А. Н. Голубинский, С. В. Дворянкин

ОБРАБОТКА АКУСТИЧЕСКИХ СИГНАЛОВ ПРИ КОНТАКТНО-РАЗНОСТНОМ МЕТОДЕ ИДЕНТИФИКАЦИИ ЛИЧНОСТИ

Речевой сигнал широко применяется в разнообразных системах управления, безопасности и связи [1, 2], при этом качество работы систем безопасности и управления, основанных на аутентификации (идентификации и верификации) личности по голосу будет тем выше, чем точнее модель, описывающая речевой сигнал [3–6].

Одним из существенных недостатков известных систем идентификации и верификации по голосу является трудность сохранения в тайне речевого сигнала как биометрического образа, а также малая степень защиты от имитации голоса с помощью различных звуковоспроизводящих устройств [7]. Это обусловлено тем, что речевой сигнал представляет собой изменения давления воздушной среды его распространения, формируемые речевым трактом человека. Использование современных звукозаписывающих и звуковоспроизводящих устройств позволяет злоумышленнику фальсифицировать в процедурах аутентификации биометрический образ зарегистрированного в системе пользователя.

Развитием идентификации личности по голосу является контактно-разностный метод [8], использующий акустические характеристики человеческого тела в качестве биометрического параметра для идентификации [7]. При этом акустическую модель тела человека можно представить в виде сложной уникальной системы проводников звукового сигнала, формируемого в носоглотке человека при произнесении каких-либо звуков. Использование специальных датчиков ларингофонного типа позволяет регистрировать звуковые сигналы, распространяющиеся через биологические жидкости, мягкие и твердые ткани человека, с последующим формированием индивидуального биометрического образа. При идентификации сигналы снимаются с выбранной области регистрации колебаний — это может быть, например, голова, плечо, локоть, запястье руки, колено и т. п. Следует отметить, что для различных точек наблюдения (областей регистрации) сигнала, вследствие разных трактов звукопередачи, акустические характеристики принимаемых колебаний будут отличаться. Таким образом, если злоумышленник не знает область регистрации акустического сигнала, то фальсификация такого принимаемого колебания существенно усложняется.

Сигнал, регистрируемый ларингофоном, описать с достаточной степенью точности весьма проблематично вследствие того, что тело человека является весьма сложной по своей структуре средой распространения звуковых волн. Описание акустических свойств человеческого тела должно учитывать форму и структуру неоднородной биологической среды распространения звуковых волн, включающую в себя: кожу, мышцы, хрящи, костную ткань, жировую ткань, сосуды, кровь и т. д. Вопросы аппроксимации амплитудно-частотной характеристики тела человека исследовались в работе [9].

Цель работы — создание принципов совместной обработки речевых сигналов и акустических сигналов, регистрируемых на теле человека при произнесении им звуков; разработка меры различимости при параметризации биометрического образа амплитудно-частотной характеристикой тела человека применительно к контактно-разностному методу идентификации личности.

При произнесении речевого сигнала с практической полосой частот примерно от $f_{\text{Н}} = 70$ Гц до $f_{\text{В}} = 4$ кГц также распространяются звуковые колебания посредством мышц человека, его костей, мягких тканей, сосудов, кожи и т. д. Регистрируемое колебание в области приема представляет собой сигнал, прошедший некоторый тракт звукопередачи тела человека, интенсивность которого уменьшается из-за влияния различных факторов. Различные спектральные составляющие акустических колебаний, проходящих через тело человека, ослабляются по-разному, вследствие

рассеяния (дифракции) на неоднородностях среды распространения, поглощения звуковых волн в твердых телах, жидкостях и газах, а также ряда других факторов [9]. Затухание звука в диапазоне от 70 Гц до 4 кГц может быть также аномально высоким на некоторых частотах, что интерпретируется рядом низкочастотных релаксационных процессов различных ионов в крови человека; поглощением воздушными пузырьками и газовыми включениями; рассеянием звука на неоднородностях; дифракционными потерями, обусловленными утечкой энергии из канала распространения; нелинейными эффектами, возникающими при распространении звука. Также можно выдвинуть гипотезу о том, что аномально высокое затухание связано с поглощением в пористых средах (резонансные и тепловые эффекты в костных и жировых тканях и др.) и многократным рассеянием звуковых волн на различных неоднородностях тела человека.

Применительно к контактно-разностному методу акустической идентификации частотные свойства человеческого тела можно характеризовать амплитудно-частотной характеристикой (АЧХ) $|K(j\omega)|$, определяемой как отношение давления регистрируемой волны к начальному давлению.

Формально частотный коэффициент передачи (ЧКП), характеризующий уникальность акустического сигнала, сформированного речевым аппаратом и телом человека, можно записать следующим образом:

$$K(j\omega) = \frac{S_{\text{вых}}(j\omega)}{S_{\text{вх}}(j\omega)} = \frac{S_{\text{г}}(j\omega)\cdot K_{\text{т}}(j\omega)\cdot K_{\text{л}}(j\omega)}{S_{\text{г}}(j\omega)\cdot K_{\text{рт}}(j\omega)\cdot K_{\text{изл}}(j\omega)\cdot K_{\text{м}}(j\omega)}, \quad (1)$$

где $S_{\text{ВХ}}(j\omega)$ и $S_{\text{ВЫХ}}(j\omega)$ — спектральные плотности входного и выходного сигналов соответственно; $S_{\text{Г}}(j\omega)$ — спектральная плотность сигнала, генерируемого голосовыми связками (сигнала генератора); $K_{\text{Т}}(j\omega)$ — ЧКП тела человека; $K_{\text{рт}}(j\omega)$ и $K_{\text{изл}}(j\omega)$ — соответственно ЧКП речевого тракта и излучателя (в виде гортани, рта, зубов, языка человека, который излучает акустические сигналы во внешнее пространство); $K_{\text{Л}}(j\omega)$ и $K_{\text{М}}(j\omega)$ — соответственно ЧКП ларингофона и микрофона. В результате АЧХ:

$$|K(j\omega)| = \frac{|K_{\text{т}}(j\omega)|\cdot|K_{\text{л}}(j\omega)|}{|K_{\text{рт}}(j\omega)|\cdot|K_{\text{изл}}(j\omega)|\cdot|K_{\text{м}}(j\omega)|}. \quad (1а)$$

Учитывая, что АЧХ аппаратной части является некоторой известной весовой функцией $|K_{\text{Л}}(j\omega)| / |K_{\text{М}}(j\omega)| = A(\omega)$, перепишем АЧХ человеческого тела в виде:

$$|K(j\omega)| = A(\omega)\frac{|K_{\text{т}}(j\omega)|}{|K_{\text{рт}}(j\omega)|\cdot|K_{\text{изл}}(j\omega)|}. \quad (1б)$$

В общем случае акустические сигналы, распространяющиеся по телу человека, адекватно можно описать математическими моделями, построенными на основе стохастического подхода [10, 11], что обусловлено наложением большого количества различных эффектов (вследствие переизлучения большим количеством неоднородностей тракта распространения волны, поглощения большим количеством поглощающих неоднородностей и т. д.). Таким образом, для характеристики селективных свойств тела человека в ряде случаев удобно использовать частотный коэффициент передачи мощности (ЧКПМ) — квадрат амплитудно-частотной характеристики $K_{\text{р}}(\omega) = |K(j\omega)|^2$ тела человека (отношение интенсивности регистрируемой волны к начальной интенсивности).

ЧКПМ можно рассчитать как отношение нормированной спектральной плотности мощности акустического сигнала ($S_N^{\text{ас}}(\omega)$), регистрируемого на теле человека, к нормированной спектральной плотности мощности речевого сигнала ($S_N^{\text{рс}}(\omega)$):

$$K_P(\omega) = |K(j\omega)|^2 = \frac{S_N^{\text{ас}}(\omega)}{S_N^{\text{рс}}(\omega)}. \quad (2)$$

Вычисление нормированной спектральной плотности мощности $S_N(\omega) = S_N(2\pi f)$ для речевого или акустического сигналов проводят по формулам:

$$S_N(f) = \frac{S(f)}{\max[S(f)]}; \tag{3}$$

$$S(f) = 2\Delta\left[1 + 2\sum_{l=1}^{V-1} R_l W(l)\cos(2\pi f \Delta l)\right], \tag{4}$$

где $\Delta = 1/f_d$ интервал дискретизации, при частоте дискретизации f_d; V — точка отсечения корреляционного окна $W(l)$ для получения состоятельной оценки спектральной плотности мощности. Например, для корреляционного окна Тьюки [12]:

$$W(l) = \begin{cases} \frac{1}{2}\left(1 + \cos\left(\frac{\pi l}{V}\right)\right), & |l| \le V; \\ 0, & |l| > V. \end{cases} \tag{5}$$

Частоту дискретизации можно рассчитать по теореме Котельникова $f_d = 2f_B$, выбрав для f_B = 4 кГц стандартное значение f_d = 8 кГц.

В выражении (4) R_l — коэффициент корреляции центрированного речевого сигнала:

$$R_l = \frac{K_l}{K_0}, \tag{6}$$

где K_l — функция корреляции:

$$K_l = \frac{1}{N}\sum_{i=1}^{N-l}(y_i - \bar{y})(y_{i+l} - \bar{y}); \tag{7}$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad j = \overline{0, L}, \tag{8}$$

здесь N — число отсчетов речевого сигнала; L — число отсчетов коэффициента корреляции;

$$y_i = x_i - \bar{x}, \quad i = \overline{1, N}, \tag{9}$$

где x_i — начальные отсчеты речевого сигнала; $\bar{x}$ — математическое ожидание:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{10}$$

Следует отметить, что можно выделить два направления развития контактно-разностного метода идентификации личности. Первым из них является кратковременный анализ ЧКПМ $K_{\rho L}(\omega)$ при временных интервалах локальной квазистационарности существенных параметров $\Delta t_l \sim 10$ мс (данный интервал может варьироваться в диапазоне, как правило, от 5 до 30 мс). В результате формиру-ется вектор существенных параметров, представляющий собой набор ЧКПМ $\{K_\rho(\omega)\} = (K_{\rho 1}(\omega), K_{\rho 2}(\omega), ..., K_{\rho L}(\omega), ..., K_{\rho L}(\omega))$, усредненных за малые квазистационарные (по параметрам индивидуальности) участки времени. При этом в качестве невязки между существенными параметрами x-го и y-го диктора может выступать, например, сумма модулей разности сравниваемых ЧКПМ ($K_\rho^{(x)}(\omega)$ и $K_\rho^{(y)}(\omega)$) по соответствующим временным интервалам (x_α и y_β):

$$\varepsilon(x_\alpha, y_\beta) = \int_{\omega_н}^{\omega_в} \left| K_{P\alpha}^{(x)}(\omega) - K_{P\beta}^{(y)}(\omega) \right| d\omega, \tag{11}$$

где $\omega_H = 2\pi f_{\text{Н}}$ и $\omega_B = 2\pi f_{\text{В}}$ — соответственно нижняя и верхняя круговые частоты спектра мощности акустического сигнала. Если возможно или целесообразно использование весовых коэффициентов для соответствующих частот $w(\omega)$ (определяющих степень вклада в общую невязку компоненты на соответствующей частоте), то применяется взвешенная сумма модулей разности сравниваемых ЧКПМ по соответствующим временным интервалам:

$$\varepsilon(x_\alpha, y_\beta) = \int_{\omega_{\text{н}}}^{\omega_{\text{в}}} w(\omega) \cdot \left| K_{P\alpha}^{(x)}(\omega) - K_{P\beta}^{(y)}(\omega) \right| \mathrm{d}\omega \,. \tag{11a}$$

Также может быть использована квадратичная невязка, которая удобна для аналитических расчетов ввиду отсутствия модуля:

$$\varepsilon(x_\alpha, y_\beta) = \int_{\omega_{\text{н}}}^{\omega_{\text{в}}} \left(K_{P\alpha}^{(x)}(\omega) - K_{P\beta}^{(y)}(\omega) \right)^2 \mathrm{d}\omega \,, \tag{12}$$

или взвешенная квадратичная невязка:

$$\varepsilon(x_\alpha, y_\beta) = \int_{\omega_{\text{н}}}^{\omega_{\text{в}}} w(\omega) \cdot \left(K_{P\alpha}^{(x)}(\omega) - K_{P\beta}^{(y)}(\omega) \right)^2 \mathrm{d}\omega \,. \tag{12a}$$

Далее обработка для кратковременного анализа по принятию решения о результате идентификации может проводиться, например, с помощью нейронных сетей, скрытых марковских моделей или же гибрида скрытая марковская модель — нейронная сеть.

Второй подход основан на анализе усредненных характеристик на интервалах вокализации речевых сегментов. При этом наиболее эффективно использовать гласные и сонорные согласные звуки, в которых присутствует наибольший вклад голоса (а следовательно, и существенная характеризация уникальности акустических сигналов) при их генерации. В данном случае усредненные ЧКПМ рассчитываются на временном интервале $\Delta t \sim 100$ мс (данный интервал также может варьироваться в диапазоне, как правило, от 50 до 300 мс), что обычно превышает на порядок интервалы при кратковременном анализе. При данном подходе для принятия решения о результате идентификации на основе речевого и акустического сигналов (сформированных для вокализованного сегмента речи) может использоваться, например, среднеквадратическая невязка:

$$\varepsilon(x, y) = \frac{1}{\omega_{\text{в}} - \omega_{\text{н}}} \int_{\omega_{\text{н}}}^{\omega_{\text{в}}} \left(K_P^{(x)}(\omega) - K_P^{(y)}(\omega) \right)^2 \mathrm{d}\omega =$$

$$= \frac{1}{\omega_{\text{в}} - \omega_{\text{н}}} \left\{ \int_{\omega_{\text{н}}}^{\omega_{\text{в}}} \left(K_P^{(x)}(\omega) \right)^2 \mathrm{d}\omega + \int_{\omega_{\text{н}}}^{\omega_{\text{в}}} \left(K_P^{(y)}(\omega) \right)^2 \mathrm{d}\omega - 2 \cdot \int_{\omega_{\text{н}}}^{\omega_{\text{в}}} K_P^{(x)}(\omega) \cdot K_P^{(y)}(\omega) \mathrm{d}\omega \right\}. \tag{13}$$

Следует отметить, что в общем случае функциональная зависимость ЧКПМ $K_{\text{р}}(\omega)$ является довольно сложной, в результате чего целесообразно использовать различного рода аппроксимации, обладающие приемлемой точностью [9].

Учитывая различные виды затухания — вследствие расширения волнового фронта акустической волны; рассеяния на неоднородностях среды распространения; поглощения звуковых волн в твердых телах, а также в жидкостях и газах (что, в свою очередь, обусловлено сдвиговой вязкостью; объемной вязкостью; теплопроводностью среды; релаксационным поглощением), можно предложить следующую степенную аппроксимацию ЧКПМ (квадрата АЧХ) полиномом M-й степени [9]:

$$K_P(2\pi f) \equiv \left|K(j2\pi f)\right|^2 = K_0 \cdot \left(1 - \sum_{m=1}^{M} \gamma_m \cdot f^m\right), \qquad (14)$$

где $\gamma_1, \gamma_2, ..., \gamma_M$ — коэффициенты аппроксимации ЧКПМ тела человека, характеризующие вектор $\{\gamma_m\}$ уникальных биометрических параметров человека; также должно выполняться условие $\sum_{m=1}^{M} \gamma_m \cdot f^m \in [0;1]$.

Расчет вектора уникальных биометрических параметров $\{\gamma_m\}$ осуществляют на основе минимизации невязки (например, по методу наименьших квадратов) между вычисленным экспериментально частотным коэффициентом передачи мощности $K_\rho(\omega)$ и полиномом заданной M-й степени, которая, как правило, принимается равной от шести до десяти.

В режиме идентификации формируется биометрический образ заявляемой личности (в виде ЧКПМ тела человека), далее происходит автоматический ввод образов эталонов из хранимой базы данных. После этого формируется результат сравнения входного биометрического образа (вычисленного на основе входных речевого и акустического сигналов) неизвестного диктора и поступающего образа очередного эталона. Для сравнения входных акустических параметров идентифицируемого диктора и эталона используется мера различимости между ЧКПМ тела идентифицируемого человека и ЧКПМ эталона, которую можно определить как взвешенную Евклидову невязку входных параметров и эталона. Используя аппроксимацию (14) и нормируя ЧКПМ на максимальное значение (соответствующие дикторам x и y константы K_0), для невязки (13) получим следующее выражение:

$$\varepsilon(x,y) = \frac{2\pi}{f_\text{в} - f_\text{н}} \sum_{m=1}^{M} \sum_{v=1}^{M} \left(\gamma_m^{(y)} - \gamma_m^{(x)}\right) \cdot \left(\gamma_v^{(y)} - \gamma_v^{(x)}\right) \cdot \frac{f_\text{в}^{m+v+1} - f_\text{н}^{m+v+1}}{m+v+1}. \qquad (15)$$

Если, например, диктор y является идентифицируемым, а диктор x соответствует n-му эталонному диктору, то невязку, соответствующую мере различимости идентифицируемого и эталонного дикторов, можно переписать следующим образом:

$$D_n = \frac{2\pi}{f_\text{в} - f_\text{н}} \sum_{m=1}^{M} \sum_{v=1}^{M} \left(\gamma_m^{\text{ид}} - \gamma_{n,m}^{\text{эт}}\right) \cdot \left(\gamma_v^{\text{ид}} - \gamma_{n,v}^{\text{эт}}\right) \cdot \frac{f_\text{в}^{m+v+1} - f_\text{н}^{m+v+1}}{m+v+1} = \qquad (15\text{а})$$

$$= \sum_{m=1}^{M} \sum_{v=1}^{M} w_{m,v} \cdot \left(\gamma_m^{\text{ид}} - \gamma_{n,m}^{\text{эт}}\right) \cdot \left(\gamma_v^{\text{ид}} - \gamma_{n,v}^{\text{эт}}\right).$$

Меру различимости (15а) можно рассматривать как частный случай расстояния Махаланобиса [5] с весовыми коэффициентами:

$$w_{m,v} = \frac{2\pi}{f_\text{в} - f_\text{н}} \cdot \frac{f_\text{в}^{m+v+1} - f_\text{н}^{m+v+1}}{m+v+1}. \qquad (16)$$

Предложенная мера различимости обладает существенным преимуществом — весовые коэффициенты могут быть непосредственно рассчитаны исходя из аппроксимирующей функции и физических параметров спектра акустического сигнала, а не вычисляются эмпирически на этапе обучения системы.

Заметим, что в ряде случаев, если аппроксимация ЧКПМ не является полиномиальной (14), а определяется произвольной функциональной зависимостью, невязка как мера различимости идентифицируемого и эталонного дикторов может быть рассчитана по следующему простому выражению:

$$D_n = \sum_{m=1}^{M} w_m \cdot (\gamma_m^{\text{ид}} - \gamma_{n,m}^{\text{эт}})^2 , \quad (17)$$

где w_m — весовые коэффициенты, определяемые на этапе обучения (введения эталонов) системы. Число сравниваемых параметров (M) должно быть равно шести и более. В случае идентификации мера D_n вычисляется для каждого хранящегося в базе данных набора параметров n-го эталонного диктора ($n = 1; N^{\text{БД}}$).

В системе идентификации личности по голосу, тех говорящих, которые заявляют истинную идентичность, можно называть «Своими», в то время как говорящих, которые заявляют ложную идентичность, можно называть «Чужими». При оценке говорящих система идентификации говорящего может делать ошибки двух типов: (а) ложное отклонение и (б) ложный допуск. Ошибка ложного отклонения (ошибка первого рода — вероятность ложной тревоги) имеет место, когда «Свой» заявляет истинную идентичность, но система идентификации говорящего его отвергает. Когда «Чужой» получает допуск с помощью системы идентификации говорящего, имеет место ошибка ложного допуска (ошибка второго рода — вероятность пропуска цели). Также можно характеризовать обнаружение сигнала средней вероятностью ошибки, которая определяется как половина от суммы ошибок первого и второго рода. Решение принять или отвергнуть идентичность зависит от порога идентификации. В зависимости от цены ошибки каждого типа система может быть спроектирована так, чтобы достичь компромисса между одним типом ошибки и другим.

Для принятия решения о результате идентификации необходимо вычислить значение наименьшей меры различимости $\min\limits_n D_n$, которое сравнивается с заранее заданным значением порога идентификации D_0. Порог D_0 выбирается исходя из ошибок первого и второго рода (или средней вероятности ошибки) на этапе практического тестирования системы. Возможна ситуация выбора порога идентификации, при котором получается равный уровень ошибок обоих родов.

Минимум меры различимости D_n ($D_{\min(n)} = \min\limits_n [D_n]$) приводит к решению о соответствии идентифицируемого диктора n-му эталонному диктору из базы данных при условии [2, 5]:

$$\min_n [D_n] \le D_0 , \quad (18)$$

если же

$$\min_n [D_n] > D_0 , \quad (19)$$

то принимается решение о несоответствии идентифицируемого диктора ни одному из имеющихся эталонов.

Использование предлагаемой биометрической параметризации в виде ЧКПМ тела человека позволит повысить потенциальную надежность идентификации личности, при этом обеспечивая высокую помехоустойчивость распознавания при работе с наличием шумов. Это достигается за счет использования существенных уникальных акустических параметров тела человека, таких как числовые коэффициенты амплитудно-частотной характеристики человеческого тела. При этом использование различных областей регистрации акустического сигнала с тела человека позволяет обеспечить повышенную надежность идентификации личности за счет уникальности акустических характеристик звуковых трактов.

Таким образом, предложен способ совместной обработки речевых сигналов и акустических сигналов, регистрируемых на теле человека при произнесении им звуков, применительно к контактно-разностному методу голосовой идентификации. Разработаны новые меры различимости для идентификации личности при параметризации биометрического образа частотным коэффициентом передачи мощности тела человека.

СПИСОК ЛИТЕРАТУРЫ:

1. *Сорокин В. Н.* Фундаментальные исследования речи и прикладные задачи речевых технологий // Речевые технологии. 2008. № 1. С. 18–48.
2. *Болл Р. М., Коннел Дж. Х., Панканти Ш., Рахта Н. К., Сеньор Э. У.* Руководство по биометрии. М.: Техносфера, 2007. – 368 с.
3. *Zavarehei E., Vaseghi S., Yan Q.* Noisy speech enhancement using harmonic-noise model and codebook-based post-processing // IEEE Trans. on Speech and Audio Process. 2007. Vol. 15. № 4. P. 1194–1203.
4. *Levinson S.C.* Mathematical models for speech technology. Chichester: John Wiley & Sons, 2005. – 284 p.
5. *Рабинер Л. Р., Шафер Р. В.* Цифровая обработка речевых сигналов. М.: Радио и связь, 1981. – 496 с.
6. *Назаров М. В., Прохоров Ю. Н.* Методы цифровой обработки и передачи речевых сигналов. М.: Радио и связь, 1985. – 176 с.
7. Патент РФ №2263358: МПК G 10 L 15/06, G 10 L 17/00. Способ автоматического распознавания человека с использованием акустических сигналов, снимаемых с тела человека / С. Л. Бочкарев, В. В. Андрианов, И. В. Бочкарев – № 2003136444/09; Заявл. 11.12.03; Опубл. 27.10.05; Бюл. № 30.
8. *Дворянкин С. В., Гаврилюк С. Б., Калиновский А. Г.* Об удаленной биометрической идентификации обучаемых на основе контактно-разностной речевой обработки парольных фраз // XI Международная научная конференция «Цивилизация знаний: Проблемы модернизации России»: Сборник материалов. М.: РосНОУ, 2010. С. 67–68.
9. *Голубинский А. Н., Дворянкин С. В.* К вопросу о параметризации результатов акустического зондирования тела человека (АЧХ) при реализации контактно-разностного метода аудиоидентификации // Спецтехника и связь. 2011. № 2. С. 38–43.
10. *Айфичер Э. С., Джервис Б. У.* Цифровая обработка сигналов: практический подход. М.: Изд. дом «Вильямс», 2004. – 992 с.
11. *Бендат Дж., Пирсол А.* Прикладной анализ случайных данных. М.: Мир, 1989. – 540 с.
12. *Дженкинс Г., Ваттс Д.* Спектральный анализ и его приложения. М.: Мир, 1971. Вып. 1. – 316 с.