УДК 004.522

Разработка метода голосового доступа к информационным системам с применением нейросетевого инструмента анализа

Т.А. Газизов, А. Н. Назмутдинова

Московский физико-технический институт (государственный университет)

Речевые характеристики каждого индивидуума вариативны, и для точного определения этих характеристик требуется проведение неоднократной записи голоса. В данной работе исследуется возможность распознавания с большой вероятностью личности того, чей голос записан на аудиозапись, с помощью цифровой обработки голоса в аудиозаписи и использования искусственной нейронной сети как инструмент анализа. Актуальность данного метода заключается в том, что технология распознавания голоса – практичное решение для доступа ко многим устройствам.

Проект представляет собой многомодулевую программу, способную записывать аудиофайл и принимать решение об успешности идентификации говорящего, имеющего доступ, или об отказе в доступе другим лицам. Идентификация осуществляется путем анализа обученной нейронной сетью вектора признаков записи. Исследование показало, что можно проводить идентификацию личности человека и создавать голосовые пароли, опираясь на анализ записи голоса.

В ходе работы было выделено два основных фактора, влияющих на точность идентификации личности:

1. Качественная запись и цифровая обработка аудиофайла
2. Эффективное обучение нейронной сети

Аудиозапись, используемая в исследовании, оцифровывается с применением метода импульсно-кодовой модуляции. После записи на микрофон получается сигнал длительностью 1с, который дискретизируется с частотой 16кГц, и получается 16000 отсчетов. В итоге получается 16-битный знаковый (PCM-signed) wav-файл с частотой дискретизации 16 кГц. Затем полученные значения нормируются, чтобы свести к минимуму зависимость от интенсивности звука и частоты звуковых колебаний говорящего, и полученный дискретный сигнал разбивается по кадрам для получения менее объемных данных (рис.1). Далее производится «порезка» звукового сигнала на кадры длины 128мс с половинным перекрытием. Необходимость в перекрытии вызвана искажением звука в случае, если бы кадры были расположены рядом.

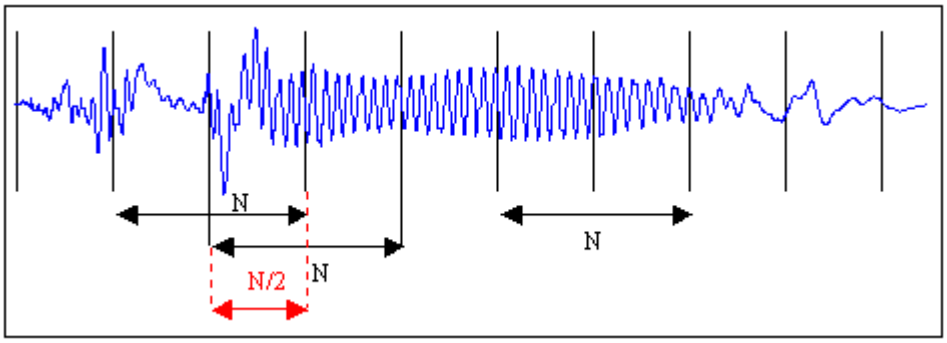


Рис.1

Затем для сглаживания краевых эффектов для каждого кадра применяется окно Хэмминга. Каждый из кадров домножается на весовую функцию (Хэмминга).

$$\omega(n) = 0.53836 - 0.46164\cos(\frac{2\pi n}{N-1})$$

Где n - порядковый номер элемента в кадре, для которого вычисляется новое значение амплитуды, N - длина кадра (количество значений сигнала, измеренных за период).

Следующим шагом является получение кратковременной спектрограммы каждого кадра в отдельности. Для этого применяется дискретное преобразование Фурье (рис.2).

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

N — как и ранее, длина кадра (количество значений сигнала, измеренных за период), x_n — амплитуда n-го сигнала, X_k — N комплексных амплитуд синусоидальных сигналов, слагающих исходный сигнал.

Кроме этого, возведем каждое значение X_k в квадрат для дальнейшего логарифмирования. На приложенных картинках оцифрованный сигнал и его Фурье-спектр. Первый скриншот для имеющего доступ человека, второй – для не имеющего.

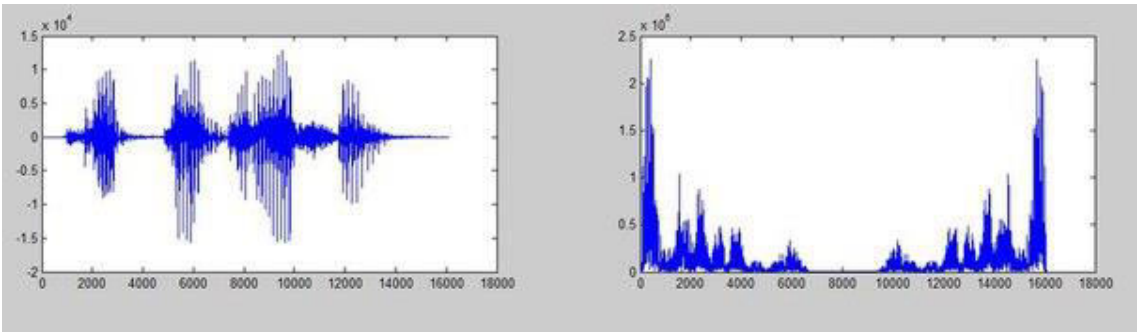


Рис. 2

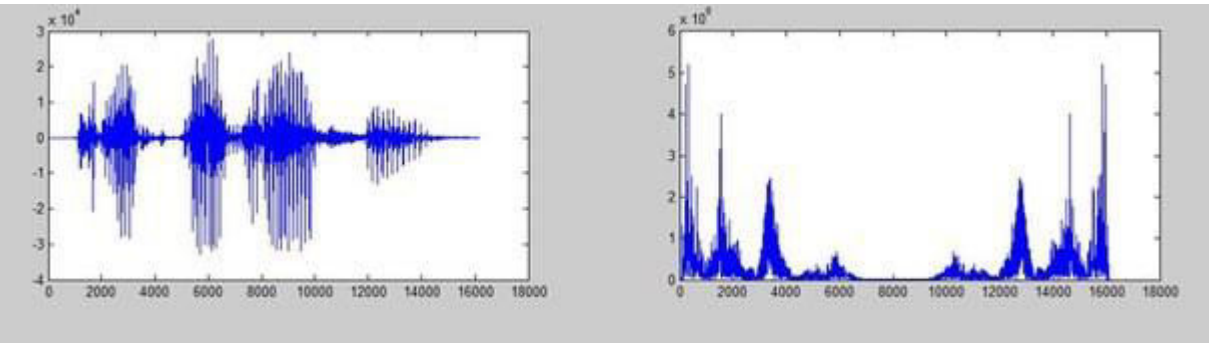


Рис. 3

Чтобы произвести более глубокий анализ той части спектра, которая играет наибольшую роль при распознавании, используется мел-шкала.

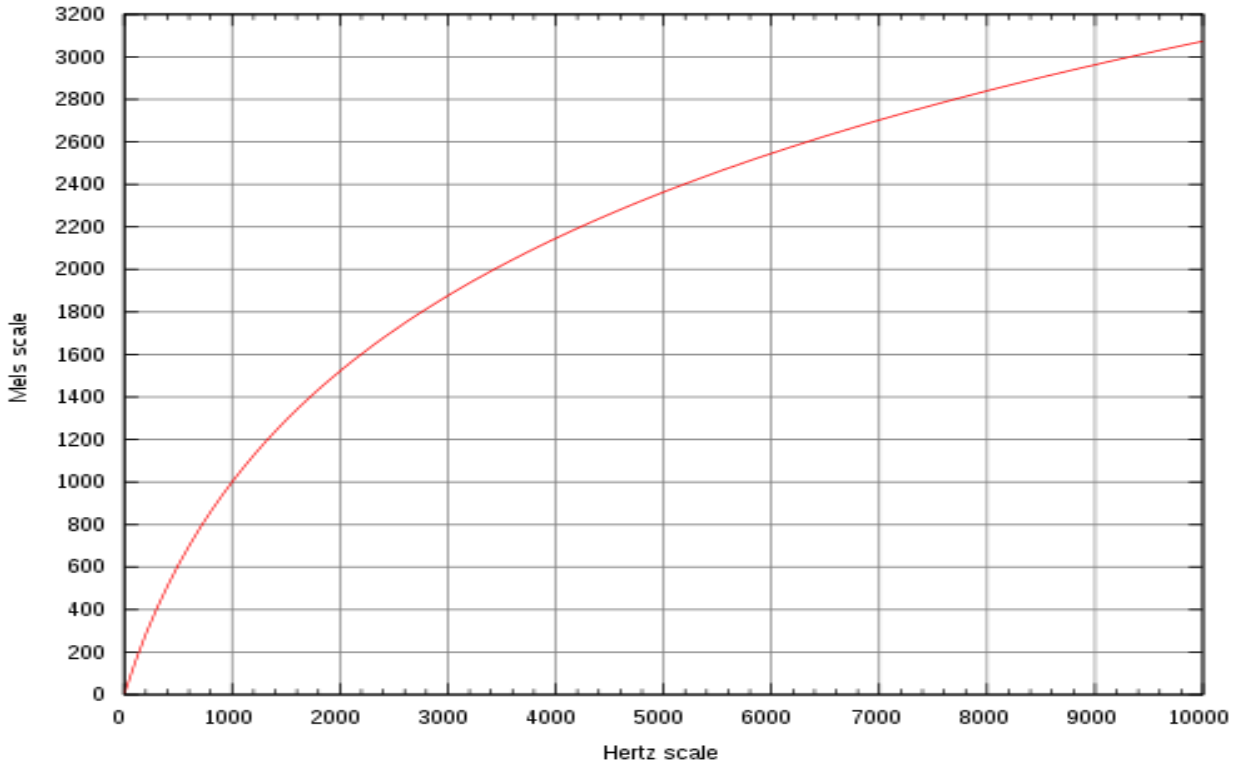


Рис.4

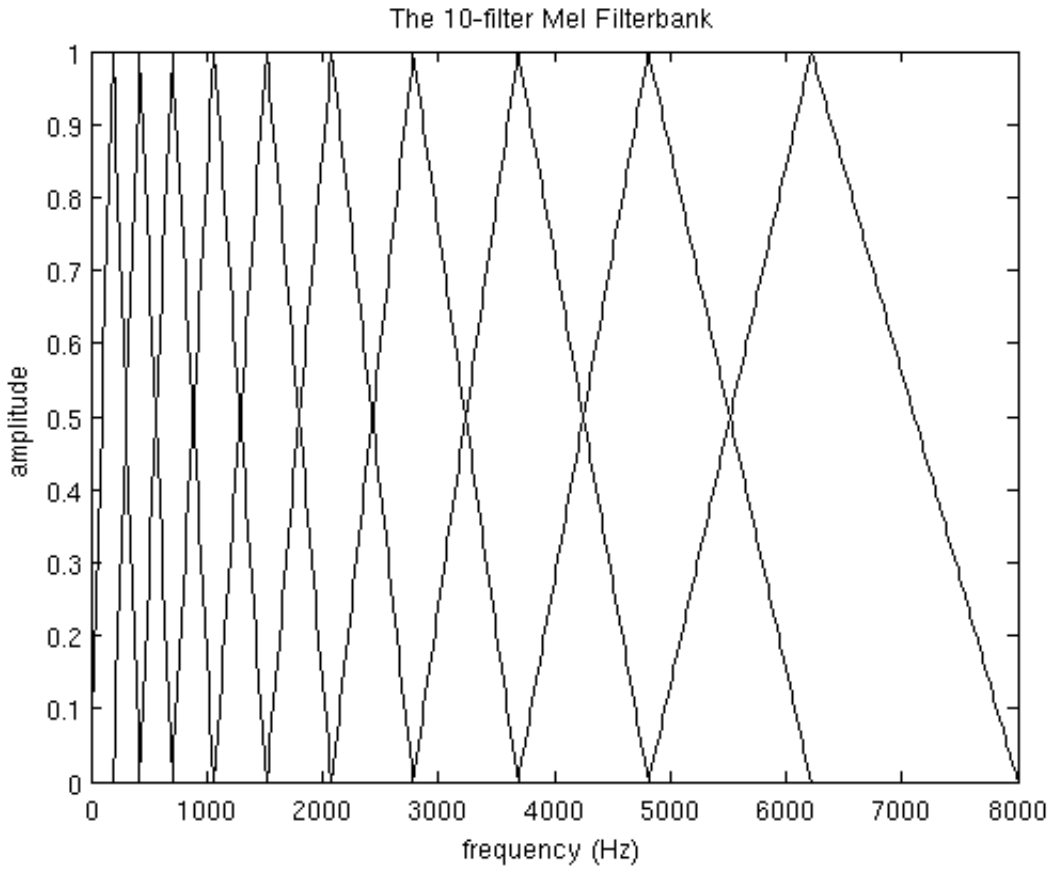
Мел-шкала ведет себя линейно до 1000 Гц, а после проявляет логарифмическую природу. Переход к новой шкале описывается зависимостью:

$$m = 2595 \log\left(1 + \frac{f}{700}\right) = 1127 \ln(1 + \frac{f}{700})$$

m — частота в мелах, f — частота в герцах.

Мел — это «психофизическая единица высоты звука», основанная на субъективном восприятии среднестатистическими людьми. Зависит в первую очередь от частоты звука. Для исследования представляет интерес диапазон от 300 до 8000 Гц (в мел-шкале диапазон превращается в [401.25; 2834.99]), поэтому если разложить полученный спектр по мел-

функциям, получится более глубокий анализ той части спектра, которая играет наибольшую роль при распознавании. Мел-фильтр представляет собой треугольную оконную функцию, которая позволяет просуммировать количество энергии на какой-то конкретной частоте и тем самым получить мел-коэффициент.



Далее, для того, чтобы построить 10 треугольных фильтров, требуется 12 опорных точек:

m[i] = [401.25, 622.50, 843.75, 1065.00, 1286.25, 1507.50, 1728.74, 1949.99, 2171.24, 2392.49, 2613.74, 2834.99]

В частотном диапазоне этот набор будет выглядеть следующим образом:

h[i] = [300, 517.33, 781.90, 1103.97, 1496.04, 1973.32, 2554.33, 3261.62, 4122.63, 5170.76, 6446.70, 8000]

Таким образом выровнялся рост значимости на низких и высоких частотах. Наложив полученную шкалу на спектр каждого из выделенных «кусков» записи, можно получить опорные точки по оси частот. Зная опорные точки на оси X, легко построить необходимые нам фильтры по следующей формуле:

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

Далее «склеиваются» для каждой записи все ее «мел коэффициенты» и результат называется «вектором признаков». В итоге мы получаем 12*9 составляющих вектора, которые сравниваются с «эталонным» вектором признаков записанного имеющим доступ человеком кодового слова:

$$S[m] = \log(\sum_{k=0}^{N-1} |X[k]|^2 * H_m[k]), \quad 0 \le m < M$$

Для реализации сравнения голосовых записей и вынесения итога идентификации личности была написана программа, создающая искусственную нейронную сеть. Для обучения этой нейронной сети используются числовые векторы-признаки - результат цифровой обработки аудиозаписей различных кодовых слов, произнесенных различными людьми. Обученная нейронная сеть позволяет распознать личность и выдать положительный результат, если правильное кодовое слово было произнесено человеком, которому нейронная сеть обучена выдавать доступ, и отрицательный результат в остальных случаях. Для изменения кодового слова или личности имеющего доступ человека необходимо внести новый «правильный» вектор признаков в нейронную сеть и переобучить сеть так, чтобы предыдущий «правильный» вектор признаков воспринимался как неправильный.

Результаты экспериментов, проведённые на 200 различных записях в нейронную сеть и 70 тестах, показали успешный результат в 90% тестов.

Таким образом отработан метод, позволяющий успешно распознавать личность пользователя по его голосу.

Литература.

1. *Simon Haykin* Neural Networks. A comprehensive foundation. 2nd Edition., 1999, 823 p
2. *Станислав Осовский* Нейронные сети для обработки информации, 2004, 344 с
3. *Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman* Speaker identification using mel frequency cepstral coefficients. - 3rd International Conference on Electrical & Computer Engineering ICECE 2004
4. *Крахмалев А.К.* Использование речевой информации для биометрической идентификации в системах контроля доступа - сборник “Связь и автоматизация МВД России – 2008”