

Д.А. Иванов, А.П. Никитин

Метод текстозависимой аутентификации по голосу

В данной работе был предложен метод биометрической аутентификации пользователя по параметрам его голоса. В качестве основных параметров разрабатываемой модели были выбраны индивидуальные особенности скорости и длительности произношения отдельных элементов речи.

При разработке алгоритма сравнения образцов голоса было обнаружено, что анализ скорости и длительности произношения отдельных элементов речи в текстозависимой процедуре аутентификации пользователя сводится к нахождению процентного содержания совпадающих по содержанию и местоположению отрезков записи с учетом изменяемости общей длины произносимой парольной фразы (сравнение наборов мел-частотных кепстральных коэффициентов).

Ключевые слова: биометрика, аутентификация по голосу, метод мел-частотных кепстральных коэффициентов

В большинстве работ при решении задачи распознавания диктора используются параметры в виде коэффициентов кепстра. Кепстральные коэффициенты вычисляются по огибающей спектра, полученного через преобразование Фурье, с помощью гребенки фильтров. Также возможно вычисление по передаточной функции речевого тракта, найденной методом линейного предсказания.

Для процедуры аутентификации по голосу следует использовать свойства, обладающие высокой индивидуальностью¹. Частотные характеристики отдельных звуков хотя и обладают большим разнообразием, однако из-за явления аккомодации нахождение и выделение самих звуков в случайной записи осложнено, а сами акустические параметры голоса пользователя при произношении некоторых звуков могут не отражать анатомию речевого тракта².

Следует отметить, что относительные скорости произношения звуков, определенные в данной работе как отношение длительности выделенного звука (элемента речи) к длительности всего выбран-

ного отрезка речи, хотя и являются вариативными от состояния человека в спокойной обстановке, имеют относительно небольшую изменяемость для одного индивида³.

Поэтому в данной работе рассматривается текстозависимый вариант метода биометрической аутентификации. Для него можно однозначно определять относительные скорости произношения отрезков речи.

В данной работе были выбраны следующие параметры для проведения процедуры аутентификации: спектральные характеристики парольной фразы и ее отрезков, относительные скорости произношения отрезков фразы, как включающие в себя индивидуальные особенности артикуляции пользователя. Спектральные характеристики самой фразы в данном случае используются для проверки самого пароля.

Следует учитывать, что дальнейшее увеличение частоты дискретизации по достижению значения 32 000 Гц увеличивает точность распознавания только пользователей с высоким, профессионально поставленным голосом. В действительности основной диапазон человеческой речи 80–6000 Гц, причем значения меньше 160 Гц в основном достигаются низкими мужскими голосами. Частоты в диапазоне 6000–16 000 Гц в основном заполняются согласными звонкими или «щелкающими» звуками, учет данного диапазона также увеличивает точность распознавания голоса.

В зависимости от сферы применения следует учитывать и объем анализируемых данных, в общем случае пропорциональный произведению частоты дискретизации, глубины звучания и длительности записи. На рисунке 1 показаны графики зависимости размера односекундной записи от частоты дискретизации для различных значений глубины звучания.

Для увеличения скорости и точности процедуры аутентификации пользователя из записи следует вырезать не содержащие информацию отрезки в начале и конце записи.

Метод мел-частотных кепстральных коэффициентов (далее MFCC – Mel-frequency cepstral coefficients) извлечения признаков является одним из самых распространенных как в системах распознавания дикторов, так и в системах распознавания речи⁴. Он обладает следующими особенностями:

- благодаря использованию спектра сигнала, при дальнейшем анализе учитывается его волновая природа;
- спектр проецируется на мел-шкалу, мел является психофизической единицей высоты звука, это позволяет учитывать особенности восприятия человека и выделять наиболее значимые в речи частоты, а следовательно, содержащие больший объем инфор-

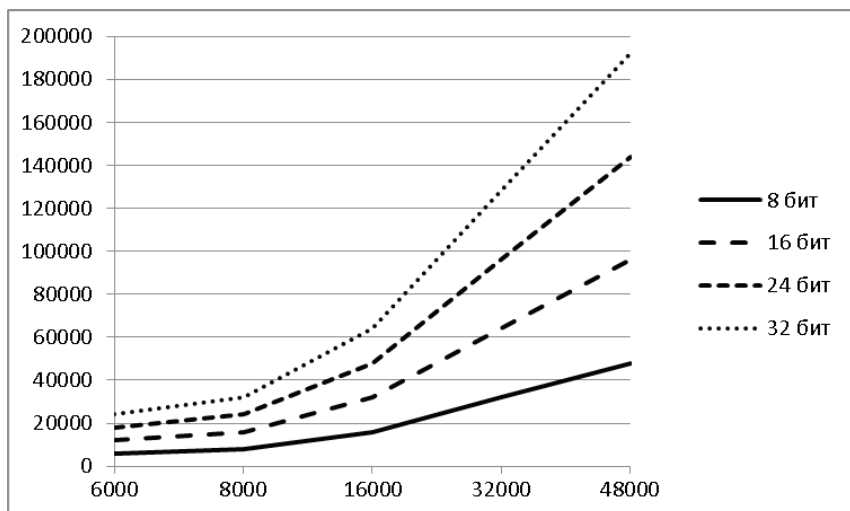


Рис. 1. График зависимости размера односекундной записи от частоты дискретизации для различных значений глубины звучания

- количество вычисляемых коэффициентов ограничивается разрабатчиком системы, что позволяет масштабировать объем анализируемых данных.

На вход алгоритма подается последовательность отсчетов участка сигнала (фреймов), исследуемого на данной итерации, $x_0, \dots, x_{N-1}$. К данной последовательности применяется весовая функция и затем дискретное преобразование Фурье. Весовая функция используется для уменьшения искажений в Фурье-анализе, вызванных конечностью выборки. На практике в качестве весовой функции обычно используется окно Хэмминга, которое имеет вид, отраженный в формуле 1.

$$w_n = 0,54 - 0,46 \times \cos\left(2\pi \frac{n}{N-1}\right), \quad (1)$$

где $n = 0, \dots, N-1$, N — длина фрейма, выраженная в отсчетах. На рисунке 2 показано графическое представление спектра сигнала, полученного с использованием окна Хемминга.

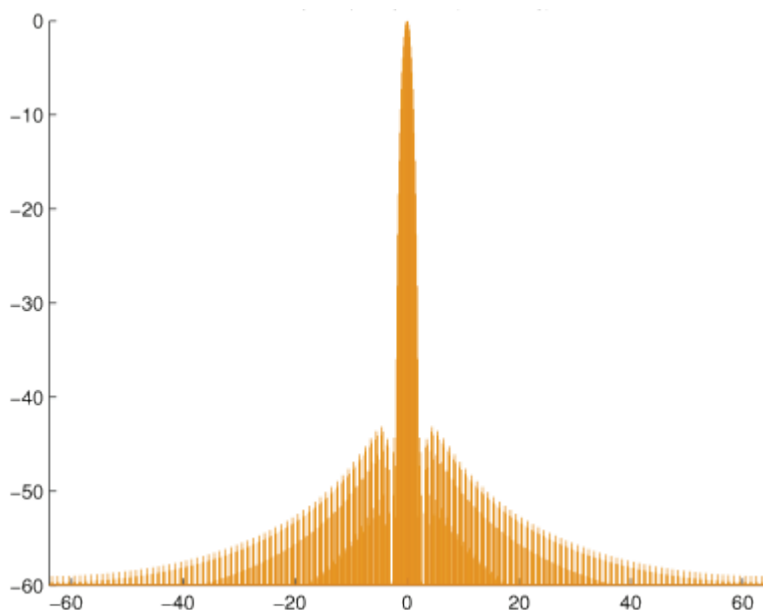


Рис. 2. Графическое представление спектра сигнала, полученного с использованием окна Хемминга

Тогда дискретное преобразование Фурье взвешенного сигнала можно записать в виде формулы (2).

$$X_k = \sum_{n=0}^{N-1} x_n w_n e^{\frac{-2\pi i}{N} kn}, \quad (2)$$

где $k = 0, \dots, N-1$, причем X_k – зависимость магнитуды сигнала от его частоты, такое представление ДПФ и называется оконным преобразованием Фурье.

Полученное представление сигнала в частотной области разбивают на диапазоны с помощью банка (гребенки) треугольных фильтров. Количество фильтров рекомендуется подбирать равным выбранному количеству мел-частотных кепстральных коэффициентов. Границы фильтров рассчитывают в шкале мел.

Перевод в мел-частотную область осуществляют по формуле (3).

$$B(f) = 1127 \times \ln\left(1 + \frac{f}{700}\right). \quad (3)$$

Тогда, $B^{-1}(b) = 700(e^{\frac{b}{1127}} - 1)$ – обратное преобразование.

Пусть N_{FB} – количество фильтров, $(f_{\text{low}}, f_{\text{high}})$ – исследуемый диапазон частот. Тогда данный диапазон переводят в шкалу мел, разбивают на N_{FB} равномерно распределенных перекрывающихся диапазонов и вычисляют соответствующие границы в области линейных частот. Чтобы наложить полученную шкалу на спектр сигнала, необходимо использовать пропорцию (формула (4)).

$$f(j) = \left\lceil \frac{F_d}{N} j \right\rceil, \quad (4)$$

где F_d – частота дискретизации, $j = 0, \dots, N/2$.

Тогда фильтры будут иметь следующий вид (формула (5)):

$$H_m(k) = \begin{cases} 0 & k < f(m-1), \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k < f(m), \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \leq k \leq f(m+1), \\ 0 & k > f(m+1), \end{cases} \quad (5)$$

где $m = 0, \dots, N_{FB}$ – порядковый номер фильтра.

Фильтры применяются к энергии спектра. После применения фильтров полученные значения логарифмируются, это позволяет понизить чувствительность коэффициентов к шумам. Таким образом, промежуточные результаты будут вычисляться по формуле (6):

$$S(m) = \log \left(\sum_{k=0}^{N-1} |X_k|^2 * H_m(k) \right). \quad (6)$$

Заключительным этапом в вычислении MFCC-коэффициентов является дискретное косинусное преобразование (формула (7)).

$$c(i) = \sum_{m=0}^{N_{FB}-1} S(m) \cos \left(\frac{\pi i \left(m + \frac{1}{2} \right)}{N_{FB}} \right), \quad (7)$$

где $i = 0, \dots, N_{FB}$ – порядковый номер MFCC-коэффициента.

Коэффициент $s(0)$ обычно не используется, так как представляет собой энергию сигнала.

Создание карты пользователя состоит из выработки следующего набора данных.

- Уникальный идентификатор пользователя.
- Длина, частота дискретизации и глубина звучания очищенной записи. В случае использования одинакового оборудования с общими параметрами записи образца во время регистрации пользователя и процедуры аутентификации частота дискретизации и глубина звучания будут общими для всей системы.
- Набор мел-частотных кепстральных коэффициентов для фрейма, равного по длине всей очищенной записи. Такое решение позволяет сравнить акустические параметры пароля, это, во-первых, не требует применения алгоритмов распознавания речи для сравнения паролей, а во-вторых, дает нам возможность судить об общем совпадении манеры произношения пароля.
- Массив наборов мел-частотных кепстральных коэффициентов для фреймов заданной длины с половинным перекрытием.

Для проведения процедуры аутентификации пользователь вводит свой идентификатор и произносит парольную фразу. По идентификатору из базы данных извлекается карта пользователя.

Для полученной записи от начала последовательно со сдвигом на одно значение вправо производится расчет мел-частотных кепстральных коэффициентов для фреймов. После расчета каждого набора коэффициентов производится сравнение расстояния между полученными значениями и значениями первого набора коэффициентов из карты пользователя до нахождения его первого минимума (последующие минимумы в общем случае могут обозначать начало второго слова). Аналогичным образом находится конец записи.

Далее происходит обрезка дорожки по полученным значениям, расчет набора MFCC для всей записи и его сравнение с набором из карты пользователя.

Проводится сравнение длины полученной после обрезки записи со значением из карты пользователя, исходя из отношения длин выбирается степень перекрытия фреймов для расчета массива наборов MFCC (при разнице длин, равной половине значения из карты пользователя, рекомендуется запросить повторный ввод образца голоса, данное событие может сигнализировать о нахождении пользователя в состоянии стресса, наркотического или алкогольного опьянения и о прочих отклонениях от адекватного поведения).

Набор MFCC можно рассматривать как координаты точки в пространстве, тогда для сравнения двух фреймов достаточно

рассчитать расстояние между двумя наборами коэффициентов. Для этого воспользуемся формулой (8).

$$d = \sqrt{\sum_{n=1}^{N_{FB}} (mfcc_1(n) - mfcc_2(n))^2}, \quad (8)$$

где $mfcc_1(n)$ – коэффициент с порядковым номером n , принадлежащий набору i .

Следующим этапом сравнения полученной записи с картой пользователя является последовательный расчет MFCC и нахождение ближайших наборов для всех наборов из карты, сравнение относительного положения наиболее близких пар.

Если расстояние между двумя наборами MFCC меньше заданного коэффициента (учет погрешности), фреймы можно считать равными. Сравнение относительного положения происходит с погрешностью на половину среднего арифметического от значений перекрытия фреймов (половинное перекрытие при расчете наборов коэффициентов для карты).

Вывод о подлинности производится на основании сравнения отношения количества совпавших по относительному положению фреймов к их общему количеству с выбранным пороговым значением.

Таким образом, алгоритм работы можно разбить на две части: создание карты пользователя и процедура аутентификации.

Создание карты пользователя:

- запись парольной фразы;
- очистка записи от отрезков, не несущих полезной информации;
- расчет и сохранение общих частотных характеристик парольной фразы на основе выработки общего набора MFCC;
- расчет и сохранение наборов MFCC для фреймов;
- присваивание уникального идентификатора выработанной карте.

Процедура аутентификации:

- считывание идентификатора пользователя;
- запись образца парольной фразы;
- открытие карты пользователя;
- нахождение начала и конца парольной фразы на основании данных из карты. В случае неудачи при проведении процедуры производится отказ от предоставления доступа;
- производится очистка записи с целью выборки отрезка, предположительно содержащего парольную фразу;
- расчет общих частотных характеристик проверяемого отрезка образца и сравнение с данными из карты пользователя. При пре-

- вышении контрольного значения производится отказ от предоставления доступа;
- поиск схожих фреймов и расчет их относительного положения;
 - расчет процента найденных, схожих по значению и относительному положению, фреймов. Сравнение с пороговым значением предоставления прав доступа. При недостижении порогового значения производится отказ от предоставления доступа;
 - предоставление прав доступа пользователю.

В основу разработанного прототипа программной реализации были положены следующие параметры:

- 1) исследуемый диапазон принят за 32–16 000 Гц;
- 2) все записи производятся с частотой дискретизации 32 000 Гц и глубиной звучания 32 бита. Такая частота дискретизации позволяет по теореме Котельникова восстановить без искажений основной диапазон частот речи для анализируемой записи. Глубина звука выбирается исходя из условий задачи, для проверки модели оптимально использовать максимально доступное значение. Такой шаг увеличит точность расчета наборов MFCC;
- 3) длина фрейма выбирается меньшей или равной половине значения $sl = \frac{60}{t * s}$, где sl – средняя длительность звука, t – средний темп речи (слов в минуту), s – среднее количество звуков в слове. Для русского языка соответственно: $sl = 70$ мс. Таким образом, длина фрейма при данных условиях не должна превышать 1129 семплов, выберем ее равной 1024 семпла;
- 4) в расчетах MFCC используется 24 фильтра.

Для проверки корректности предложенного метода были проведены следующие эксперименты с разной длиной записи (525–812 мс):

- 1) проведение процедуры аутентификации пользователя;
- 2) симуляция атаки на процедуру аутентификации (нарушитель знает пароль пользователя);
- 3) симуляция ошибки пользователя в парольной фразе;
- 4) симуляция атаки на процедуру аутентификации (нарушитель не знает пароль пользователя).

С целью оптимизации сравнения записей в случаях, когда обнаруживалось, что процент совпадения записей меньше пятидесяти одного, дальнейшее сравнение и расчет точного значения процента совпадения не проводились.

Во время экспериментов было обнаружено, что суждение об общей правильности произнесенной парольной фразы на основе вычисления расстояния между наборами MFCC для двух полных записей не обладает достаточной точностью. По этой причине от данной проверки пришлось отказаться.

Также с помощью экспериментов 3 и 4 было выяснено, что при выполнении условий достаточности различия действительного и произнесенного экземпляра пароля, а также корректного выбора порогового значения процента совпадения записей, ошибок при выполнении процедуры аутентификации выявлено не было. Стоит заметить, что необходимая степень различия для данного исхода не выявлялась.

При проведении экспериментов с образцами женских голосов был определен чрезмерный рост вероятности ошибки первого рода, природа данного явления содержится в особенности конкретных мел-частотных фильтров, определяющих большую точность расчета MFCC в низких частотах.

В таблице представлены вероятности ошибок первого и второго рода при различных выборках участников и пороговых значениях процента совпадения записей.

Таблица

Вероятности ошибок первого и второго рода (в %)

Выборка	Пороговое значение	Вероятность ошибки I рода	Вероятность ошибки II рода
Все	55	36.(66)	2.(27)
	51	30	11.(36)
Без женских голосов	55	16.(66)	0
	51	16.(66)	8.(33)

На точность процедуры аутентификации пользователя по голосу с помощью приведенной выше модели системы аутентификации в большей степени влияют следующие факторы:

- длина парольной фразы;
- частотный состав звуков, составляющих парольную фразу.

Как показали эксперименты, точность распознавания диктора зависит, прежде всего, от длины парольной фразы и работы артикуляционных органов во время произношения парольной фразы. Для уменьшения влияния данных факторов при выработке карты пользователя следует использовать методы нейронных сетей и

машинного обучения. Также в карты пользователя следует внести диапазон анализируемых частот с учетом ограничений, рассчитанных на основании пола пользователя.

Для учета различий между программно-аппаратными средствами при записи проверочной информации и проведении процедуры аутентификации авторами рекомендуется применение методов «компенсации канала».

Примечания

- ¹ *Островский А.А., Жариков Д.Н., Лукьянов В.С., Попов Д.С.* Динамические методы биометрической аутентификации // Известия Волгоградского государственного университета. 2010. Т. 6.
- ² *Рамишвили Г.С.* Автоматическое опознавание говорящего по голосу. М.: Радио и связь, 1981.
- ³ *Кедрова Г., Анисимов Н., Захаров Л.* Сопоставительное МРТ-исследование артикуляционных моделей гласных звуков в разных языках // Вестник Московского государственного лингвистического университета. Серия «Языкознание». 2012. Т. 13.
- ⁴ *Сорокин В.Н., Вьюгин В.В., Тананькин А.А.* Распознавание личности по голосу: Аналитический обзор // Информационные процессы. 2012. Т. 12.