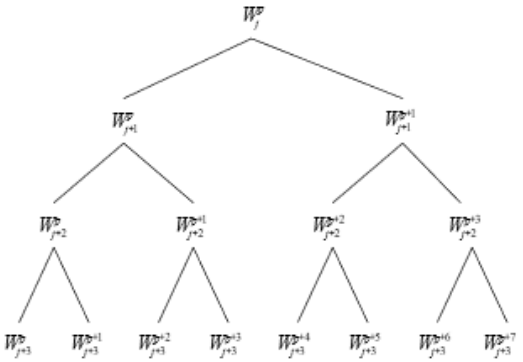

Рис. 12. Бинарное дерево вейвлет-пакетов [27]

ЗАКЛЮЧЕНИЕ

Результат этапа параметризации речи – создание вектора признаков, подаваемого на вход классификатора. Алгоритм MFCC часто встречается в предлагаемых научным сообществом автоматических системах распознавания речи [28–32]. Тем не менее процесс параметризации нередко включает в себя не один метод, а комбинацию из нескольких [33–37]. Большое разнообразие сочетаний подходов можно наблюдать у участников ASVspoof: Automatic Speaker Verification Spoofing and Countermeasures Challenge [38], решающих задачу противодействия спуфинговым атакам.

Предоставленная информация об основных техниках анализа акустических особенностей речи призвана помочь сориентироваться в выборе технологии при проектировании важного элемента речевой системы.

СПИСОК ЛИТЕРАТУРЫ

1. *Крылова И.Ю., Рудакова О.С.* Биометрические технологии как механизм обеспечения информационной безопасности в цифровой экономике // Молодой ученый. – 2018. – № 45. – С. 74–79.

2. *Фант Г.* Акустическая теория речеобразования / пер. с англ. Л.А. Варшавского и В.И. Медведева ; под ред. В.С. Григорьева. – М.: Наука, 1964. – 284 с.

3. *Фланаган Дж.Л.* Анализ, синтез и восприятие речи: пер. с англ. / под ред. А.А. Пирогова. – М.: Связь, 1968. – 396 с.

4. *Кодзасов С.В., Кривнова О.Ф.* Общая фонетика. – М.: Рос. гос. гуманит. ун-т, 2001. – 592 с.

5. *Рабинер Л.Р., Шафер Р.В.* Цифровая обработка речевых сигналов: пер. с англ. / под ред. М.В. Назарова и Ю.Н. Прохорова. – М.: Радио и связь, 1981. – 496 с.

6. *Wang F., Xu W.* A comparison of algorithms for the calculation of LPC coefficients // Proceedings of International Conference on Information Science, Electronics and Electrical Engineering. – Sapporo, Japan, 2014. – P. 300–302.

7. *Oppenheim A., Schafer R.* From frequency to quefrency: a history of the cepstrum // IEEE Signal Process Magazine. – 2004. – Vol. 21, N 5. – P. 95–106.

8. A tutorial on text-independent speaker verification / F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, D.A. Reynolds // EURASIP Journal on Advances in Signal Processing. – 2004. – Vol. 2014, N 4. – P. 430–451.

9. Speaker identification features extraction methods: a systematic review / S. Tirumala, S. Shahamiri, A. Garhwal, R. Wang // Expert Systems with Applications. – 2017. – Vol. 90. – P. 250–271.

10. *Chauhan P.M., Desai N.P.* Mel Frequency Cepstral Coefficients (MFCC) based speaker identification in noisy environment using wiener filter // Proceedings of International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE 2014). – Coimbatore, India, 2014. – P. 1–5. – DOI: 10.1109/ICGCCEE.2014.6921394.

11. *Sharma D., Ali I.* A modified MFCC feature extraction technique for robust speaker recognition // Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI 2015). – Kochi, India, 2015. – P. 1052–1057.

12. Linear versus mel frequency cepstral coefficients for speaker recognition / X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, S. Shamma // Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2011). – Waikoloa, HI, USA, 2011. – P. 559–564. – DOI: 10.1109/ASRU.2011.6163888.

13. *Boril H., Hansen J.* Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments // IEEE Transactions on Audio, Speech, and Language Processing. – 2010. – Vol. 18, N 6. – P. 1379–1393.

14. *Sahidullah M., Kinnunen T., Hanilci C.* A comparison of features for synthetic speech detection // Proceedings of Interspeech (ISCA 2015). – Dresden, Germany, 2015. – P. 2087–2091.

15. *Shao Y., Wang D.L.* Robust speaker identification using auditory features and computational auditory scene analysis // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008). – Las Vegas, NV, USA, 2008. – P. 1589–1592. – DOI: 10.1109/ICASSP.2008.4517928.

16. Frame theory for signal processing in psychoacoustics / P. Balazs, N. Holighaus, T. Necciari, D. Stoeva // Applied and Numerical Harmonic Analysis. – 2017. – Vol. 5. – P. 225–268.

17. *Bhattacharjee U.* A comparative study of LPCC and MFCC features for the recognition of assamese phonemes // International Journal of Engineering Research & Technology (IJERT). – 2013. – Vol. 2, iss. 1.

18. *Hermansky H.* Perceptual linear predictive (PLP) analysis of speech // The Journal of the Acoustical Society of America. – 1990. – Vol. 87, N 4. – P. 1738–1752.

19. Bark scale // Wikipedia: The Free Encyclopedia: website. – URL: https://en.wikipedia.org/w/index.php?title=Bark_scale&oldid=904712246 (accessed: 18.12.2019).

20. ГОСТ Р ИСО 226–2009. Акустика. Стандартные кривые равной громкости: дата введения 2010–12–01. – М.: Стандартинформ, 2010.

21. *Kim C., Stern R.* Power-normalized cepstral coefficients (PNCC) for robust speech recognition // IEEE/ACM Transaction on Audio, Speech, and Language Processing. – 2016. – Vol. 24, N 7. – P. 1315–1329.

22. *Todisco M., Delgado H., Evans N.* A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients // Odyssey 2016. The Speaker and Language Recognition Workshop. – At Bilbao, Spain, 2016. – P. 283–290.

23. *Sadjadi S.O., Hansen J.H.* Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification // Speech Communication. – 2015. – Vol. 72. – P. 138–148.

24. Investigation of spectral centroid magnitude and frequency for speaker recognition / J.M.K. Kua, T. Tharmarajah, M. Nosratighods, E. Ambikairajah, J. Epps // Odyssey 2010. The Speaker and Language Recognition Workshop. – Brno, Czech Republic, 2010. – P. 34–39.

25. *Deshpande M., Holambe R.* Speaker identification using admissible wavelet packet based decomposition // International Journal of Electrical and Computer Engineering. – 2010. – Vol. 4, N 1. – P. 83–86.

26. Speaker identification system using wavelet transform and neural network / K. Daqrouq, T. Abu Hilal, M. Sherif, S. El-Hajjar, A. Al-Qawasmi // Proceedings of International Conference on Advances in Computational Tools for Engineering Applications (ACTEA 2009). – Beirut, Lebanon, 2009. – P. 559–564.

27. *Ganchev T., Siafarikas M., Fakotakis N.* Speaker verification based on wavelet packets // Proceedings of Text, Speech and Dialogue (TSD 2004). – Brno, Czech Republic, 2004. – P. 299–306.

28. *Kang W.H., Kim N.S.* Unsupervised learning of total variability embedding for speaker verification with random digit strings // Applied Sciences. – 2019. – Vol. 9, N 8.

29. *Michelsanti D., Tan Z.* Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification // Proceedings of Interspeech 2017. – Stockholm, Sweden, 2017. – P. 2008–2012.

30. *Wang Y., Lawlor B*. Speaker recognition based on MFCC and BP neural networks // Proceedings of 28th Irish Signals and Systems Conference (ISSC 2017). – Killarney, Co. Kerry, Ireland, 2017. – P. 1–4.

31. Deep neural network embeddings for text-independent speaker verification / D. Snyder, D. Garcia-Romero, D. Povey, S. Khudanpur // Proceedings of Interspeech 2017. – Stockholm, Sweden, 2017. – P. 999–1003.

32. *Ozaydin S.* Design of a text independent speaker recognition system // International Conference on Electrical and Computing Technologies and Applications (ICECTA 2017). –Ras Al Khaimah, UAE, 2017. – P. 1–5.

33. *Daqrouq K., Tutunji T.A.* Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers // Applied Soft Computing. – 2015. – Vol. 27. – P. 231–239.

34. Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions / A.K.H. Al-Ali, D. Dean, B. Senadji, V. Chandran, G.R. Naik // IEEE Access. – 2017. – Vol. 5. – P. 15400–15413.

35. *Chelali F.Z., Djeradi A.* Text dependant speaker recognition using MFCC, LPC and DWT // International Journal of Speech Technology. – 2017. – Vol. 20, N 3. – P. 725–740.

36. *Mohammadi M., Sadegh Mohammadi H.R.* Robust features fusion for text independent speaker verification enhancement in noisy environments // Proceedings of Iranian Conference on Electrical Engineering (ICEE 2017). – Tehran, Iran, 2017. – P. 1863–1868.

37. Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification / M.T.S. Al-Kaltakchi, W.L. Woo, S.S. Dlay, J.A. Chambers // Proceedings of 4th International Conference on Biometrics and Forensics (IWBF). – Limassol, Cyprus, 2016. – P. 1–6.

38. The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection / T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, K.A. Lee // Proceedings of Interspeech 2017. – Stockholm, Sweden, 2017. – P. 2–6.

Судьенкова Анна Владимировна, магистрант кафедры защиты информации факультета автоматики и вычислительной техники Новосибирского государственного технического университета. Область научных интересов: цифровая обработка сигналов, интеллектуальные измерительные системы. E-mail: sudenkova.2018@stud.nstu.ru.

DOI: 10.17212/2307-6879-2019-3-4-139-164

Overview of methods for extracting acoustic speech features in speaker recognition*

A.V. Sudjenkova

Novosibirsk State Technical University, 20 Karl Marx Prospekt, Novosibirsk, 630073, master student department of information security. E-mail: sudenkova.2018@stud.nstu.ru.

Biometric technologies are a perspective direction in the field of information security. Voice biometry is very popular nowadays, and works on improving the quality of voice systems does not lose its relevance. The choice of a method for extracting speech features is one of the key steps in the design of voice automation systems. The article considers acoustic parameters caused by physiological properties of a human speech tract: fundamental frequency, spectral envelope, formants and antiformants. The topic of the article focuses on methods of their extraction. Most of them are different variants of cepstral analysis, because they are the most common in modern developments, both in the form of popular mel-frequency cepstral coefficients and in new modifications. Attention is also paid to linear prediction algorithms, spectral centroid and wavelet analysis. Parameterization of speech characteristics is included in recognition of speech, emotions, language, gender. Although the article contains a list of the main approaches to the extraction of acoustic features of speech in order to recognize the speaker, the content can be useful in the above tasks of processing speech signals.

Keywords: speaker recognition, speech analysis, cepstral coefficients, linear prediction, perceptual linear predictive, spectral centroid, wavelet analysis

REFERENCES

1. Krylova I.Yu., Rudakova O.S. Biometricheskie tekhnologii kak mekhanizm obespecheniya informatsionnoi bezopasnosti v tsifrovoi ekonomike [Biometric technologies as a mechanism for ensuring information security in the digital economy]. *Molodoi uchenyi – Young Scientist*, 2018, no. 45, pp. 74–79.

2. Fant G. *Acoustic theory of speech production.* The Hague, Mouton, 1960 (Russ. ed.: Fant G. *Akusticheskaya teoriya recheobrazovaniya*. Moscow, Nauka Publ., 1964. 284 p.).

3. Flanagan J.L. *Speech analysis, synthesis and perception*. Berlin, Springer-Verlag, 1965. (Russ. ed.: Flanagan Dzh.L. *Analiz, sintez i vospriyatie rechi.* Moscow, Svyaz' Publ., 1968. 396 p.).

4. Kodzasov S.V., Krivnova O.F. *Obshchaya fonetika* [General phonetics]. Moscow, Russian State University for the Humanities Publ., 2001. 592 p.

5. Rabiner L.R., Schafer R.W. *Digital processing of speech signal.* New Jersey, Prentice-Hall, 1978 (Russ. ed.: Rabiner L.R., Shafer R.V. *Tsifrovaya obrabotka rechevykh signalov*. Moscow, Radio i svyaz' Publ., 1981. 496 p.).

* *Received 26 September 2019.*

6. Wang F., Xu W. A comparison of algorithms for the calculation of LPC coefficients. *Proceedings of International Conference on Information Science, Electronics and Electrical Engineering*, Sapporo, Japan, 2014, pp. 300–302.

7. Oppenheim A., Schafer R. From frequency to quefrency: a history of the cepstrum. *IEEE Signal Process Magazine*, 2004, vol. 21, no. 5, pp. 95–106.

8. Bimbot F., Bonastre J., Fredouille C., Gravier G., Magrin-Chagnolleau I., Meignier S., Merlin T., Ortega-García J., Petrovska-Delacrétaz D., Reynolds D.A. A tutorial on text-independent speaker verification. *EURASIP Journal on Advances in Signal Processing*, 2004, vol. 2014, no. 4, pp. 430–451.

9. Tirumala S., Shahamiri S., Garhwal A., Wang R. Speaker identification features extraction methods: a systematic review. *Expert Systems with Applications*, 2017, vol. 90, pp. 250–271.

10. Chauhan P.M., Desai N.P. Mel Frequency Cepstral Coefficients (MFCC) based speaker identification in noisy environment using wiener filter. *Proceedings of International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE 2014)*, Coimbatore, India, 2014, pp. 1–5. DOI: 10.1109/ICGCCEE.2014.6921394.

11. Sharma D, Ali I. A modified MFCC feature extraction technique For robust speaker recognition. *Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI 2015)*, Kochi, India, 2015, pp. 1052–1057.

12. Zhou X., Garcia-Romero D., Duraiswami R., Espy-Wilson C., Shamma S. Linear versus mel frequency cepstral coefficients for speaker recognition. *Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2011)*, Waikoloa, HI, USA, 2011, pp. 559–564. DOI: 10.1109/ASRU.2011.6163888.

13. Boril H., Hansen J. Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, vol. 18, no. 6, pp. 1379–1393.

14. Sahidullah M., Kinnunen T., Hanilci C. A comparison of features for synthetic speech detection. *Proceedings of Interspeech 2015*, Dresden, Germany, 2015, pp. 2087–2091.

15. Shao Y., Wang D.L. Robust speaker identification using auditory features and computational auditory scene analysis. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, Las Vegas, NV, USA, 2008, pp. 1589–1592. DOI: 10.1109/ICASSP.2008.4517928.

16. Balazs P., Holighaus N., Necciari T., Stoeva D. Frame theory for signal processing in psychoacoustics. *Applied and Numerical Harmonic Analysis*, 2017, vol. 5, pp. 225–268.

17. Bhattacharjee U. A comparative study of LPCC and MFCC features for the recognition of assamese phonemes. *International Journal of Engineering Research & Technology (IJERT)*. – 2013. – vol. 2, iss. 1.

18. Hermansky H. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 1990, vol. 87, no. 4, pp. 1738–1752.

19. Bark scale. *Wikipedia: The Free Encyclopedia*: website. Available at: https://en.wikipedia.org/w/index.php?title=Bark_scale&oldid=904712246 (accessed 18.12.2019).

20. State standard R ISO 226–2009. *Acoustics. Normal equal-loudness-level contours*. Moscow, Standartinform Publ., 2010. (In Russian).

21. Kim C., Stern R. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transaction on Audio, Speech, and Language Processing*, 2016, vol. 24, no. 7, pp. 1315–1329.

22. Todisco M., Delgado H., Evans N. A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients. *Odyssey 2016: The Speaker and Language Recognition Workshop*, At Bilbao, Spain, 2016, pp. 283–290.

23. Sadjadi S.O., Hansen J.H. Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification. *Speech Communication*, 2015, vol. 72, pp. 138–148.

24. Kua J.M.K., Tharmarajah T., Nosratighods M., Ambikairajah E., Epps J. Investigation of spectral centroid magnitude and frequency for speaker recognition. *Odyssey 2010. The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010, pp. 34–39.

25. Deshpande M., Holambe R. Speaker identification using admissible wavelet packet based decomposition. *International Journal of Electrical and Computer Engineering*, 2010, vol. 4, no. 1, pp. 83–86.

26. Daqrouq K., Abu Hilal T., Sherif M., El-Hajjar S., Al-Qawasmi A. Speaker identification system using wavelet transform and neural network. *Proceedings of International Conference on Advances in Computational Tools for Engineering Applications (ACTEA 2009)*, Beirut, Lebanon, 2009, pp. 559–564.

27. Ganchev T., Siafarikas M., Fakotakis N. Speaker Verification Based on Wavelet Packets. *Proceedings of Text, Speech and Dialogue (TSD 2004)*, Brno, Czech Republic, 2004, pp. 299–306.

28. Kang W.H., Kim N.S. Unsupervised learning of total variability embedding for speaker verification with random digit strings. *Applied Sciences*, 2019, vol. 9, no. 8.

29. Michelsanti D., Tan Z. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. *Proceedings of Interspeech 2017*, Stockholm, Sweden, 2017, pp. 2008–2012.

30. Wang Y., Lawlor B. Speaker recognition based on MFCC and BP neural networks. *Proceedings of 28th Irish Signals and Systems Conference (ISSC 2017)*, Killarney, Co. Kerry, Ireland, 2017, pp. 1–4.

31. Snyder D., Garcia-Romero D., Povey D., Khudanpur S. Deep neural network embeddings for text-independent speaker verification. *Proceedings of Interspeech 2017*, Stockholm, Sweden, 2017, pp. 999–1003.

32. Ozaydin S. Design of a text independent speaker recognition system. *International Conference on Electrical and Computing Technologies and Applications (ICECTA 2017)*, Ras Al Khaimah, UAE, 2017, pp. 1–5.

33. Daqrouq K., Tutunji T.A. Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers. *Applied Soft Computing*, 2015, vol. 27, pp. 231–239.

34. Al-Ali A.K.H., Dean D., Senadji B., Chandran V., Naik G.R. Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions. *IEEE Access*, 2017, vol. 5, pp. 15400–15413.

35. Chelali F.Z., Djeradi A. Text dependant speaker recognition using MFCC, LPC and DWT. *International Journal of Speech Technology*, 2017, vol. 20, no. 3, pp. 725–740.

36. Mohammadi M., Sadegh Mohammadi H.R. Robust features fusion for text independent speaker verification enhancement in noisy environments. *Proceedings of Iranian Conference on Electrical Engineering (ICEE 2017)*, Tehran, Iran, 2017, pp. 1863–1868.

37. Al-Kaltakchi M.T.S., Woo W.L., Dlay S.S., Chambers J.A. Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification. *Proceedings of 4th International Conference on Biometrics and Forensics (IWBF)*, Limassol, Cyprus, 2016, pp. 1–6.

38. Kinnunen T., Sahidullah M., Delgado H., Todisco M., Evans N., Yamagishi J., Lee K.A. The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection. *Proceedings of Interspeech 2017*, Stockholm, Sweden, 2017, pp. 2–6.

Для цитирования:

Судьенкова А.В. Обзор методов извлечения акустических признаков речи в задаче распознавания диктора // Сборник научных трудов НГТУ. – 2019. – № 3–4 (96). – С. 139–164. – DOI: 10.17212/2307-6879-2019-3-4-139-164.

For citation:

Sudjenkova A.V. Obzor metodov izvlecheniya akusticheskikh priznakov rechi v zadache raspoznavaniya diktora [Overview of methods for extracting acoustic speech features in speaker recognition]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Transaction of scientific papers of the Novosibirsk state technical university*, 2019, no. 3–4 (96), pp. 139–164. DOI: 10.17212/2307-6879-2019-3-4-139-164.