УДК 004.896

РАСЧЕТ MEL – КОЭФФИЦИЕНТОВ ЗВУКОВОГО СПЕКТРА ДЛЯ МОДУЛЯ ИДЕНТИФИКАЦИИ ГОЛОСОВЫХ КОМАНД УПРАВЛЕНИЯ МОБИЛЬНЫМРОБОТОМ

к.т.н. А.А. Левтеров[1], Ю.А. Нечитайло[2], М.А. Лычман[2]
Национальный университет гражданской защиты Украины
Харьковский национальный автомобильно-дорожный университет

*В статье приведен расчет MEL – кепстральных коэффициентов (*MFCC*) для системы распознавания голосовых команд низкочастотного диапазона для управления мобильным роботом. Приведены спектрограммы команды «СТОП». Описан метод идентификации команд методом сравнения с применением алгоритма DWT.*

MEL calculation – the capstral of coefficients (MFCC) for the voice recognition system commands of low-frequency range for control of the mobile robot has been given in article. Spectrograms of the "STOP" team have been provided. The method of identification commands by method comparison with application of an algorithm of DWT has been described.

*У статті наведено розрахунок MEL – кепстральних коефіцієнтів (*MFCC*) для системи розпізнавання голосових команд низькочастотного діапазону для управління мобільним роботом.. Приведені спектрограми команди «СТОП». Описано метод ідентифікації команд методом порівняння із застосуванням алгоритма DWT.*

Ключевые слова: кепстральный коэффициент, мобильный робот, голосовая команда

При беспроводном управлении мобильным дорожным роботом в условиях сильных радиопомех или расположенных близко от зоны управления вещающих станций в диапазоне радиочастот управления может возникнуть ситуация, когда команды управления неверно распознаются навигационной системой, и, следовательно, дальнейшее управление роботом становится невозможным. В качестве решения данной проблемы предлагается использовать систему управления (как аварийную) звуковыми командами, в том числе и голосовыми, в диапазоне частот от 100 до 3000Гц, с применением направленного микрофона.

При такой системе управления возникает проблема распознавания голосовых команд, поскольку возможность управления должна быть у любого оператора, находящегося в зоне управления.

Для решения этой задачи предлагается использовать систему распознавания голоса с выделением характерных частот акустического спектра для каждой команды.

В системах распознавания речи применяют различные методы, например, распознавание по образцу, применение нейронных сетей, генетические алгоритмы, вейвлет-преобразования, а также статистический анализ. Предлагаемые методы в [4, 15, 19] являются весьма ресурсозатратными. Для систем же управления роботом требуется быстрая обработка поступающих команд управления. Для повышения быстродействия системы предлагается применить MEL-преобразование [3]. Данный метод не лишен недостатков, но является приемлемым в данном случае.

Для того чтобы система идентифицировала верно команды управления необходимо: на фоне шумовых помех выделить голосовую команду и при помощи MEL фильтрации ее идентифицировать.

Для анализа данных воспользуемся методом фреймов и разбиения слов. Опытным путём установлено, что оптимальная длина фрейма должна соответствовать промежутку в 0,01с, а наложение фреймов (перекрытие) — 1/2 от длины фрейма. С учётом того, что средняя длина слова составляет приблизительно 0,4÷0,8 с, такой шаг дает примерно от 80 до160 фреймов (не менее) на слово [3].

Если команда состоит из нескольких слов, то необходимо выделить (разделить каждое слово) в поступающем сигнале управления [5]. В нашем случае паузами между словами будет фон зоны управления роботом.

Существует несколько способов разделения сигнала на отдельные слова. Согласно [6, 8] выбираем метод анализа энтропии, которая определяет, как сильно изменяется сигнал управления в рамках заданного фрейма. Для расчета энтропии конкретного фрейма предположим, что принятый сигнал управления будет нормирован и его значения лежат в диапазоне [-1;1], тогда энтропия фрейма:

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

Для того, чтобы отделить полезный сигнал от фона, нужно его сравнивать с ранее известным фоном либо сигналом, в зависимости от того, что нужно выделить. В некоторых источниках [6] рекомендуют брать порог энтропии равным среднему между её максимальным и минимальным значениями или подбирать значение её порога в виде константы, характерной для данного фона. В случае четкой команды оператора проблемы данного подхода, изложенные в [9, 12, 14, 17], будут сведены к минимуму.

Разделив сигнал на отдельные фреймы, получим набор фреймов, соответствующих определенным словам (командам).

Для распознавания команд воспользуемся методом MFCC (MEL-frequency cepstral coefficients) [10].

Пусть фрейм сигнала имеет вид вектора

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

где N — размер фрейма (256 или 512, но не более 1024).

Теперь получим спектр сигнала с помощью дискретного преобразования Фурье[13] (см. рис. 1, 2):

$$X[k]=\sum_{n=0}^{N-1}x[n]\cdot e^{-2\pi kn/N},0\le k<N \qquad (3)$$

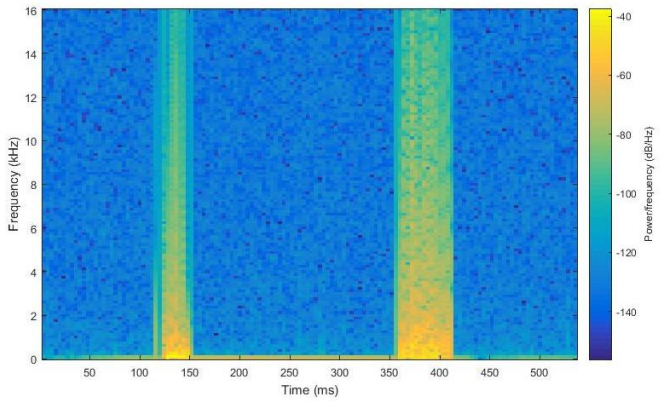

Рисунок 1 – Спектр голосовой команды «СТОП»

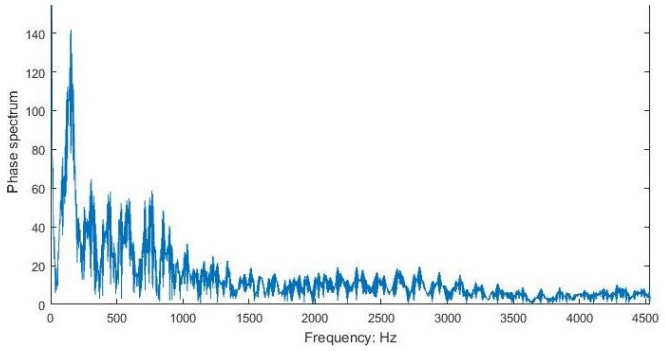

Рисунок 2 – Фазовый спектр голосовой команды «СТОП»

Далее применим оконную функцию, например, Хэмминга[11](вид оконной функции выбирается исходя из формы сигнала), что бы произвести операцию смуфинга (сглаживания) значения на границах фреймов.

$$H[k]=0.54-0.46\cdot\cos(2\pi k/(N-1)) \qquad (4)$$

Тогда результатом будет вектор вида:

$$X[k]=X[k]\cdot H[k],0\le k<N \qquad (5)$$

Результат применения оконной функции Хемминга приведен на рис.3.

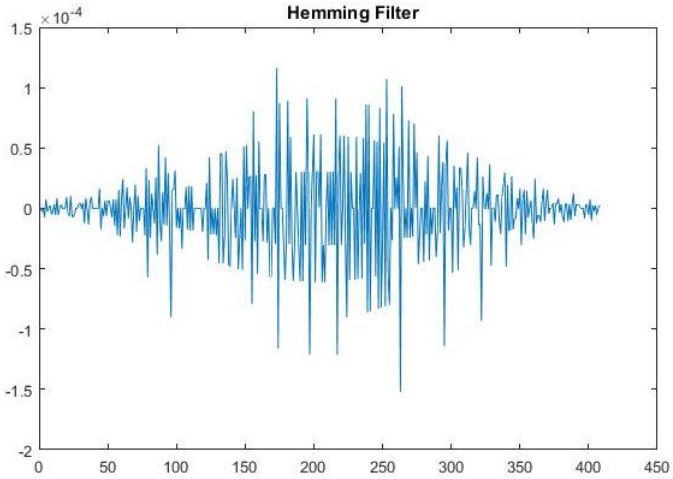

Рисунок3 – Сигнал после применения оконной функции Хемминга

Далее перейдем к вычислению MEL – коэффициентов.

Преобразуем частоту сигнала в MEL-шкалу частот по следующей формуле:

$$M=1127\cdot\ln(1+f/700) \qquad (6)$$

Размер фрейма 512 элементов и известно, что частота звука в данном фрейме 25000 Гц. Для помехоустойчивости приемника сигнала управляющие команды, по возможности, будут выбраны в диапазоне частот от 200 до 3000 Гц. Количество MEL-коэффициентов, положим M = 24, как рекомендуют в [2,7], но следует учесть, что с увеличением числа коэффициентов возрастает и время обработки.

Разложим полученный спектр по MEL-шкале, что приведет к получению набора фильтров.

Преобразуем частотный диапазон 200 – 3000 Гц согласно (6) в MEL-частоту, что соответствует диапазону (283,23 – 1876,46 Гц). Для построения 24 фильтров потребуется 26 опорных точек (см. табл. 1).

Таблица 1

i	f_i^{mel}	i	f_i^{mel}	i	f_i^{mel}	I	f_i^{mel}	i	f_i^{mel}	i	f_i^{mel}
1	**283,23**	6	601,88	11	920,52	16	1239,17	21	1557,81	26	**1876,46**
2	346,96	7	665,60	12	984,25	17	1302,90	22	1621,54		
3	410,69	8	729,33	13	1047,98	18	1366,63	23	1685,27		
4	474,42	9	793,06	14	1111,71	19	1430,35	24	1749,00		
5	538,15	10	856,79	15	1175,44	20	1494,08	25	1812,73		

Проделав обратное преобразование, при помощи формулы (2):

$$f=700\cdot\left(e^{M/1127}-1\right), \qquad (7)$$

получим набор частот (см. табл. 2).

Таблица 2

i	f_i^r	i	f_i^r	i	f_i^r	i	f_i^r	i	f_i^r	i	f_i^r)
1	**200**	6	494,07	11	884,24	16	1401,90	21	2088,71	26	**3000**
2	252,35	7	563,54	12	976,41	17	1524,18	22	2250,94		
3	307,76	8	637,05	13	1073,94	18	1653,58	23	2422,62		
4	366,39	9	714,83	14	1177,14	19	1790,50	24	2604,28		
5	428,43	10	797,14	15	1286,34	20	1935,39	25	2796,51		

Теперь построим функцию $f^{mel}\left(f^r\right)$ (см. рис. 4).

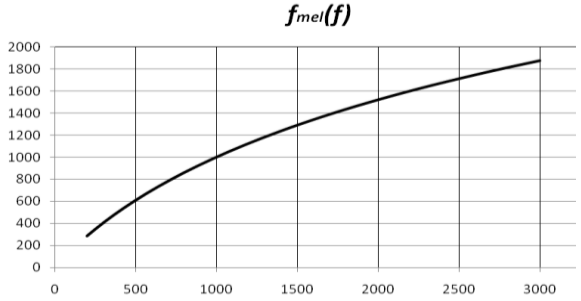

Рисунок 4 – Зависимость MEL частоты от частоты сигнала

Далее наложим полученную шкалу на спектр фрейма:

$$f_i = \left\lfloor \frac{(f_s+1)\cdot f_i^r}{F_{sr}} \right\rfloor \quad (8)$$

где f_s– размер фрейма, равный 512; f_i^r – значение i- ой (i=1,2,..24) частоты после обратного преобразования по формуле (7); F_{sr} – максимальная частота в анализируемом спектре, в нашем случае 25000 Гц (выбрана исходя из акустического диапазона в зоне управления).

В результате получим набор значений F_i (см. табл. 3)

Таблица 3

i	$f(i)$	i	$f(i)$	i	$f(i)$	i	$f(i)$	i	f_i	i	f_i
1	4	6	10	11	18	16	28	21	42		61
2	5	7	11	12	20	17	31	22	46		
3	6	8	13	13	22	18	33	23	49		
4	7	9	14	14	24	19	36	24	53		
5	8	10	16	15	26	20	39	25	57		

Зная опорные точки спектра, построим необходимые фильтры по следующей формуле:

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

Применение фильтра заключается в попарном перемножении его значений со значениями спектра. Результатом этой операции является mel-коэффициент. Т.к. число фильтров равно M, то, следовательно, коэффициентов тоже M.

$$S[m] = \ln\left(\sum_{k=0}^{N-1} |X[k]|^2 \cdot H_m[k], 0 \le m < M \right) \quad (10)$$

Стоить учесть, что MEL -фильтры применяют не к значениям спектра, а к его энергии. Принято считать, что так снижается чувствительность коэффициентов к шумам [18].

Дискретное косинусное преобразование (DCT) используется для того, чтобы получить MEL-коэффициенты (см. рис. 5), т.е. повысить значимость первых коэффициентов и уменьшить значимость последних. В данном случае применим:

$$C[l] = \sum_{m=0}^{M-1} S[m] \cdot \cos\left(\pi l (m+0.5)/M\right), 0 \le l < M \quad (11)$$

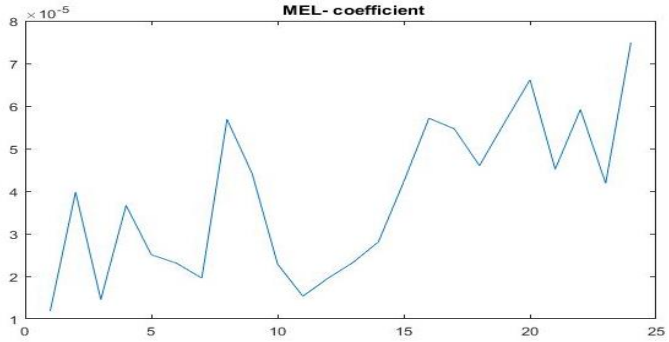

Рисунок 5 – MEL коэффициенты

В итоге для каждого фрейма получим набор из M MFCC-коэффициентов, который будет использован для дальнейшего анализа.

Поскольку задача заключается в распознавании слова из некоторого набора (может быть, в том числе, и голосовым фоном на фоне помех окружающей среды-зоны управления роботом), то далее поставим в

соответствие каждому слову L наборов mfcc-коэффициентов соответствующих записей. Это соответствие обозначим как «CSC» (compliance set coefficients). Теперь задача сводится к подбору наиболее «близкой» CSC для некоторого набора MFCC-коэффициентов (распознаваемого слова). Задача решается в два этапа:

–для каждой CSC находим среднее расстояние между идентифицируемым mfcc-вектором и векторами CSC;

–в качестве верной выбираем ту CSC, среднее расстояние до которой будет наименьшим.

Следует отметить, что модель с MFCC-вектором –последовательность MFCC-субвекторов размерности M, полученных из фреймов. т.е. в качестве значений расстояний должны использовать расстояния между MFCC-субвекторами фреймов. Т.к. одно и тоже слово может произносится разными операторами, размер MFCC-вектора для одного и того же слова может быть разный, но эта проблема решается при помощи DTW (Dynamic Time Warping) алгоритма [16]. Он рассчитывает оптимальную деформацию времени между сравниваемыми временными последовательностями [1].

Выводы. Полученный набор MEL значений (24 коэффициента) позволяет заменить использование большого массива отсчетов сигнала или спектра сигнала, что значительно сокращает время обработки команд модулем идентификации мобильного робота, и тем самым повышает быстродействие системы «оператор-робот». Также снижение времени реакции данной системы связано с повышением эффективности применения голосовых команд в условиях сильных радиопомех при управлении роботом.

ЛИТЕРАТУРА

1. Ghazi Al-Naymat, Sanjay Chawla, Javid Taheri. SparseDTW: A Novel Approach to Speed up Dynamic Time Warping. arXiv:1201.2969v1 [cs. DB].–2012.–P. 1-17.

2. Алюнов Д.Ю., Сергеев Е.С., Пигачев П.В., Мытников А.Н. Реализация алгоритма обработки и распознавания речи // Современные наукоемкие технологии. – 2016. – № 3-2. – С. 225-230

3. X. Huang, A. Acero, and H. Hon. Spoken Language Processing: A guide to theory, algorithm, and system development. Prentice Hall PTR Upper Saddle River, NJ, USA.–2001.–960p.

4. Young S. A Review of Large-Vocabulary Continuous Speech Recognition, IEEE Signal Processing Magazine.– 1996, pp. 45-57

5. Бондаренко, И.Ю. Анализ эффективности метода нечёткого сопоставления образов для распознавания изолированных слов / И.Ю. Бондаренко, О.И. Федяев // Интеллектуальный анализ информации ИАИ-2006: сб. трудов VI междунар. науч. конференции; под ред. Т.А. Таран. – К.: Просвіта, 2006. С. 20–27.

6. Рабинер Л.Р., Цифровая обработка речевых сигналов: [пер. с англ.] / Л.Р. Рабинер, Р.В. Шафер. – М.: Радио и связь, 1981. – 251 с.

7. О.С. Агашин, О.Н. Корелин. Методы цифровой обработки речевого сигнала в задаче распознавания изолированных слов с применением сигнальных процессоров./ Труды Нижегородского государственного технического университета им. Р.Е. Алексеева.– № 4.– 2012.– С. 32-44

8. L. Rabiner, Biing-Hwang Juang. Fundamentals of Speech Recognition.1993. — 507 с.

9. Маркел Дж.Д., Грэй А. Х. Линейное предсказание речи: Пер. с английского / Под редакцией Ю. Н. Прохорова и В. С. Звездина. — М.: Связь, 1980.– 308с.

10. Taabish Gulzar, Anand Singh. Comparative AnalysisofL PCC, MFCC and BFCC // International Journal of Computer Applications. — 2014. — № 101(12). — С. 22–27.

11. Смит С. Цифровая обработка сигналов. Практическое руководство для инженеров и научных работников. — М.: Додэка-XXI, 2012. — 720 с.

12. Воробьева С. А. Методы распознавания речи // Молодой ученый. — 2016. — №26. — С. 136-141.

13. Зорич В. А. Математический анализ. — М.: Физматлит, 1984. — 544 с.

14. Первушин Е. А. Обзор основных методов распознавания дикторов// Математические структуры и моделирование. –2011.– Вып. 24.С. 41–54.

15. Сорокин В. Н., Вьюгин В. В., Тананыкин А. А. Распознавание личности по голосу: аналитический обзор // Информационные процессы. — 2012. — Т. 12, № 1. – 2012.–С. 1-30

16. Keogh E, Ratanamahatana C. Exact indexing of dynamic time warping./ Knowledge and Information Systems(KIS) 7(3).–2004.–С. 358–386.

17. Огнев И. В. Распознавание речи методами скрытых марковских моделей в ассоциативной осцилляторной среде / И. В. Огнев, П. А. Парамонов // Известия высших учебных заведений. Поволжский регион. Технические науки. –2013. – № 3 (27). – С. 115–126.

18. Заковряшин А. С. Применение распределений мел-частотных кепстральных коэффициентов для голосовой идентификации личности / А. С. Заковряшин, П. В. Малинин, А. А. Лепендин - Известия АлтГУ. 2014. №1 (81) С.156-160.

19.Kai Fu Li, Hsiao-Wuen Hon. An overview of the Sphinx Speech Recognition Systems – [Электронный ресурс]. // The Robotics Institute – Carnegie Mellon University. – Режим доступа. – URL: http://www.ri.cmu.edu/pub_files/pub2/lee_k_f_1990_1/lee_k_f_1990_1.pdf (дата обращения 10.02.2017).