УДК 004.934.2

РАСПОЗНАВАНИЕ ГОЛОСОВЫХ КОМАНД УПРАВЛЕНИЯ РОБОТОМ НА ОСНОВЕ СПЕКТРА МОДУЛЯЦИИ РЕЧЕВОГО СИГНАЛА

© 2017. *Т.В. Шарий, Е.И. Черкашин*

В статье предлагается робастный метод распознавания голосовых команд управления мобильным роботом, опирающийся на спектр модуляции речевого сигнала. Описан алгоритм вычисления спектра модуляции, расчета и применения банка фильтров слуховой системы в частотной области. Исследовано влияние шума и эффекта реверберации помещений на спектр модуляции сигнала. Приведены результаты экспериментов по распознаванию голосовых команд на основе классификатора ближайших соседей и машин опорных векторов в дикторозависимой тестовой среде.

Ключевые слова: речевой сигнал; спектр модуляции; банк фильтров; классификатор.

Введение. В настоящее время автоматическое распознавание речи (Automatic Speech Recognition, ASR) [1] является одной из главных задач, решаемых в рамках передового направления информационных технологий – машинного обучения (Machine Learning) [2, 3]. Интерес к этой проблеме подтверждается постоянным ростом числа научных статей и речевых корпусов для проверки различных моделей анализа и распознавания речи, а также встраиванием модулей распознавания речи во многие современные операционные системы. В последние годы наблюдался значительный прогресс в решении всех задач, связанных с ASR. Ключевую роль при этом сыграли статистические аспекты моделей глубокого обучения, рост вычислительных мощностей и развитая инфраструктура сбора и обработки речевых данных. Вместе с тем, вопросы инженерии признаков (feature engineering) по-прежнему представляют как научный, так и практический интерес. В частности, актуален вопрос выбора робастных признаков речевого сигнала, менее чувствительных к шуму и эффектам реверберации в помещениях. Кроме того, распознавание речи востребовано в робототехнике, в которой важна вычислительная простота применяемых методов, а размер словаря часто является относительно небольшим.

В данной статье для распознавания голосовых команд управления мобильным роботом исследуется метод, основанный на вычислении и анализе спектра амплитудной модуляции речевого сигнала, представляющего голосовую команду. Спектры модуляции хорошо зарекомендовали себя в разных задачах, связанных с обработкой речи: оценивание уровня шума и разборчивости речи [4], шумоподавление [5, 6], получение цифровых отпечатков [7] и распознавание фонем [8]. В то время как традиционный оконный анализ сигнала (например, метод мел-частотных кепстральных коэффициентов (MFCC) на основе кратковременного преобразования Фурье) позволяет получить последовательность векторов признаков во времени, спектр модуляции может интегрально и относительно компактно описать всю голосовую команду.

Постановка задачи. Целью данной работы является программная реализация и исследование метода параметризации речевых сигналов на основе спектра модуляции голосовых команд и анализ показателей его эффективности с помощью классификатора ближайших соседей и машин опорных векторов. Словарь для проверки адекватности модели состоит из 11 команд управления роботом. Информационная технология включает процессы записи и сохранения речевых статистических данных, их обработки

стационарным шумом и применения эффекта реверберации, обучения классификаторов на основе вычисленных спектров модуляции, визуализации результатов.

Амплитудная модуляция речевых сигналов. Важным свойством человеческой речи является то, что большая часть лингвистической информации, на основе которой человек распознает слова в речевом потоке, содержится в относительно медленных изменениях речевого сигнала. Согласно одной из психоакустических гипотез, в слуховой системе человека существуют каналы, настроенные на восприятие модуляционных частот по аналогии со спектральными частотами в критических полосах слуха, т.к. люди способны распознавать речь в сложных условиях, в частности, в присутствии стационарного и импульсного шума, реверберации и иных искажений. Рассмотрение особенностей слуховой системы человека всегда представляло интерес для определения параметров речи, которые содержат максимальный объем лингвистической информации. Такие параметры отражают инвариантные характерные свойства акустических сигналов, устойчивые к различным преобразованиям, фильтрации и цифровому сжатию.

Модуляция речи вызывается изменениями общего уровня энергии сигнала, происходящими при чередовании фонем, а также изменениями в спектральных распределениях формант, специфических для каждого частотного диапазона. Изменения спектральных характеристик речи во времени могут быть представлены в виде спектра амплитудной модуляции. Спектр модуляции вычисляется на основе спектрального анализа огибающей сигнала:

$$M[n,k,q]=\sum_{p=0}^{P-1}E_k[n+p]e^{-j2\pi pq/P},$$

где $E_k[n]$ представляет собой сигнал траектории изменения амплитуды речевого сигнала в k-ом частотном диапазоне, q – модуляционная частота, P – размер преобразования Фурье.

Классическая схема вычисления спектра модуляции [8] представлена на рис.1.

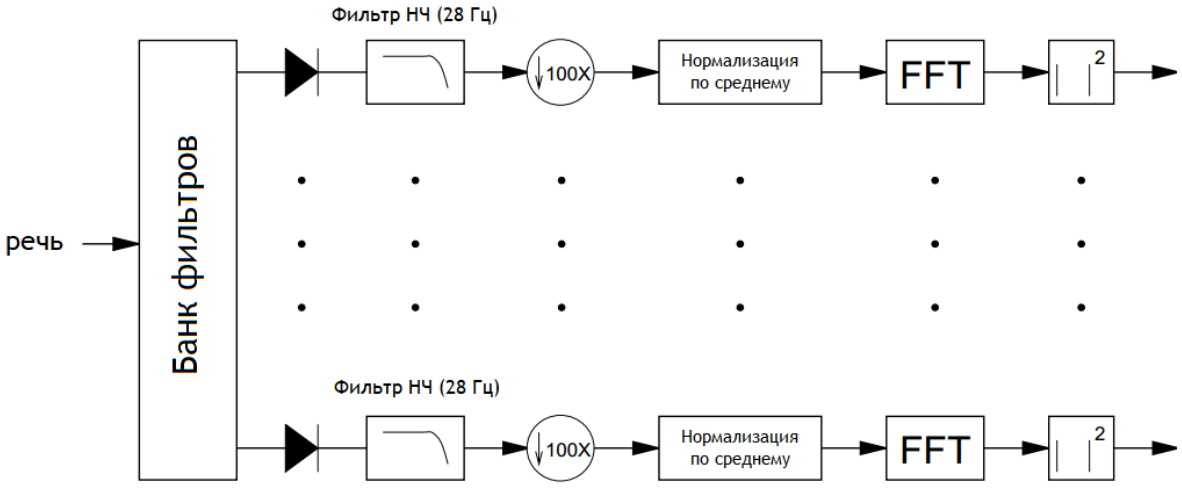

Рисунок 1 – Схема вычисления спектра амплитудной модуляции

Согласно приведенной схеме, речевой сигнал обрабатывается банком фильтров в определенном частотном диапазоне (обычно до 4000 Гц) с полосами пропускания,

соответствующими критическим полосам слуха, с минимальным перекрытием. В каждой полосе выделяется огибающая сигнала путем низкочастотной фильтрации выпрямленной траектории (однополупериодное выпрямление, half-wave rectification) с помощью КИХ-фильтров с частотой среза 28 Гц. После стократной децимации и нормализации по среднему уровню к полученным сигналам применяется повторно преобразование Фурье в окне длительностью 250 мс со взвешиванием оконной функцией Хемминга. Процедура повторяется каждые 12,5 мс с целью фиксирования временных изменений и динамических свойств сигнала.

В данной работе оценка спектра амплитудной модуляции производится похожим, но более простым в вычислительном отношении способом, подходящим для задачи распознавания голосовых команд. На первом шаге вычисляется спектрограмма сигнала как результат кратковременного преобразования Фурье с размером окна 31,25 мс без перекрытия (все числовые параметры задаются исходя из частоты дискретизации тестового сигнала 16 кГц, но могут настраиваться пользователем). К каждому спектру в спектрограмме применяется банк фильтров, описываемый ниже. Полученные значения энергии сигнала в каждой частотной полосе трактуются как отдельные сигналы с частотой дискретизации 32 Гц. Эти сигналы нормализуются по среднему уровню, после чего к ним применяется преобразование Фурье в окне длительностью 2 с, достаточной для представления речевой команды, произнесенной с обычной интонацией. Выходом приведенного алгоритма является модель спектра модуляции, представляющая собой вектор коэффициентов, образованный путем конкатенации всех спектров в каждом частотном канале.

Многочисленные исследования показывают, что медленные изменения сигнала с частотой модуляции 4 Гц особенно важны с точки зрения восприятия речи человеком и в значительно меньшей степени подвержены, например, искажениям в процессе GSM сжатия, чем традиционно используемые спектральные дескрипторы сигнала: MFCC, частота переходов уровня сигнала через ноль, спектральный центроид, спектральная равномерность и т.д. В связи с этим, предложено также исследовать «укороченную» модель спектра модуляции в виде вектора, в который входят только те коэффициенты, которые соответствуют частоте модуляции 4 Гц.

Банк фильтров. Выше было отмечено, что обработка сигнала производится в отдельных частотных диапазонах после полосовой фильтрации. С целью отражения особенностей восприятия речи человеком, а также для снижения размерности результирующего вектора признаков целесообразно использовать в качестве полосовых фильтров психоакустические банки фильтров: мел-фильтры, барк-фильтры или гамматон-фильтры. Влияние формы и числа фильтров на спектр модуляции речевого сигнала является темой отдельных исследований. В данной работе для вычисления спектра модуляции применяются 12 перекрывающихся мел-фильтров в частотном диапазоне от 100 Гц до 4200 Гц.

Фильтрация производится в частотной области путем перемножения амплитудного спектра Фурье $\{X_k\}_{k=0}^{N-1}$ сигнала в текущем окне анализа и амплитудно-частотной характеристики каждого из полосовых фильтров с последующим суммированием всех отсчетов:

$$E_i = \sum_{k=0}^{N-1} |X_k| \cdot H_i(f_k), \quad i = 1..M,$$

где N – размер преобразования Фурье, M=12 – количество мел-фильтров.

Амплитудно-частотные характеристики фильтров H_i имеют треугольный вид и изображены на рис.2. Они могут быть заданы аналитически в виде формул:

$$H_i(f_k) = \begin{cases} 0 & f_k \le f_{c_{i-1}} \\ \dfrac{f_k - f_{c_{i-1}}}{f_{c_i} - f_{c_{i-1}}} & f_{c_{i-1}} < f_k \le f_{c_i} \\ \dfrac{f_{c_{i+1}} - f_k}{f_{c_{i+1}} - f_{c_i}} & f_{c_i} < f_k \le f_{c_{i+1}} \\ 0 & f_k > f_{c_{i+1}} \end{cases} \quad (1)$$

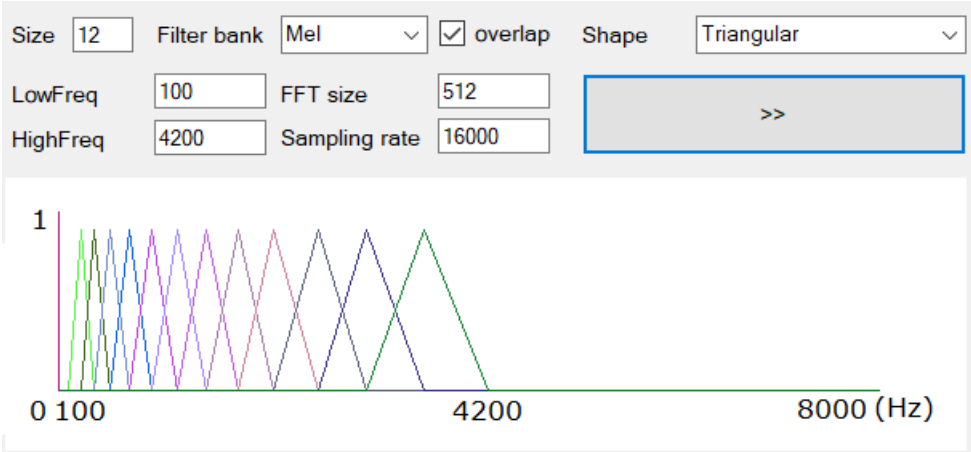

Рисунок 2 – Банк перекрывающихся треугольных мел-фильтров

В формуле (1) значения центральных (пиковых) мел-частот f_{c_i} рассчитываются путем преобразования частот из шкалы герц в шкалу мел по формуле:

$$f_{c_i} = 700 \cdot (e^{h_{c_i}/1127} - 1),$$

где h_{c_i} – i-ая частота в шкале герц. Центральные частоты распределены на шкале мел равномерно, на шкале герц – логарифмически.

Классификатор голосовых команд. Последним шагом алгоритма распознавания голосовых команд является принятие решения об отнесении спектра модуляции обработанного речевого сигнала к какой-либо команде из словаря (классификация команды). В области машинного обучения и анализа данных существует множество эффективных статистических классификаторов. Т.к. на данном этапе исследований акцент делается на цифровой обработке сигнала и анализе спектра модуляции, то были выбраны базовые (baseline) алгоритмы классификации: классификатор на основе k ближайших соседей (k-nearest neighbors, kNN) и машины опорных векторов (Support Vector Machine, SVM) [9, 10]. С увеличением объема статистических данных разница в эффективности различных классификаторов, как правило, будет уменьшаться.

Алгоритм kNN не предполагает стадии обучения, а работает сразу в режиме распознавания. После вычисления спектра модуляции сигнала находятся k уже размеченных в обучающей выборке спектров, наиболее близких к данному. В качестве метрики используется евклидово расстояние. Затем среди отобранных спектров модуляции путем взвешенного голосования окончательно определяется распознанная

команда. В данной работе k=7, но значение данного параметра может быть также установлено в результате перебора по методологии «поиска по сетке» (Grid Search).

Алгоритм обучения SVM находит коэффициенты гиперплоскости, максимально разделяющей входные образы команд в пространстве признаков (спектров модуляции). Решающая функция классификатора SVM задана формулой:

$$h(x) = \mathrm{sgn}(\sum_{i=1}^{m} \lambda_i y_i K(\mathbf{x_i}, \mathbf{x}) - \omega_0),$$

где λ_i и ω_0 – коэффициенты; $\mathbf{x_i}$ и y_i – входной вектор спектра модуляции и соответствующее ему значение из обучающей выборки (0 или 1), соответственно; $K(\mathbf{x_i}, \mathbf{x}) = \phi(\mathbf{x_i}, \mathbf{x})^T \phi(\mathbf{x_i}, \mathbf{x})$ – ядро. В качестве ядра выбраны радиально-базисные функции:

$$\phi(\mathbf{x_i}, \mathbf{x}) = \exp(-\gamma \|\mathbf{x} - \mathbf{x_i}\|^2), \quad \gamma > 0 \tag{2}$$

Обучение SVM заключается в нахождении коэффициентов λ_i и ω_0. Для этого решается задача квадратичной оптимизации с ограничениями:

$$\sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \to \max_{\lambda}$$

$$\begin{cases} \sum_{i=1}^{m} \lambda_i y_i = 0 \\ 0 \le \lambda_i \le C; \quad C > 0, \quad i = 1,2,..., m \end{cases} \tag{3}$$

Константы γ и C в формулах (2) и (3) являются свободными параметрами модели и могут быть заданы пользователем. Отдельная SVM решает задачу бинарной классификации. В случае распознавания голосовых команд количество выходных образов соответствует числу команд в словаре предметной области, поэтому модель SVM расширяется на случай мультиклассовой классификации путем композиций автономных SVM, принимающих для каждой команды решение по типу «один против остальных».

Описание эксперимента и анализ результатов. Информационная технология распознавания голосовых команд управления мобильным роботом основывается на методе, описанном выше, и поддерживается специально разработанным на языке C# программном обеспечении: приложении ModulationSpectrograph и библиотеке NWaves. Для обучения классификаторов, распознавания голосовых команд и визуализации результатов был написан скрипт recognize.py на языке Python с использованием пакетов numpy, pandas, scikit-learn, matplotlib.

Методология проведения эксперимента предполагает следующий набор действий: 1) запись голосовой команды с микрофона в приложении ModulationSpectrograph; 2) явное указание прозвучавшей команды из списка и сохранение звукового файла: программа сохранит wav-файл в специальную директорию wav с названием, которое соответствует команде, например, «БЫСТРЕЕ_001.wav»; 3) повторение первых двух шагов для формирования набора данных достаточного размера; 4) генерация в программе ModulationSpectrograph csv-файла с векторами признаков для обучения и проверки классификаторов (при этом программа обрабатывает все звуковые файлы из директории wav, вычисляет спектр модуляции в каждом случае и формирует вектор

признаков для записи в csv-файл); 5) распознавание команд и визуализация результатов с помощью скрипта recognize.py.

В эксперименте принимал участие один диктор, произнесший каждое слово из словаря по 30 раз. Обучающая и тестовая выборка были дополнены также сигналами, автоматически искаженными шумом и эффектом реверберации. Вектор признаков для SVM представляет собой в первом эксперименте модель спектра модуляции, имеющего размерность 384 (12 частотных диапазонов, в каждом из которых расчитаны 32 спектральных коэффициента), и во втором эксперименте – модель 12-мерного спектра модуляции, в котором оставлены только коэффициенты модуляции на частоте 4 Гц.

На рис.3 приведены для сравнения примеры спектра модуляции трех различных голосовых команд.

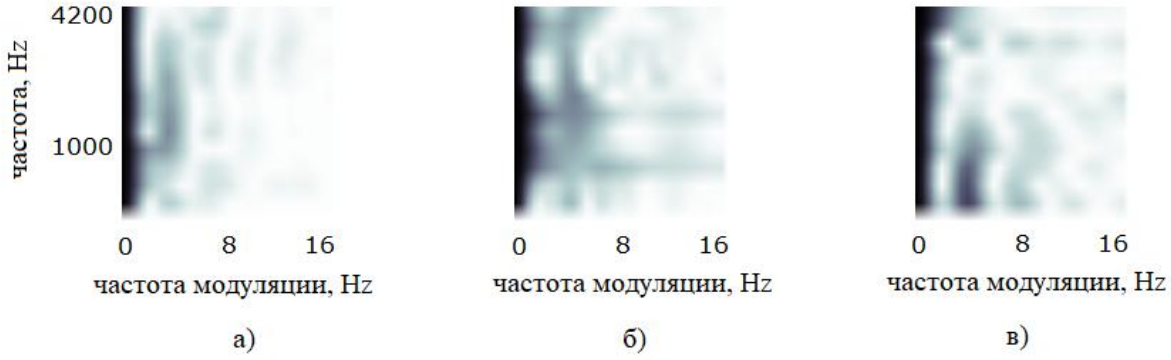

Рисунок 3 – Примеры спектра амплитудной модуляции:
а) команда «Назад»; б) команда «Опасность»; в) команда «Ищи»

На рис.4 приведены спектрограммы и спектры модуляции одной и той же команды: без обработки, с наложением белого шума уровня +6 dB и с примененным эффектом реверберации помещения.

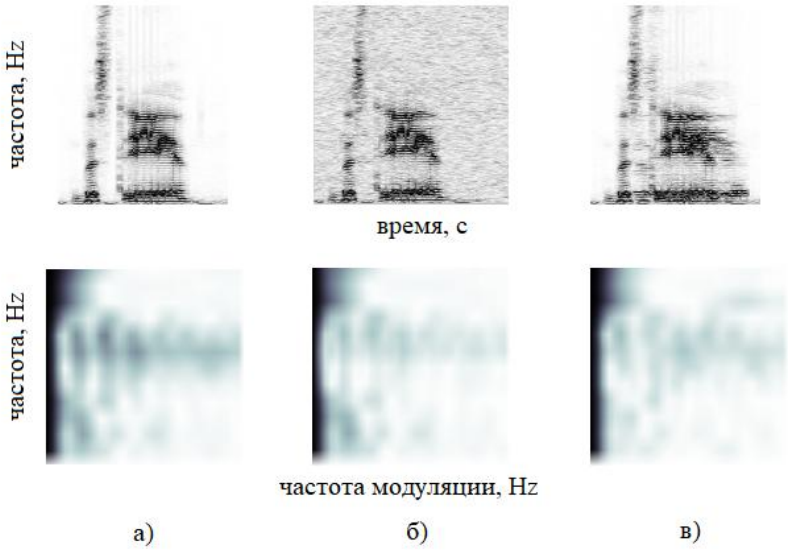

Рисунок 4 – Спектрограмма (вверху) и спектр модуляции (внизу) команды «Быстрее»:
а) без обработки; б) с наложенным шумом; в) с эффектом реверберации

Как видно из рис.4, спектр модуляции более устойчив к указанным искажениям сигнала по сравнению с кратковременным преобразованием Фурье.

Классификаторы обучались для распознавания 11 голосовых команд управления роботом: «Быстрее», «Вперед», «Ищи», «Медленнее», «Назад», «Налево», «Направо», «Опасность», «Прямо», «Разворот», «Стоп». Результаты распознавания классификатора SVM приведены в виде матрицы ошибок на рис.5.

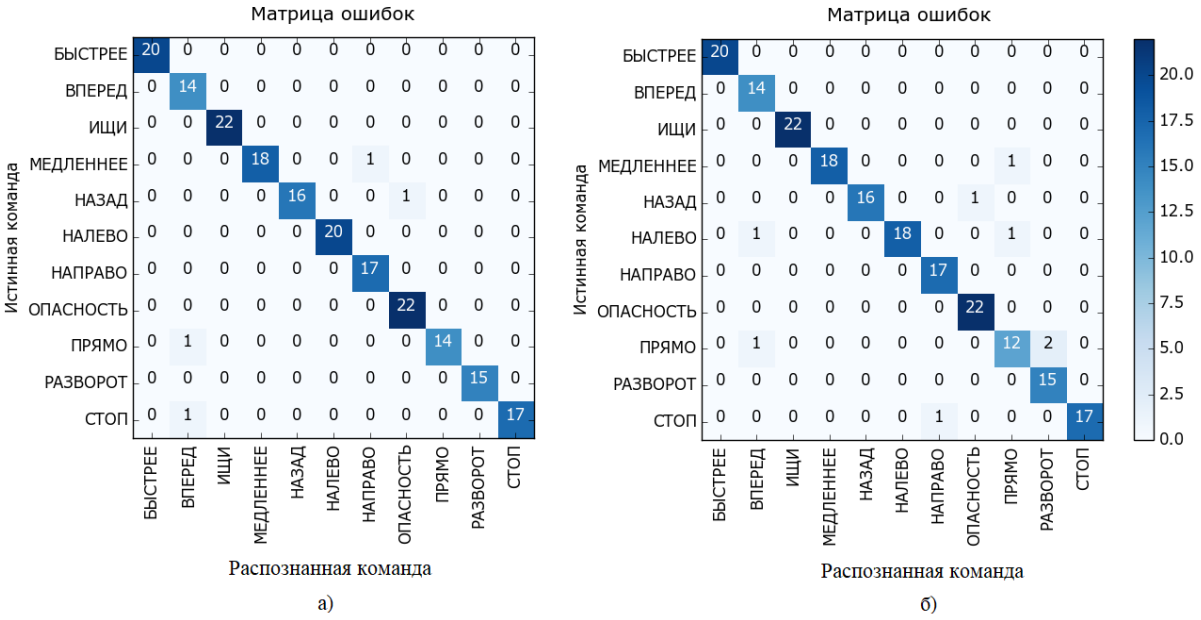

Рисунок 5 – Матрицы ошибок распознавания команд классификатором SVM:
а) в экспериментах с моделью 1; б) в экспериментах с моделью 2 (4 Гц)

Средняя точность распознавания команд расчитывалась путем кросс-валидации с размером тестовой выборки 20% от всей выборки и составила 99,2% для первой модели и 97,6% для второй модели спектра модуляции. Классификатор kNN показал похожие результаты; его средняя точность распознавания составила 97,1%.

Эксперименты производились на компьютере со следующими характеристиками: процессор Intel Core i3 с тактовой частотой 3,5 ГГц; объем оперативной памяти 16 Гб. Скорость расчета спектра модуляции оказалась равна, в среднем, 120х (120 секунд сигнала обрабатываются за 1 секунду), что дает возможность использовать алгоритм в реальном времени.

Выводы. Проведенные исследования показали перспективность подхода к распознаванию голосовых команд из небольших словарей, основанного на спектре модуляции речевого сигнала. На примере словаря робота уже на малой обучающей выборке и с использованием простейших моделей машинного обучения получена точность распознавания отдельных слов, произнесенным одним диктором, на отметке 98-99%. Визуальный анализ спектров модуляции позволяет сделать вывод о большей устойчивости данного метода параметризации речевого сигнала к стационарному шуму и эффектам реверберации помещений по сравнению с другими популярными методами, применяемыми в современных ASR-системах. Кроме того, вычислительная сложность метода относительно невелика и позволяет использовать его в реальном времени.

Дальнейшая работа связана с адаптацией алгоритма под разных дикторов за счет предварительного сдвига частоты основного тона к той, на которой система обучалась (например, методом фазового вокодера). Также заслуживают внимания исследования, связанные с варьированием формы и количества банков фильтров при вычислении

спектра модуляции, а также заменой кратковременного преобразования Фурье на первом шаге алгоритма на другие спектрально-временные преобразования (кепстр, линейное предсказание и т.д.).

СПИСОК ЛИТЕРАТУРЫ

[1] Huang X. Spoken Language Processing: A guide to theory, algorithm, and system development / X.Huang, A.Acero, H.Hon. – Prentice Hall, 2001. – 980 p.

[2] Zhang Y. Very deep convolutional networks for end-to-end speech recognition / Y. Zhang, W. Chan, N. Jaitly // *Proceedings of ICASSP 2017*. – 2017. – P. 4845-4850.

[3] Dahl G. Context-Dependent Pre-trained Deep Neural Networks for Large-Vocabulary Speech Recognition / G.E. Dahl, D. Yu, L. Deng, A. Acero // *IEEE Transactions on Audio, Speech, and Language Processing*. – Vol.20 (1). – 2012. – P. 30-42.

[4] Tchorz J. SNR estimation based on amplitude modulation analysis with applications to noise suppression / J. Tchorz, B. Kollmeier // *IEEE Transactions on Speech Audio Processing*. – Vol.11. – 2003. – P. 184-192.

[5] Азаров И.С. Алгоритм очистки речевого сигнала от сложных помех путем фильтрации в модуляционной области / И.С. Азаров, М.И. Вашкевич, Д.С. Лихачев, А.А. Петровский // *Цифровая обработка сигналов*. – №4. – 2013. – С.25-31.

[6] Paliwal K. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain / K. Paliwal, K. Wojcicki, B. Schwerin // *Speech Communication*. – Vol. 52. – 2010. – P. 450-475.

[7] Sukittanon S. Modulation frequency features for audio fingerprinting / S. Sukittanon, L. Atlas // *Proc. of the ICASSP*. – Vol.2 – 2002. – P. 1173-1176.

[8] Greenberg S. The modulation spectrogram: In pursuit of an invariant representation of speech / S. Greenberg, B.E.D. Kingsbury // *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. – 1997. – P. 1647-1650.

[9] Рашка С. Python и машинное обучение / С. Рашка. – М.: ДМК Пресс, 2017. – 418 с.

[10] Хайкин С. Нейронные сети: полный курс / С. Хайкин. – М.: Издательский дом «Вильямс», 2016. – 1104 с.

Поступила в редакцию 25.12.2017 г.

RECOGNITION OF ROBOT VOICE CONTROL COMMANDS BASED ON THE MODULATION SPECTRUM OF SPEECH SIGNAL

E.I. Cherkashin, T.V. Sharii

The article presents a robust method for recognizing robot voice control commands based on the modulation spectrum of speech signal. The algorithms are described for computing modulation spectra, calculating and applying auditory filter banks in frequency domain. The impact of noise and reverberation effects on modulation spectrum is studied. The results of experiments on the recognition of voice commands are given based on nearest neighbors and support vector machine classifiers in speaker-dependent testing environment.

Keywords: *speech signal; modulation spectrum; filter bank; classifier.*

Шарий Тимофей Вячеславович, кандидат технических наук, доцент, tsphere@mail.ru.
Черкашин Евгений Игоревич, студент, shedl33@gmail.com.

Timofei Viacheslavovich Sharii, PhD. in technical science, associate professor, tsphere@mail.ru.
Yevgenii Igorevich Cherkashin, student, shedl33@gmail.com.

Почта для кореспонденции: tsphere@mail.ru.
Телефон для контактов: (+38) 050-769-61-79.

Распознавание голосовых команд управления роботом на основе спектра модуляции речевого сигнала.
Recognition of robot voice control commands based on the modulation spectrum of speech signal.