УДК 004.934.1’1

РЕАЛИЗАЦИЯ АЛГОРИТМА ОБРАБОТКИ И РАСПОЗНАВАНИЯ РЕЧИ

Алюнов Д.Ю., Сергеев Е.С., Пигачев П.В., Мытников А.Н.

ФГБОУ ВПО «Чувашский государственный университет им. И.Н. Ульянова», Чебоксары, e-mail: dimitrie1@yandex.ru

Рассматриваются вопросы обработки, выделения информативных параметров и механизмов распознавания речи. В статье изложены основы алгоритма для практической реализации некоторых методов с учетом волновой природы звука (сигнала) и выделением наиболее существенных частот, воспринимаемых человеком при определенной энтропии, представлены расчеты MEL-фильтра, расчеты энтропии, применяемой при определении границ слов, косинусоидальное преобразование, алгоритм динамической трансформации времени, представлена акустическая модель слов. На основе представленного алгоритма разработано консольное приложение. Приложение работает со словарем, точность распознавания зависит от размера словаря. Приводятся графики зависимости MEL-шкалы от частоты, график MEL-частотных кепстральных коэффициентов, результаты вычисления сведены в таблицу. Представлен результат эксперимента на определение качества распознавания речи при использовании данного алгоритма.

Ключевые слова: распознавание речи, кепстральный анализ, обработка речи, энтропия, mel-преобразование, преобразование Фурье

IMPLEMENTATION OF THE ALGORITHM PROCESSING AND SPEECH RECOGNITION

Alyunov D.Yu., Sergeev E.S., Pigachev P.V., Mytnikov A.N.

Federal state budget educational institution of higher professional education «Chuvash State University named after I.N. Ulyanov», Cheboksary, e-mail: dimitrie1@yandex.ru

Questions of processing, allocation of informative parameters and mechanisms of recognition of the speech are considered. In article algorithm bases for practical realization of some methods taking into account the wave nature of a signal are stated and allocation of the most essential frequencies perceived by the person presented calculations of the MEL filter, calculations of the entropy applied at delimitation of words, cosinusoidal transformation, algorithm of dynamic transformation of time, the acoustic model of words is presented. On the basis of the presented algorithm the console application is developed. The appendix works with the dictionary, the accuracy of recognition depends on the dictionary size. Schedules of dependence of a MEL scale on frequency, the schedule MEL frequency the kepstaralnykh of coefficients are provided, results of calculation are tabulated. The result of experiment on determination of quality of recognition of the speech when using this algorithm is presented.

Keywords: speech recognition, cepstral analysis, speech processing, entropy, mel-transformation, the Fourier transform

При распознавании речи в первую очередь необходимо разбить ее на слова. Упростим задачу: пусть речь содержит в себе паузы (промежутки между словами), которые будут разделять слова. В этом случае нужно понять величину порога – значения, выше которого элемент сигнала является словом, а все, что ниже, – паузой между словами [2].

Энтропией будем считать меру беспорядка, меру неопределенности. В нашем случае энтропия показывает, как сильно осциллирует сигнал в пределах конкретного фрейма. Фреймом является малый отрезок, на которые мы разделяем исследуемый сигнал, следует указать, что фреймы идут не друг за другом, а накладываются. Для подсчета энтропии пронормируем сигнал, построим график плотности распределения значений сигнала в пределах одного фрейма, а энтропию рассчитаем следующим образом:

$$E = \sum_{i=0}^{N-1} P[i] \cdot \log_2 (P[i]). \qquad (1)$$

Для того чтобы отделить звук от тишины, её нужно с чем-то сравнивать. Опытным путем была подобрана величина порога, равная (0.1).

Таким образом, на вход нашей системы подается звуковой сигнал. Звук делится на фреймы – участки по 25 мс с перекрытием фреймов равным 10 мс. Для обработки звукового сигнала его следует преобразовать либо в виде спектра сигнала, либо в виде прологарифмированного спектра, с последующим масштабированием, поскольку это соответствует особенностям человеческого восприятия звука (Mel-шкала). Затем сигнал представляется в виде MFCC (Мел кепстральные коэффициенты) путем применения дискретного косинусоидального преобразования. MFCC обычно является вектором из тринадцати вещественных чисел, он представляют собой энергию спектра сигнала. Данный метод учитывает волновую природу сигнала, mel-шкала выделяет наиболее существенные частоты, воспринимаемые человеком, а количество

MFCC коэффициентов можно задать любым числом, что позволяет сжать фрейм и уменьшить количество обрабатываемой информации [3].

Рассмотрим алгоритм MFCC-преобразования получаемого звукового сигнала.

Получаемый звуковой сигнал дискретизируется:

$$x[n],\ 0 \le n < N. \qquad (2)$$

Представляем его в качестве Фурье преобразования:

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-2\pi i}{N} kn}, \quad 0 \le k < N. \qquad (3)$$

Рассчитываем гребенку фильтров, используя окно:

$$H_m = \begin{cases} 0 & k < f[m-1]; \\ \dfrac{(k - f[m-1])}{(f[m] - f[m-1])} & f[m-1] \le k < f[m]; \\ \dfrac{(f[m+1] - k)}{(f[m+1] - f[m])} & f[m] \le k \le f[m+1]; \\ 0 & k > f[m+1], \end{cases} \qquad (4)$$

где $f[m]$ будет равно

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_1) + m\frac{B(f_h) - B(f_1)}{M+1}\right); \qquad (5)$$

B(b) – представляем наши частоты в виде Мел-шкалы:

$$B^{-1}(b) = 700\left(\exp\left(\frac{b}{1125}\right) - 1\right). \qquad (6)$$

Где энергия окон будет равна

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

$$0 \le m < M. \qquad (7)$$

Получаем коэффициенты MFCC [4]:

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos\left(\pi n \left(m + \frac{1}{2}\right) \Big/ M\right),$$

$$0 \le n < M. \qquad (8)$$

Пусть наш фрейм представляется в виде дискретного вектора значения согласно формуле (2).

Вычислим спектр сигнала:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-2\pi i k n}{N}}, \quad 0 \le k < N. \qquad (9)$$

Обработаем сигнал окном Хэмминга, чтобы сгладить пульсации сигнала на краях [6].

$$H[k] = 0{,}54 - 0{,}46 \cdot \cos\left(\frac{2\pi k}{N-1}\right); \qquad (10)$$

$$X[k] = X[k] \cdot H[k],\ 0 \le k < N. \qquad (11)$$

По оси OX откладывается частота в Герцах, по оси OY – магнитуда, чтобы не связываться с комплексными величинами (рис. 1):

Mel представление показывает значимость отдельных частот звука для человека, зависит и от конкретных частот звука, и от громкости, и от тембра человека. Mel-шкала вычисляется следующим образом (прямое и обратное преобразование):

$$M = 1127 \cdot \log\left(1 + \frac{F}{700}\right); \qquad (12)$$

$$F = 700 \cdot \left(e^{M/1127} - 1\right). \qquad (13)$$

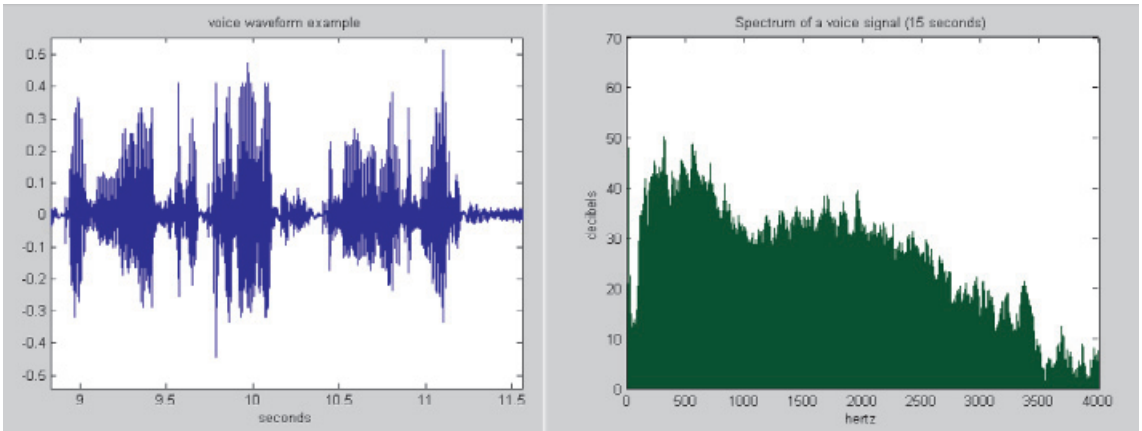

Рис. 1. Представление исходного сигнала в качестве Фурье преобразования

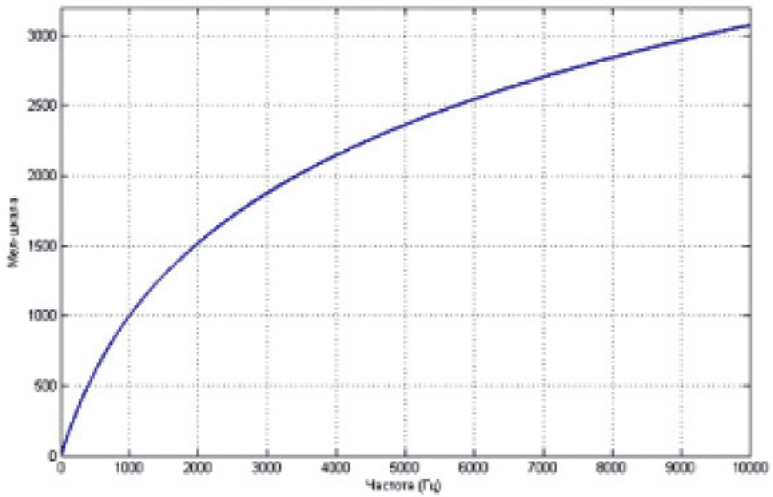

Рис. 2. График зависимости Мел-шкалы от частоты

График зависимости Мел-шкалы от частоты представлен на рис. 2.

Наибольшее распространение в системах распознавания речи получили именно эти единицы измерения, поскольку они соответствуют особенностям восприятия звука человеком.

Рассмотрим пример: дан фрейм длиной 256 отсчетов (выборок), частота звука 16 кГц. Пусть человеческая речь сосредоточена в диапазоне частот от 300 Гц до 8 кГц. Наиболее часто используемое количество Mel-коэффициентов равно десяти, его и будем использовать.

Сначала необходимо рассчитать гребенку фильтров, чтобы представить спектр в формате mel-шкалы. Мел-фильтр является треугольным окном, которое суммирует энергию на своем диапазоне частот и вычисляет mel-коэффициенты. Поскольку мы знаем количество коэффициентов, то сможем построить набор из десяти фильтров (рис. 3).

В области низких частот (те частоты, которые нам наиболее интересны) количество окон больше, что обеспечивает высокое разрешение. Это позволяет существенно повысить качество распознавания.

Для того чтобы найти энергию сигнала, перемножим вектор спектра сигнала и функцию окна, в результате чего получим вектор коэффициентов. Если их возвести в квадрат, представить в виде логарифма и получить из них кепстральные коэффициенты, то получим искомые mel-коэффициенты. Кепстральные коэффициенты можно получить как с помощью Фурье-преобразования, так и с помощью дискретного косинусоидального преобразования [6].

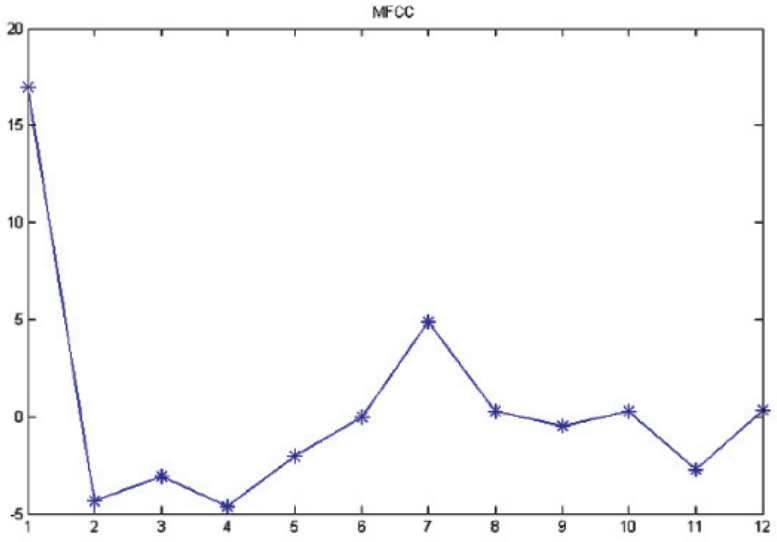

Рис. 3. Mel-частотные кепстральные коэффициенты

Диапазон частот составляет от 300 Гц до 8 кГц. На mel-шкале этот диапазон соответствует от 401,25 до 2834,99. Теперь строим двенадцать опорных точек для постройки десяти треугольных фильтров (Мел-шкала и шкала в герцах):

$$m[i] = [401{,}25;\ 622{,}50;\ 843{,}75;\ 1065{,}00;\ 1286{,}25;\ 1507{,}50;\ 1728{,}74;\ 1949{,}99;\ 2171{,}24;\ 2392{,}49;\ 2613{,}74;\ 2834{,}99]; \quad (14)$$

$$h[i] = [300;\ 517{,}33;\ 781{,}90;\ 1103{,}97;\ 1496{,}04;\ 1973{,}32;\ 2554{,}33;\ 3261{,}62;\ 4122{,}63;\ 5170{,}76;\ 6446{,}70;\ 8000]. \quad (15)$$

Как мы уже говорили, длина фрейма составляет 256 отсчетов сигнала, частота 16 кГц (откладывается по оси OX). Наложим рассчитанную шкалу на спектр сигнала.

$$f(i) = floor\left((frameSize + 1) \times h(i) / sampleRate\right), \quad (16)$$

что соответствует

$$f(i) = 4;\ 8;\ 12;\ 17;\ 23;\ 31;\ 40;\ 52;\ 66;\ 82;\ 103;\ 128. \quad (17)$$

По опорным точкам построим фильтры:

$$H_m(k) = \begin{cases} 0 & k < f(m-1); \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k \le f(m); \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \le k \le f(m+1); \\ 0 & k > f(m+1). \end{cases} \quad (18)$$

Фильтр перемножается со спектром:

$$S[m] = \log\left(\sum_{k=0}^{N-1} \left|X_a[k]\right|^2 H_m[k]\right), \quad 0 \le m < M. \quad (19)$$

Мел-фильтры применяются к энергии спектра, затем полученные значения логарифмируются.

Дискретное косинусоидальное преобразование (ДКТ) применяется для получения кепстральных коэффициентов, оно сжимает полученные результаты, повышает вклад первых коэффициентов и понижает вклад последних.

$$C[l] = \sum_{m=0}^{M-1} S[m] \cos\left(\pi l \left(m + \frac{1}{2}\right) \Big/ M\right), \quad 0 \le l < M. \quad (20)$$

Получается, что у нас имеется 12 коэффициентов (рис. 3):

В итоге небольшой конечный набор значений (например, двенадцать коэффициентов в нашем случае) позволяет заменить использование огромного числового массива отсчетов сигнала, либо спектра сигнала, либо периодограммы сигнала.

Каждому слову конечной длины соответствует набор мел-частотных кепстральных коэффициентов. Затем необходимо найти наиболее близкую модель для определенного набора мел-частотных кепстральных коэффициентов. Для этого мы ищем евклидово расстояние между вектором мел-частотных кепстральных коэффициентов и вектором исследуемой модели. Искомой является та модель, у которой рассчитываемое расстояние наименьшее.

Набор MFCC коэффициентов для одного и того же слова может отличаться, например, в том случае, если слово произносится двумя разными людьми, либо скорость произношения отличается. Для этих целей используется алгоритм динамической трансформации времени. Он рассчитывает оптимальную деформацию времени между сравниваемыми временными последовательностями [2].

	-2	10	-10	15	-13	20	-5	14	2
3	5	12	25	37	53	70	78	89	90
-13	16	28	15	43	37	70	78	105	104
14	32	20	39	16	43	43	62	62	74
-7	37	37	23	38	22	49	45	66	71
9	48	38	42	29	44	33	47	50	57
-2	48	50	46	46	40	55	36	52	**54**

Рис. 4. Результаты расчетов

Допустим, у нас есть два числовых ряда $(a_1, a_2, ..., a_n)$ и $(b_1, b_2, ..., b_m)$. Их длина может отличаться. Будем использовать Евклидово расстояние для расчета локальных отклонений между соответствующими элементами двух числовых рядов. В итоге получим матрицу отклонений $N \times M$:

$$d_{ij} = |a_i - b_j|, \quad i = \overline{1,n}, \quad j = \overline{1,m}. \quad (21)$$

Критерий оптимизации для расчета минимального расстояния:

$$a_{ij} = d_{ij} + \min\left(a_{i-1,j-1}, a_{i-1,j}, a_{i,j-1}\right), \quad (22)$$

где a_{ij} – min расстояние между последовательностями $(a_1, a_2, ..., a_n)$ и $(b_1, b_2, ..., b_m)$. Данный способ позволяет вычислить минимальную длину траектории движения от элемента a_{11} до элемента b_{nm} (рис. 4).

Алгоритмы динамической трансформации времени полезны для распознавания отдельно стоящих слов при наличии словаря. В случае проблемы распознавания и обработки слитной речи гораздо более полезны СММ (скрытые марковские модели).

На основе представленного алгоритма разработано консольное приложение, позволяющее реализовать представленный алгоритм (рис. 5).

Для запуска приложения Sound необходимо вызвать командную строку при помощи RUN.BAT. Далее из командной строки вызываем команду: Sound -h. Возможны следующие вызовы команд: Список всех доступных моделей: Sound -l; Разделение источника в образцы: Sound -i samples/female1.wav -s_split; Добавление образца в модель: Sound -i samples/female1/1.wav -a odin; Распознавание образца: Sound -i samples/female1/1.wav -r; Модульные тесты: unit_tests --gtest_filter=MATH_MFCC.

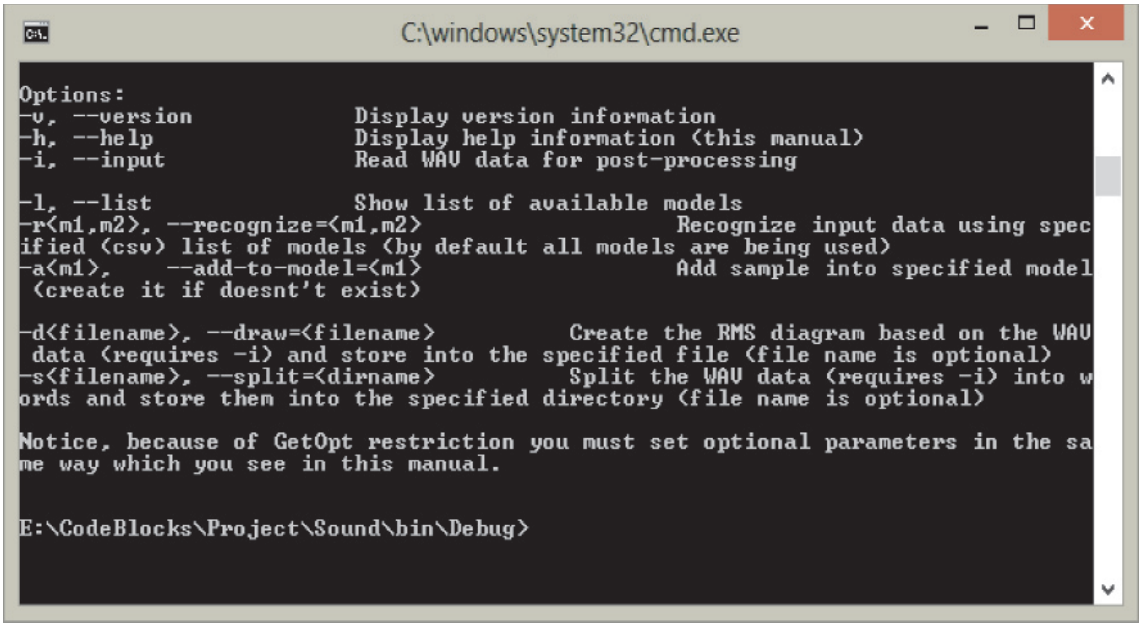

Рис. 5. Работа разработанной программы в командной строке

В первую очередь из main создается экземпляр класса Command Processor. Процессор создает команды, основанные на входных параметрах. Работа начинается вывода команды input. При этом происходит вызов метода Audio Data Command::read Data. При этом происходит запись данных в структуру wavData и поле wavData, класса context заполняется данными wavData. Данные очищаются от шумов, нормализуются.

Все этапы получения MFCC-коэффициентов выполняются в методе MFCC::transform().

Таким образом, принцип работы заключается в разбивке речи на слова на основе вычисления энтропии, эмпирическим методом было вычислено пороговое значение 0,1, затем идет разбивка на фреймы и вычисляются mel-коэффициенты. Затем идет сравнение со словарем. В качестве эксперимента брались записи слитной речи мужских и женских голосов, записанных в тихой комнате с отсутствующим шумом. Качество распознавания, показанное данным методом составило 75%.

Список литературы

1. Алюнов Д.Ю. Классификация помех и искажений в речевом сигнале // Наука и образование в жизни современного общества: сборник научных трудов по материалам Международной научно-практической конференции: в 12 частях. – 2015. – С. 14–15.

2. Кибкало А.А. Разработка системы распознавания русской речи // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. – 2003. – Вып. 3. – С. 8–20.

3. Михайлов В.Г., Златоустов Л.В. Измерение параметров речи. – М.: Радио и связь, 1987. – 167 с.

4. Сергеев Е.С., Пигачев П.В. Дифонный синтезатор речи // Теоретические и прикладные аспекты современной науки. – 2014. – № 6–3. – С. 114–116.

5. Rabiner L.R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proc. of IEEE, Feb. 1989.

6. Ingle V., Proakis J. Digital Signal Processing Using Matlab V4 – Boston: ITP, 1997.