Сравнение методов голосовой идентификации человека

Э. Х. Шамсиев, email: eldar_shamsiev@mail.ru
В. В. Мокшин, email: vladimir_kgtu@mail.ru

Альметьевский филиал КНИТУ-КАИ им. А.Н. Туполева

Аннотация. *В статье рассматриваются методы голосовой идентификации человека. Для начала описывается поэтапное решение проблем такие, как запись речи диктора; фильтрация сигнала; выявление возможных затруднений при обработке звука, а также предложены пути решения таких проблем; выявление и классификация признаков. В конце работы представлена таблица, показывающая результаты лучших методов идентификации человека, что будет в дальнейшем в полной мере отвечать на вопрос об усовершенствовании методов.*

Ключевые слова: *звук, шум, ряд, речь, преобразование, сигнал, методы классификации, голосовые признаки.*

Введение

В настоящее время голосовая идентификация имеет широкое применение, поскольку голос человека, как бы на слух ни казался похож на другие, имеет уникальные признаки, что делает его биометрическими данными. Они, в свою очередь, используются для доступа к каким-либо защищённым информационным системам. Разумеется, бывают исключительные случаи, когда не удаётся распознать человека по его же голосу или даже бывает такое, что тембр голоса одного человека может совпасть с голосом, который занесён в базу данных, другого человека и, таким образом, предоставить доступ к какой-либо информации. Вследствие этого требуется сравнить известные методы, чтобы дать основу к улучшению голосовой идентификации. Весь процесс идентификации речи основывается на классической структуре: запись сигнала, его обработка, извлечение признаков, распознавание речи.

1. Предварительная запись голоса

Обычно голос диктора записывают с помощью микрофона и в дальнейшем оцифровывают благодаря аналого-цифровым преобразователям. Также всем известно, что такой сигнал не будет идеально чистым, так как он будет содержать множество нежелательных

сигналов таких, как помех, вызванных различными факторами, которые будут перечислены в следующей главе.

2. Обработка (фильтрация) сигнала.

Необходимо обработать звук, поскольку существуют множество причин, которые могут повлиять на идентификацию речи. К примеру:

- Качество оборудования (микрофона).
- Нечёткая речь диктора (болезнь, дефект речевого аппарата и пр.).
- Незаконченное предложение диктором.
- Внешнее прямое воздействие на микрофон (сильные потоки воздуха).
- Окружающая среда (к примеру, эхо).

Необходимость очистки сигнала от шумов, лишних частот связана с тем, чтобы задать однозначность распознавания голоса диктора. Существуют множество методов фильтрации записи, но в данной работе используется некая совокупность алгоритмов, в основе которых лежит дискретное и быстрое преобразования Фурье. И на нём следует заострить внимание, так как обработка сигнала с помощью преобразования Фурье играет ключевую роль при идентификации человека, поскольку распознавание личности не будет иметь никакого смысла, если запись будет иметь много лишних звуков.

Преобразование Фурье позволяет представить непрерывную функцию $f(x)$ – он же сигнал ,- определённую на отрезке $[0,T]$ в виде суммы бесконечного числа тригонометрических функций с определёнными амплитудами и фазами, рассматриваемыми на отрезке. Такой ряд и называется рядом Фурье. Обобщая, можно сказать, что преобразование Фурье способствует разложению функции по частотам, что позволит из суммы нескольких звуковых волн найти те, которые являются «плохими» волнами (к примеру, фоновый шум), и снизить их уровень шума.

Любой сигнал можно представить математически. Так, сигнал длительностью T секунд представляется некоторой математической функцией $f(x)$, заданной на отрезке $\{0,T\}$, где x – время. Тогда такую функцию можно представить в виде суммы гармонических функций ($\sin(x)$ или $\cos(x)$) вида:

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{+\infty} A_k \cos\left(2\pi \frac{k}{\tau} x + \theta_k\right), \quad (1)$$

где:

– a_0 – коэффициент Фурье функции;

– $a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)dx$;

– k – номер тригонометрической функции;

– A_k – амплитуда k -го гармонического колебания;

– T – отрезок, где функция определена (длительность сигнала);

– θ_k – начальная фаза k-го колебания;

– $k\frac{2\pi}{\tau} = k\omega$ – круговая частота гармонического колебания.

Функцию представляют в виде суммы ряда для того, чтобы, сложив в каждой точке значение гармонических составляющих (A_k и θ_k), мы получили значение нашей функции в этой точке.

Так же ряд можно представить в двух видах:

$$f(x) = \sum_{k=-\infty}^{+\infty} f_k e^{i2\pi\frac{k}{\tau}x}, \qquad (2)$$

где A_k , k -я комплексная амплитуда.

Или же:

$$f(x) = \frac{a_0}{2} + \sum_{k=1} [a_k \cos(2\pi\frac{k}{\tau}x) + b_k \cos(2\pi\frac{k}{\tau}x)] \qquad (3)$$

Связь между коэффициентами в (1) и (3) выражается следующими формулами:

$$A_k = \sqrt{a_k^{\,2} + b_k^{\,2}} \qquad (4)$$

$$\theta_k = \arctan(\frac{b_k}{a_k}) \qquad (5)$$

На примере функции $f(x) = \sin(x)$ стоит добавить, что периоды гармонических составляющих кратны величине отрезка $\{0, T\}$, на котором определена исходная функция f(x). Другими словами, периоды гармоник кратны длительности измерения сигнала. Например, период первой гармоники ряда Фурье равен интервалу T, на котором

определена функция $f(x)$. Период второй гармоники ряда Фурье равен интервалу $\frac{T}{2}$. И так далее.

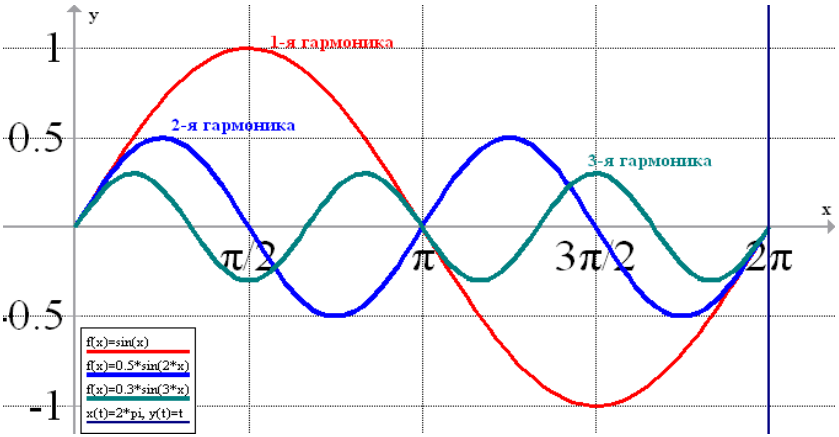

Рис. 1. График гармоник

Соответственно, частоты гармонических составляющих кратны величине $\frac{1}{T}$. То есть частоты гармонических составляющих F_k равны $F_k = \frac{k}{T}$, где k пробегает значения от 0 до ∞, например, $k = 0, F_0 = 0; k = 1, F_1 = \frac{1}{T}; F_k = \frac{k}{T}$.

Пусть наша исходная функция представляет собой сигнал, записанный в течение $T = 1$ сек. Тогда период первой гармоники (рис. 1) будет равен длительности нашего сигнала $T_1 = T = 1$ сек и частота гармоники равна 1 Гц. Период второй гармоники будет равен длительности сигнала, деленной на 2 ($T_2 = \frac{T}{2} = 0,5$ сек), и частота равна 2 Гц. Для третьей гармоники $T_3 = \frac{T}{3} = 0,33$ сек, и частота равна 3 Гц. И так далее. Тогда шаг между гармониками равен 1 Гц.

Помимо этого, следует помнить о теореме Котельникова (частота дискретизации должна как минимум вдвое превышать максимальную частоту сигнала), иначе при фильтрации мы получим ложные результаты.

Согласно теореме Котельникова, если непрерывный сигнал имеет спектр, ограниченный частотой $F_{макс}$, то он может быть полностью и однозначно восстановлен по его дискретным отсчетам, взятым через интервалы времени.

$$T = \frac{1}{2 F_{макс}} \tag{6}$$

То есть с частотой $F_d \geq 2 F_{макс}$, где F_d — частота дискретизации; $F_{макс}$ — максимальная частота спектра сигнала. Другими слова, частота оцифровки сигнала должна как минимум в 2 раза превышать максимальную частоту сигнала, который мы хотим измерить. Иначе же мы будем сталкиваться с эффектом «алиасинга», при котором сигнал высокой частоты после оцифровки превращается в несуществующий сигнал низкой частоты.

На рис. 2 красным цветом изображена синусоида с высокой частотой – это реальный сигнал, а синим цветом изображена синусоида с низкой частотой – это фиктивный сигнал, который возникает из-за того, что за время взятия отсчёта успевает пройти больше чем полпериода высокочастотного сигнала. Вследствие этого частоту дискретизации нужно устанавливать вдвое выше максимальной частоты сигнала [1, 2].

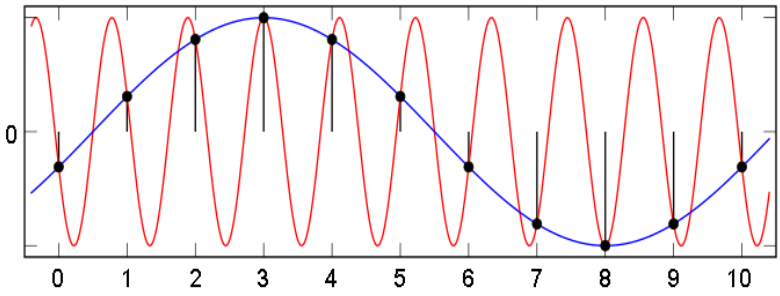

Рис. 2. Появление ложного сигнала при недостаточно высокой частоте дискретизации

После того, как мы на примере простейших функций рассмотрели преобразование Фурье, переходим непосредственно к решению реальных голосовых сигналов.

– запись разбивается на блоки 25.6 мс со смещением каждые 10 мс. Получаются блоки по 409 отсчётов каждый;
– предварительное усиление (фильтр 1-го порядка);
– обработка окном Хемминга;
– дополнение нулями и БПФ длиной 512;
– усреднения амплитуд спектра в пределах полос с треугольными весовыми функциями;
– вычисление максимальной нормы всех векторов сигнала;
– нормализация всех векторов сигнала;

Данные алгоритмы в дальнейшем будут интерпретированы и перенесены в среду разработки MatLab R2021a.

На рис. 3 представлен результат выполнения программы на MatLab [3].

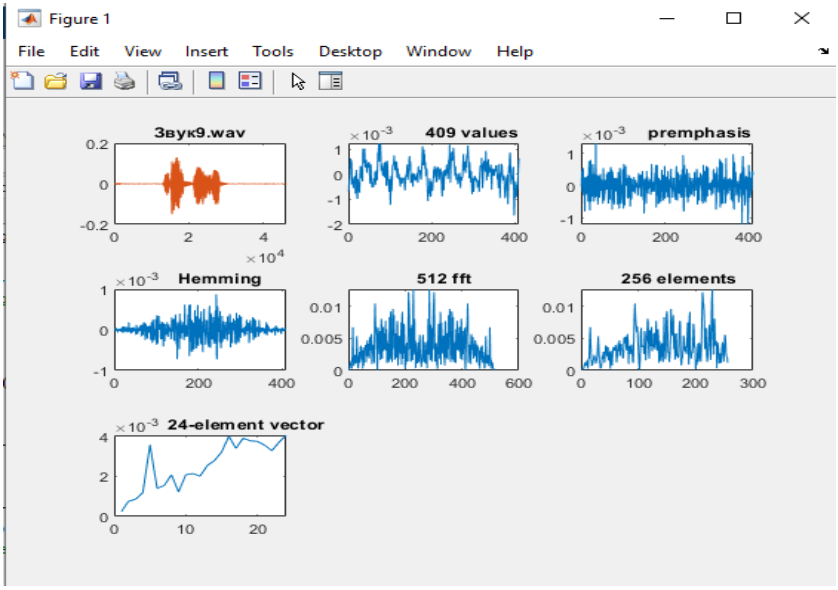

Рис. 3. Результат фильтрации записи

3. Извлечение признаков

Необходимо извлечь те признаки, по которым можно будет определить голос того или иного диктора. В данной статье рассматривается наиболее популярный признак: MFCC-коэффициенты.

MFCC-коэффициенты. Такой метод считается наиболее популярным среди известных методов, поскольку обеспечивает максимальную точность при распознавании личности диктора. Изначально подаётся последовательность отсчётов участка сигнала.

$$x[n], \quad 0 \le n < N \tag{7}$$

Далее эта последовательность представляется в виде ряда Фурье, и применяется соответствующее дискретное преобразование.

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-2\pi i}{N} kn}, \quad 0 \le k < N \tag{8}$$

Помимо этого, используется так называемая весовая функция, которая предназначена для уменьшения искажений во время преобразований Фурье. Как правило, весовой функцией выступает окно Хэмминга. Далее этот сигнал разбивают на диапазоны с помощью банка треугольных фильтров.

$$H_m = \begin{cases} 0, & k < f[m-1] \\ \dfrac{(k - f[m-1])}{(f[m] - f[m-1])}, & f[m-1] \le k < f[m] \\ \dfrac{(f[m+1] - k)}{(f[m+1] - f[m])}, & f[m] \le k \le f[m+1] \\ 0, & k > f[m+1] \end{cases} \tag{9}$$

Границы этих фильтров представляются в шкале мэл, которая описывает восприятие звуков на разных частотах человеческим ухом. Выполняем перевод в мэл-частотную область согласно формуле

$$B(f) = 1127 \ln\left(1 + \frac{f}{700}\right) \tag{10}$$

Обратное преобразование

$$B^{-1}(b) = 700\left(e^{\frac{b}{1127}} - 1\right) \tag{11}$$

После того как выбран диапазон частот, его переводят в шкалу мэл и разбивают на несколько равномерных перекрывающихся диапазон. Также необходимо вычислить энергию для каждого окна

$$S[m] = \ln\left(\sum_{k=0}^{N-1} \left|X_a[k]\right|^2 H_m[k]\right), \quad 0 \le m < N \tag{12}$$

Остаётся применить дискретное косинусное преобразование (ДКП)

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos\left(\frac{\pi n\left(m + \frac{1}{2}\right)}{M}\right), \quad 0 \le n < M \tag{13}$$

Как примечание, M выбирают до 12-го порядка [4].

4. Классификация признаков

После того как основным методом извлечения признаков был выбран MFCC-коэффициенты, приступаем к классификациям признаков, которые будут иллюстрировать числовые показатели качества голосовой идентификации.

Метод опорных векторов. Данный метод используется преимущественно в задачах бинарной классификации, заключающийся в поиске разделяющей гиперплоскости между двумя классами. Кроме того, уникальность такого метода заключается в том, что при линейной разделимости среди других выделяется такая гиперплоскость, которая максимально отдалена от обоих классов. Но если линейная разделимость отсутствует, то используют вспомогательные переменные, которые характеризуют допустимые ошибки классификации. Также существует способ, описанный многими учёными, который позволяет перейти в пространство большей размерности. Это приведёт к такой выборке, которая может быть линейно разделена. Способ заключается в том, что выборка переходит от скалярного произведения к нелинейной функции ядра. Однако основная мысль этого метода – чем больше расстояние между разделяющей гиперплоскостью и объектами разделяемых классов, тем меньше будет средняя ошибка классификатора [5].

Модель наивного байесовского классификатора. Данная модель – это простой вероятностный классификатор, основанный на применении теоремы Байеса со строгими предположения о независимости [6].

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}, \tag{14}$$

где $P(A \mid B)$ – вероятность события A при условии, что событие B выполнилось, $P(B \mid A)$ – наоборот, вероятность события B, при условии, что событие A истинно. $P(A)$, $P(B)$ – вероятность события A, B соответственно.

Модель логистической регрессии – статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путём его сравнения с логистической кривой. Результат представляется как бинарное событие (1 или 0). Логистическая регрессия применяется для прогнозирования некоторого события по значениям множества признаков. Для этого вводится зависимая переменная y, принимают лишь одно из двух значений: 0 (ложь / событие не произошло) и 1 (истина / событие произошло) и множество независимых переменных – вещественных $x_1, x_2, x_3, \cdots, x_n$ на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной. Если предположить, что событие $y = 1$, то

$$\Pr\{y = 1 \mid x\} = f(z), \tag{15}$$

где $x = \theta^T x = \theta_1 x_1 + \cdots + \theta_n x_n$, x и θ – вектора значений независимых переменных $x_1, x_2, x_3, \cdots, x_n$ и параметров – вещественных числе $\theta_1, \theta_2, \theta_3, \cdots, \theta_n$ соответственно, а функция $f(z)$ – логистическая функция – она же сигмоида [7].

После того как были отобраны основные сравниваемые классификации признаков, необходимо их сопоставить на практике. Для анализа используются 20 разных голосов людей [8].

Таблица

Используемые для анализа признаки и методы

Признак	**Метод опорных векторов**	**Модель наивного байесовского классификатора**	**Модель логистической регрессии**
MFCC-коэффициенты	0,873	0,33	0,888

По таблице видно, что наибольший показатель предоставляет модель логистической регрессии, которая даёт точность в 88%, а если округлять, то и вовсе 89%.

Заключение

Таким образом, в данной статье мы разобрали, по какому принципу устроена голосовая идентификация человека; как осуществить запись голоса; её предварительную обработку (фильтрацию) и с помощью каких фундаментальных знаний это сделать; самые популярные методы извлечения признаков; методы классификации; сравнили эти методы и пришли к заключению, какие из них требуют совершенствования.

Список литературы

1. Практическое применение преобразования Фурье для анализа сигналов. Введение для начинающих / Андрей @diakin. [Электронный ресурс] – статья на сайте Habr. – Режим доступа: https://habr.com/ru/post/269991/
2. Шамсиев Э.Х. Повышение качества голосовой записи в летательных аппаратах на примере языка программирования Python 3 / Э. Х. Шамсиев, В. В. Мокшин // Всероссийская научно-практическая конференция «Современные технологии в кораблестроительном и авиационном образовании, науке и производстве», посвященная памяти Р.Е. Алексеева (Нижний Новгород, 16-17 декабря 2021 г.). – Нижний Новгород, 2021. – С. 10.
3. Предварительная обработка речевых сигналов с помощью Matlab / Роман Іваськевич @REssential. [Электронный ресурс] – статья на сайте Habr. – Режим доступа: https://habr.com/ru/post/159605/
4. Мел-кепстральные коэффициенты (MFCC) и распознавание речи. [Электронный ресурс] – статья на сайте Habr. – Режим доступа: https://habr.com/ru/post/140828/
5. Метод опорных векторов (SVM). [Электронный ресурс] – статья в Wikipedia – Режим доступа: https://ru.wikipedia.org/wiki/Метод_опорных_векторов
6. Наивный байесовский классификатор. [Электронный ресурс] – статья в Wikipedia – Режим доступа: https://ru.wikipedia.org/wiki/Наивный_байесовский_классификатор
7. Логистическая регрессия. [Электронный ресурс] – статья в Wikipedia – Режим доступа: https://ru.wikipedia.org/wiki/Логистическая_регрессия
8. С.Г. Лялин. Выбор метода извлечения признаков для голосовой идентификации. Научный журнал: Научно-технический вестник Поволжья №6 2020, 2020. – С. 88-93.
9. Тассов К.Л., Дятлов Р.А. Метод идентификации человека по голосу. Инженерный журнал: наука и инновации, 2013, вып. 6. URL: http://engjournal.ru/catalog/it/biometric/1103.html