УДК 004.522

Метод идентификации человека по голосу

© К.Л. Тассов, Р.А. Дятлов

МГТУ им. Н.Э. Баумана, Москва, 105005, Россия

В статье описан метод решения проблемы идентификации человека по голосу. Рассматриваются методики анализа и системы коэффициентов звукового потока. Приведены алгоритмы предварительной обработки сигнала и выделения критериев. Описана модификация сети встречного распространения и карты Кохонена.

Ключевые слова: *голос, идентификация по голосу, устранение шумов, фильтрация речи, распознавание диктора, алгоритм фильтрации, билатеральный фильтр, алгоритм устранения тихих областей сигнала, кепстральные коэффициенты, частота основного тона, автокорреляционный метод определения частоты основного тона.*

Введение. В настоящее время актуальной является разработка систем, предназначенных для идентификации диктора. Эти системы имеют широкую область применения: криминалистика (фоноскопическая экспертиза), криптография, охранные системы и др. При их разработке важную роль играет выбор системы признаков и методов идентификации, использующих эти признаки.

Весь процесс обработки речевого сигнала можно разбить на несколько этапов:

- предобработка сигнала;
- выделение критериев;
- распознавание диктора.

Каждый этап представляет алгоритм или некоторую совокупность алгоритмов, что в итоге дает требуемый результат. На каждом этапе результат работы будет представлять собой входные параметры для следующего.

Предобработка сигнала. Необходимо понимать, что в результате оцифровки аналогового сигнала, полученного с микрофона, в сигнале будет содержаться шум, мешающий последующей обработке. Так как громкость высказывания зависит от окружающей среды и других факторов и не является постоянной величиной для двух высказываний, помимо устранения шума необходимо нормализовать амплитудную характеристику входного сигнала.

На данном этапе оцифрованные данные подвергаются фильтрации и устранению областей, не содержащих полезный сигнал. В качестве алгоритма устранения таких областей применяется авторский

метод устранения тихих областей сигнала. Для устранения высокочастотного шума применяют алгоритм билатеральной фильтрации [1].

Билатеральное фильтрование — это нелинейная техника фильтрования, которая расширяет понятие «сглаживание Гаусса», увеличивая показатели фильтра соответствующей им относительной амплитудой. Значения сигнала, которые сильно отличаются по амплитуде от центральной величины в окне, увеличиваются в меньшей степени, даже несмотря на то, что они могут находиться в непосредственной близости к центральной величине, что фактически является искривлением нелинейного фильтра Гаусса. Данный разновес основывается на значении амплитуды сигнала. В этом случае используются два фильтра Гаусса в локализованном соседстве дискретных значений сигнала: один — во временном домене (фильтр домена), другой — в домене амплитудной характеристики (ранговый фильтр).

Пусть входной сигнал $x(t)$, тогда весовые коэффициенты — $w(t)$ и выходной сигнал — $s(t)$. Для окна размером N:

$$w(k) = e^{-\frac{(x_0 - x(k))^2}{2A^2}} \cdot e^{-\frac{(t_0 - k)^2}{2T^2}}, \tag{1}$$

где A — коэффициент рангового фильтра; T — коэффициент фильтра домена; x_0 и t_0 — амплитудная и временная характеристики сигнала в центре окна.

$$s(i) = \sum_{k=0}^{N} x(k) \frac{w(k)}{wSum}, \tag{2}$$

где $wSum$ — сумма коэффициентов

$$wSum = \sum_{k=0}^{N} w(k). \tag{3}$$

На рис. 1 представлен пример сигнала до фильтрации, а на рис. 2 — после нее.

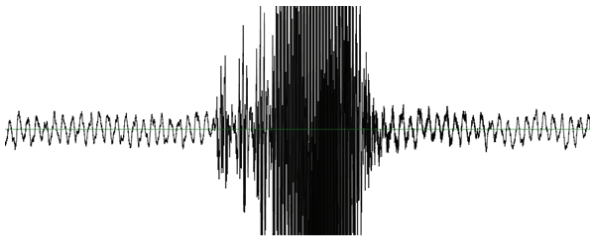

Рис. 1. Спектр амплитудно-временной характеристики сигнала

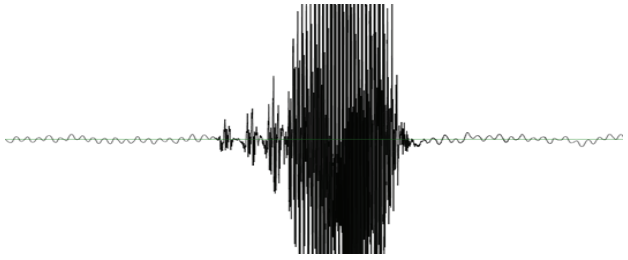

Рис. 2. Спектр амплитудно-временной характеристики сигнала после процесса фильтрации

На следующем шаге отфильтрованный сигнал подвергается устранению областей, не содержащих полезный сигнал. Для этого все значения амплитудно-временного спектра переносятся в положительную область по оси амплитуд, и на всем временном отрезке, окнами в 25 мс, происходит усреднение значений амплитуд сигнала. Спектр принимает вид, представленный на рис. 3. По данному спектру можно судить о присутствии полезного сигнала.

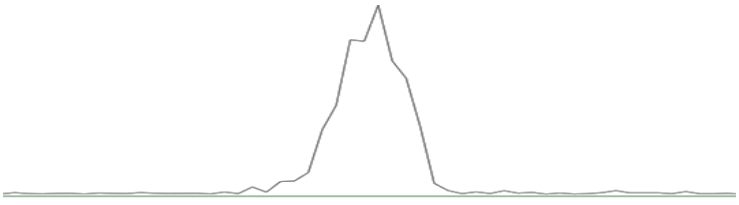

Рис. 3. Амплитудный спектр сигнала, усредненного окнами в 25 мс

Далее необходимо определить верхнюю границу величин, не содержащих полезный сигнал. Для этого все значения спектра сортируют в порядке возрастания, и применяя метод золотого сечения находят два пороговых значения кусочно-линейной функции, для которых ошибка по амплитудной оси относительно исходного спектра минимальна. На рис. 4 представлена последняя итерация работы алгоритма в графическом виде, а на рис. 5 — результат работы вышеописанного алгоритма.

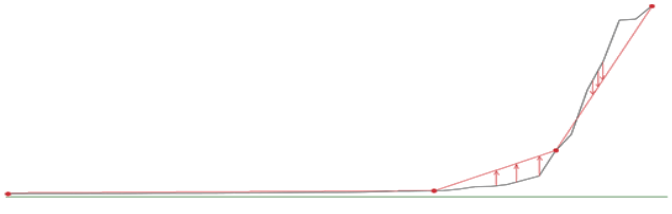

Рис. 4. График последней итерации алгоритма определения верхней границы бесполезного сигнала

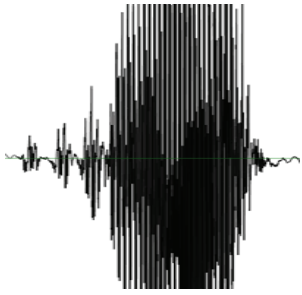

Рис. 5. Результат работы алгоритма устранения областей, не содержащих полезный сигнал

Выделение критериев. Для идентификации диктора по голосу необходимо выделить характеризующие признаки сигнала. На данном этапе обработки сигнала происходит определение частот основного тона по всей временной оси окнами по 18 мс. В качестве критериев принимаются следующие величины:

• начальная частота — значение первого отсчета;

• конечная частота — значение последнего отсчета;

• максимальная частота — максимальное значение частоты основного тона по всем отсчетам;

• минимальная частота — минимальное значение частоты основного тона по всем отсчетам;

• средняя частота — среднее значение частоты основного тона по всем отсчетам;

• время максимума — координата максимального значения в процентах от общего количества отсчетов.

Для определения частоты основного тона используется алгоритм, основанный на процессе автокорреляции сигнала.

В основе метода выделения основного тона по автокорреляционной функции лежит теорема, утверждающая, что автокорреляционная функция периодического сигнала тоже периодическая и эти два периода совпадают. Автокорреляционная функция определяется по формуле

$$R_n(k) = \sum_{m=0}^{N-k-1} x(n+m)x(n+m+k), \qquad (4)$$

где N — длина кадра анализа; n — текущая координата начала кадра анализа во всем сигнале; k — номер коэффициента функции автокорреляции.

Функция $R(k)$ достигает максимума при $k = 0$, следующий локальный максимум функция для периодического сигнала $X(n)$ с периодом P имеет место при $k = P$. Таким образом, определив положение максимума автокорреляционной функции вокализованного речевого сигнала, можно определить период основного тона. На рис. 6 представлен

спектр амплитудно-временной характеристики речевого сигнала для слова «три» и спектр частот основного тона для данного сигнала.

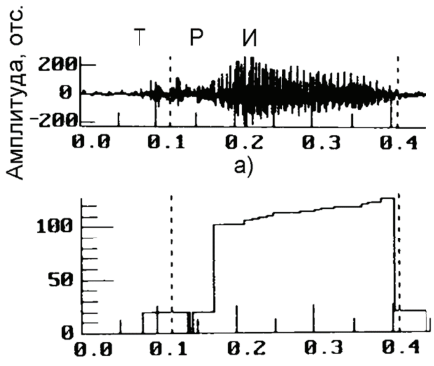

Рис. 6. Спектр частот основного тона для слова «три»

Также в качестве характеризующих признаков используются мел-частотные кепстральные коэффициенты (МЧКК), [2, 3] основанные на двух ключевых понятиях — кепстр и мел-шкала.

Кепстр — это результат дискретного косинусного преобразования от логарифма амплитудного спектра сигнала. Мел-шкала моделирует частотную чувствительность человеческого слуха. Специалистами по психоакустике было установлено, что изменение частоты в 2 раза в диапазоне низких и высоких частот человек воспринимает по-разному. В частотной полосе до 1000 Гц субъективное восприятие удвоения частоты совпадает с реальным увеличением частоты в 2 раза, поэтому до 1000 Гц мел-шкала близка к линейной. Для частот выше 1000 Гц мел-шкала является логарифмической (рис. 7).

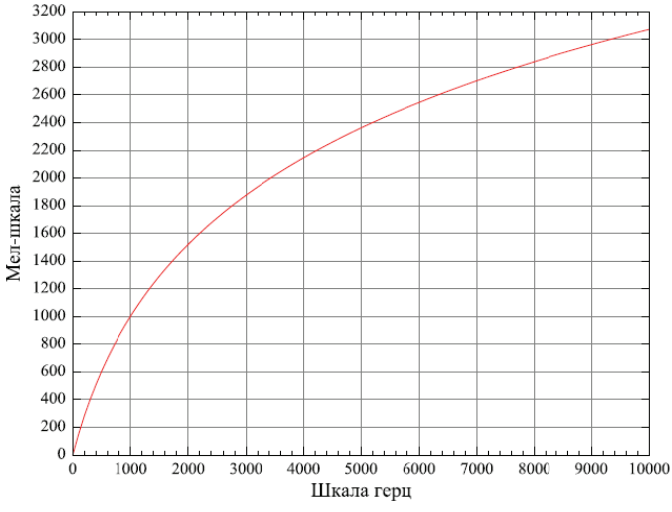

Рис. 7. Мел-шкала

Перевод из шкалы герц в шкалу мелов и обратно происходит по следующим формулам:

$$F_{mel}\left(f_{hz}\right)=1127,01048\ln\left(1+\frac{f_{hz}}{700}\right); \tag{5}$$

$$F_{hz}\left(f_{mel}\right)=700\left(e^{f_{mel}/1127,01048}-1\right). \tag{6}$$

МЧКК — это значения кепстра, распределенные по мел-шкале с использованием банка фильтров.

Существует алгоритм нахождения МЧКК.

1. Прошедший предварительную обработку сигнал $s[t]$ разбивается на K кадров по N отсчетов, пересекающихся на половину длины:

$$s\left[t\right]\to S_n\left[t\right],\ n=1,...,K\ .$$

2. В каждом кадре проводится получение комплексного представления сигнала по частотам.

3. Находится спектральная плотность мощности получившегося сигнала:

$$P_n\left[k\right]=A_n\left[k\right]^2; \tag{7}$$

$$A_n\left[k\right]=\sqrt{\mathrm{Re}\,X_n\left[k\right]^2+\mathrm{Im}\,X_n\left[k\right]^2}\ . \tag{8}$$

4. Применение банка фильтров (рис. 8):

а) задается количество фильтров, а также начальная f_l и конечная f_h частоты (f_h не должна превосходить половины частоты дискретизации);

б) далее они переводятся в мелы:

$$f_l^m=F_{mel}\left(f_l\right),$$

$$f_h^m=F_{mel}\left(f_h\right);$$

в) на мел-шкале отрезок $\left[f_l^m,f_h^m\right]$ разбивается на $P+1$ равных непересекающихся подотрезков $\left[f_j^m,f_{j+1}^m\right]$, $1\le 1\le P+1$ длины

$$len=\frac{f_h^m-f_l^m}{P+1}; \tag{9}$$

г) находятся их центры:

$$C^m\left[i\right]=f_l^m+i\cdot len,1\le i\le P; \tag{10}$$

и, переводя в шкалу Гц,

$$C[i] = F_{hz}\left(C^m[i]\right), 1 \le i \le P. \quad (11)$$

Это центральные частоты треугольных фильтров.

д) центры треугольных фильтров переводятся из герц в номера отсчетов массива $P_n[k]$:

$$f_{smp}[i] = \frac{M}{F_S} C[i], 1 \le i \le P; \quad (12)$$

где F_S — частота дискретизации исходного сигнала;

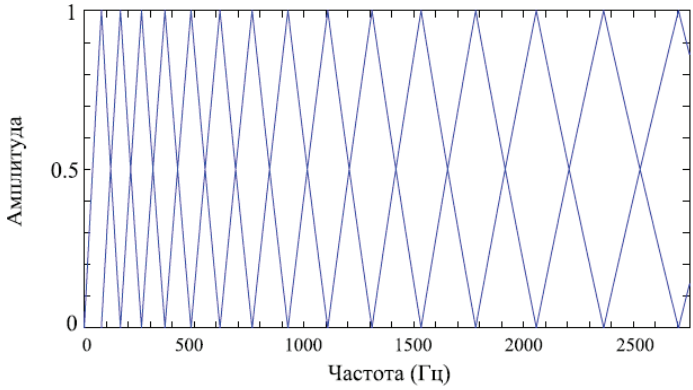

Рис. 8. Банк фильтров

е) для каждого фильтра отсчеты спектральной плотности мощности умножаются на соответствующий фильтр:

$$X_n[i] = \sum_{k=1}^{M} P_n[k] H_i[k], 1 \le i \le P. \quad (13)$$

$$H_i[k] = \begin{cases} 0, k < f_{smp}[i-1], \\ \dfrac{k - f_{smp}[i-1]}{f_{smp}[i] - f_{smp}[i-1]}, f_{smp}[i-1] \le k \le f_{smp}[i], \\ \dfrac{f_{smp}[i+1] - k}{f_{smp}[i+1] - f_{smp}[i]}, f_{smp}[i] \le k \le f_{smp}[i+1], \\ 0, k > f_{smp}[i+1]. \end{cases} \quad (14)$$

Взятие логарифма:

$$X_n[i] = ln\left(X_n[i]\right), 1 \le i \le P. \quad (15)$$

Дискретное косинусное преобразование:

$$C_n[j] = \sum_{k=1}^{P} X_n[k] \cos\left(j\left(k - \frac{1}{2}\right)\frac{\pi}{P}\right), 1 \le j \le J \ , \qquad (16)$$

где $C_n[j]$ — массив кепстральных коэффициентов; J — желаемое число коэффициентов ($J < P$).

Распознавание диктора. На данном этапе обработки данных происходит идентификация диктора по характеризующим признакам. Для этого применяется модификация сети встречного распространения Кохонена — Гроссберга [4]. Слой Кохонена в описываемой модификации представляет собой самоорганизующуюся карту Кохонена [5, 6]. На рис. 9 представлена топология данной сети.

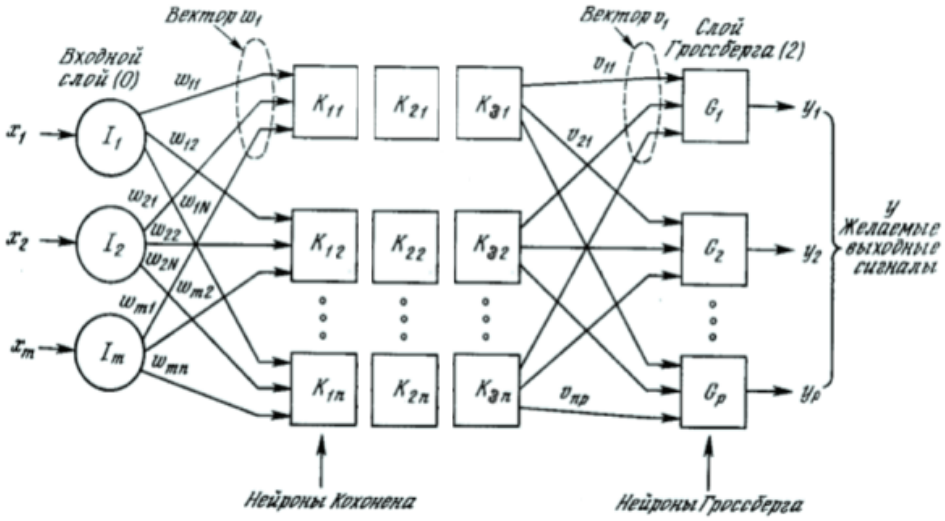

Рис. 9. Модификация сети встречного распространения

Во встречном распространении объединены два хорошо известных алгоритма: самоорганизующаяся сеть Кохонена и звезда Гроссберга.

В процессе обучения входные векторы ассоциируются с соответствующими выходными векторами. Они могут быть двоичными, состоящими из нулей и единиц, или непрерывными. Когда сеть обучена, приложение входного вектора приводит к требуемому выходному вектору. Обобщающая способность сети позволяет получать правильный выход даже при приложении входного вектора, который является неполным или слегка неверным.

Самоорганизующиеся карты Кохонена представляют собой нейронные сети, обучаемые без учителя. Они используются для классификации, организации и визуального представления больших объемов данных. Важной особенностью карт Кохонена является их способность отображать многомерные пространства признаков на плос-

кость, представив данные в виде двумерной карты, при помощи которой значительно упрощаются кластеризация и корреляционный анализ данных.

Алгоритм обучения основывается на соревновательном обучении без учителя. Он обеспечивает сохраняющее топологию отображение из пространства большой размерности в элементы карты, или нейроны, образующие двумерную решетку. Таким образом, это отображение является отображением пространства большей размерности на плоскость.

Свойство сохранения топологии означает, что карта Кохонена распределяет сходные векторы входных данных по нейронам, т.е. точки, расположенные в пространстве входов близко друг к другу, отображаются на близко расположенные элементы карты.

В совокупности, вышеописанные методы являются мощным инструментом классификации применительно к предметной области.

Результаты исследования функционирования метода. Темой исследования была выбрана оценка влияния наличия разнородных звуков в высказывании на качество идентификации диктора. Для проведения экспериментов были отобраны 20 дикторов: 10 женщин и 10 мужчин. Были разработаны следующие высказывания:

- «шиншилла шила шубу» — в дальнейшем «Высказывание 1»;
- «Клара украла кораллы» — в дальнейшем «Высказывание 2»;
- «шалость олигарха» — в дальнейшем «Высказывание 3».

Для каждого из дикторов было сделано по 600 записей каждого из высказываний с частотой дискретизации 44100 Гц. Тестовая и обучающая выборки для каждого из дикторов составили по 300 записей. В таблице приведены результаты экспериментов.

Таблица

Результаты экспериментов для каждого из высказываний

Высказывание	Кол-во	Позитивное срабатывание	Негативное срабатывание
Высказывание 1	6000	4878 (81.3%)	1122 (18.7%)
Высказывание 2	6000	5928 (98.8%)	72 (1.2%)
Высказывание 3	6000	5772 (96.2%)	228 (3.8%)
Итог	18000	16578 (92.1%)	1422 (7.9%)

Заключение. В представленной работе произведен анализ предметной области, описаны алгоритмы предварительной обработки сигнала, алгоритмы выделения критериев и концепции функционирова-

ния классификатора, а также описан авторский алгоритм устранения областей, не содержащих полезный сигнал. Описываемый метод был реализован, и проведены исследования влияния разнородных звуков на качество идентификации диктора. Из результатов исследования видно, что преобладание тональных звуков в идентифицирующем выражении значительно улучшает качество работы изложенного метода.

ЛИТЕРАТУРА

[1] Ле Н.В. Предварительная обработка речевых сигналов для системы распознавания речи. *Молодой ученый*, 2011, № 5, т. 1, с. 74–76.

[2] Запрягаев С.А., Коновалов А.Ю. Распознавание речевых сигналов. *Вестник ВГУ*, 2009, № 2, с. 39–48.

[3] *Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk Speech Recognition using MFCC. International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012)*, July 28–29, 2012б Pattaya (Thailand).

[4] *Сети встречного распространения*. [Электрон. ресурс]. http://neuronets.chat.ru/nets.html

[5] *Пригодность самоорганизующихся нейронных сетей (карт) Кохонена для задач визуализации и разведочного анализа данных*. [Электрон. ресурс] http://www.neuropro.ru/memo32.shtml

[6] Ларионов И.Б. *Карты Кохонена как способ восстановления мультимедийной информации. Омский государственный университет им. Ф.М. Достоевского*, 2010. [Электрон. ресурс] http://jre.cplire.ru/koi/oct10/3/text.html

Статья поступила в редакцию 10.06.2013

Ссылку на эту статью просим оформлять следующим образом:

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

Тассов Кирилл Леонидович родился в 1966 г., окончил МГТУ им. Н.Э. Баумана в 1991 г. Старший преподаватель кафедры «Программное обеспечение ЭВМ и информационные технологии» МГТУ им. Н.Э. Баумана. Автор научных работ в области теории распознавания образов и цифровой обработки сигналов. e-mail: ktassov@policesoft.ru

Дятлов Роман Андреевич родился в 1990 г. Студент кафедры «Программное обеспечение ЭВМ и информационные технологии» факультета «Информатика и системы управления» МГТУ им. Н.Э. Баумана с 2007 г. Область научных интересов: разработка средств анализа данных. e-mail: djatlik@mail.ru