ИНФОРМАТИКА

Вестник Сыктывкарского университета.
Серия 1: Математика. Механика. Информатика.
Выпуск 2 (31). 2019

УДК 004.522

РАЗРАБОТКА СИСТЕМЫ РАСПОЗНАВАНИЯ РЕЧИ ДЛЯ ДОМАШНЕЙ АВТОМАТИЗАЦИИ

А. В. Горьев, В. А. Устюгов

В статье описаны математические основы, необходимые для построения систем распознавания речи. Описан вариант реализации алгоритма распознавания речи, основанный на сравнении мел-частотных кепстральных коэффициентов выборок звуковых сигналов. Представлена реализация программного детектора речевой активности, позволяющего существенно снизить объем требуемых для решения задачи вычислительных ресурсов.

Ключевые слова: распознавание речи, мел-частотные коэффициенты, кепстр.

В наши дни бурно развивается сфера автоматизации, которая, безусловно, является одним из главных направлений научно-технического прогресса. Автоматизация призвана максимально упростить рабочие процессы и существенно увеличить эффективность их организации при помощи различных инструментов. Сегодня таким инструментом выступает цифровизация — интеграция цифровых технологий в физические процессы. Участвуя в скоростной гонке эффективности взаимодействия человека и машины, привычные нам диалоговый и графический интерфейсы постепенно отступают, уступая некоторые позиции речевому интерфейсу. Системы с речевым интерфейсом не только упрощают

взаимодействие человека с различными устройствами, но и открывают целый ряд различных возможностей: голосовая биометрия, синтез и анализ речи, обучение и образование, телефония и многое другое.

Речевые технологии, в частности, дошли и до домашней автоматизации, что позволяет избавиться от постоянных рутинных задач, внезапных форс-мажоров, принести удобство и комфорт в повседневный быт. Рассматривая рынок устройств для домашней автоматизации, можно выделить два класса систем: online-системы, требующие непрерывного доступа в сеть «Интернет», и offline-системы, которые могут работать в автономном режиме без интернет-соединения. В настоящее время, как правило, используются online-системы распознавания речи, однако это не всегда удобно, так как интернет-технологии носят неповсеместный характер. В таком случае на помощь приходят offline-системы распознавания, например программный комплекс с открытым исходным кодом **PocketSphinx**. Недостатком данной системы можно считать отсутствие в стартовом наборе функций возможности распознавания русскоязычной речи. Реализовать такой функционал возможно лишь путём изменения программного кода, что не под силу рядовому пользователю. Такое решение не может быть ориентировано на массовый рынок.

В работе предпринята попытка разработки легко конфигурируемой offline-системы распознавания речи, пригодной для запуска на микрокомпьютерах, широко применяющихся для целей домашней автоматизации.

1. Мел-частотная шкала

Достоверно известно, что амплитудно-частотная характеристика человеческого слухового аппарата имеет нелинейный характер [1]. В связи с этим фактом использование привычных нам физических величин, таких как амплитуда и высота звука, является непродуктивной мерой. Для решения данной проблемы были введены альтернативные единицы измерения — *Фон* и *Мел.* Мел — это эмпирически полученная едини-

ца измерения высоты звука, основанная на психофизических параметрах восприятия. Фон — логарифмическая единица для оценки уровня громкости звука, учитывающая чувствительность человеческого слуха на разных частотах.

Мел удобно применять в системах анализа речи, так как его использование учитывает ряд особенностей слухового анализатора человека, делает чувствительность алгоритмов более близкой к человеческим параметрам восприятия [2]. Перевод частоты из Гц в Мелы осуществляется с помощью выражения (1), а обратное преобразование — с помощью выражения (2).

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) = 1127 \cdot \ln\left(1 + \frac{f}{700}\right), \tag{1}$$

$$f = 700 \cdot (10^{m/2595} - 1), \tag{2}$$

где f — частота, измеряемая в герцах, m — частота в мелах. График, демонстрирующий мел-частотную шкалу, приведен на рис. 1.

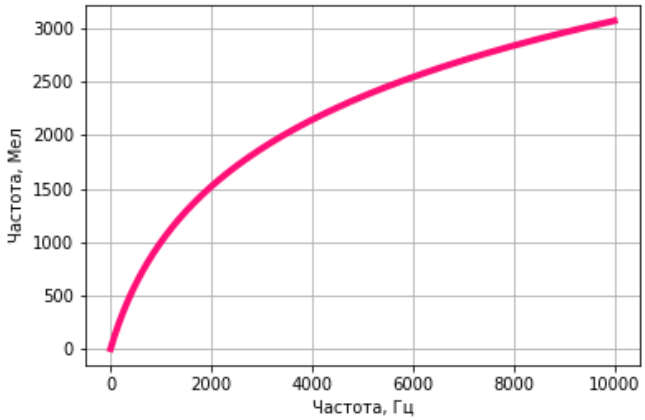

Рис. 1. Мел-частотная шкала

2. Спектр

В теории обработки сигналов под спектром понимают распределение энергии сигнала по частотам. Изучение спектра позволяет качественно оценить частотный состав сигнала [3].

Для определения спектра звукового сигнала (3) на микрокомпьютере применяется стандартный метод дискретного преобразования Фурье (ДПФ), результат которого определяется как дискретная последовательность $X(m)$ в частотной области

$$X(m) = \sum_{n=0}^{N-1} x(n) exp\left(-i\frac{2\pi nm}{N}\right), \tag{3}$$

где N — количество отсчетов сигнала.

Заметим, что ДПФ представляет собой неэффективный инструмент с точки зрения использования памяти цифровой системы и её вычислительного времени. Когда количество точек ДПФ увеличивается до тысяч значений, количество вычислительных операций крайне велико. Для устранения данной проблемы Кули и Тьюки предложили так называемое быстрое преобразование Фурье (далее — БПФ), требующее в несколько раз меньшее количество вычислительных операций, чем стандартное ДПФ. При этом результат БПФ является не приближённым к ДПФ, а в точности повторяет его. Благодаря такой экономии ресурсов вычислительной системы БПФ можно реализовать на микроконтроллерах и микрокомпьютерах.

Обычно для вычисления БПФ N выбирают степенью двойки. Это необходимо для того, чтобы $N/2$ точечное ДПФ разбить на элементарные блоки по 2 отсчёта, вычислить преобразование для них, а затем простой перестановкой знака поворотных множителей вычислить остальные слагаемые и собрать всё воедино.

3. Утечка спектра

Утечка спектра — это явление, заключающееся в «растекании спектра» по бинам преобразования (бин — интервал на шкале частот), если

частота входного сигнала в точности не равна центральной частоте одного из бинов. Например, если частотная шкала имеет шаг в n кГц, а входящий сигнал — частоту $1.5n$ кГц, то возникнет утечка. Это неизбежный эффект при выполнении ДПФ реальных последовательностей конечной длины, потому что ограничение сигнала во времени — это свертка спектра идеального бесконечного сигнала со спектром прямоугольного окна. Окном называется специальная весовая функция, которая позволяет выделить некоторую часть сигнала [4]. В частности, прямоугольное окно позволяет ограничить бесконечную последовательность.

В общем случае амплитудно-частотная характеристика ограниченной во времени косинусоиды, взвешенной прямоугольным окном, аппроксимируется функцией $sinc(x)$:

$$X(m) = \frac{N}{2} \cdot \left[\frac{\sin(\pi k - \pi m)}{\pi k - \pi m}\right] = \frac{N}{2} \cdot \mathrm{sinc}(\pi k - \pi m), \tag{4}$$

где

$$sinc(x) = \begin{cases} \dfrac{\sin(x)}{x} & , x \neq 0; \\ 1 & , x = 0. \end{cases} \tag{5}$$

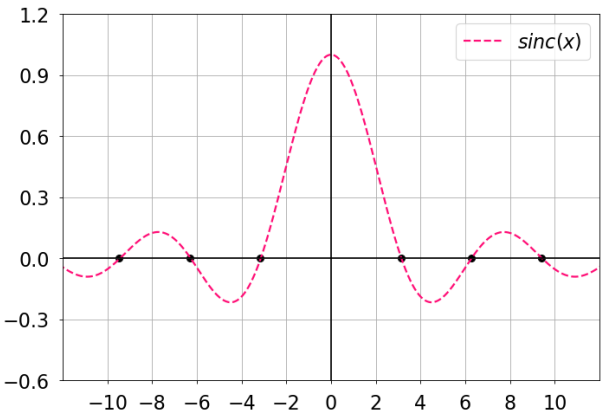

Рис. 2. График функции $sinc(x)$

Функция $sinc(x)$ есть результат преобразования Фурье прямоугольного окна. Именно резкие переходы от 0 к 1 являются причиной возникновения боковых лепестков функции $sinc(x)$. Для борьбы с утечкой, обусловленной боковыми лепестками, необходимо использовать окна, отличные от прямоугольного. Таких окон существует множество, каждое со своими достоинствами и недостатками. При выборе окна стоит искать компромисс между шириной главного лепестка, уровнем первого лепестка и скоростью убывания высоты боковых лепестков с ростом частоты.

4. Кепстр

Зачастую данных, полученных при помощи спектрального анализа сигнала, недостаточно для хорошей оценки сигнала. Одним из инструментов, который помогает повысить точность анализа, является *кепстр*, или спектр спектра. Первоначальный спектр в данном случае представляется как самостоятельный сигнал. Используя кепстры, можно получить информацию в более компактном виде, что облегчает анализ данных.

Свёрткой называется выражение вида

$$f_2(t) = \int_0^t f_1(\tau) \cdot w(t-\tau) d\tau. \tag{6}$$

Свёртка позволяет найти реакцию системы $f_2(t)$ при произвольном воздействии $f_1(t)$, если известная её импульсная характеристика $w(t)$.

Сигнал на выходе голосового тракта, работу которого можно интерпретировать как действие фильтра, можно представить в виде свёртки:

$$f(t) = s(t) \otimes w(t), \tag{7}$$

где $s(t)$ — изначальный вид сигнала, $w(t)$ — характеристика фильтра, которая определяется параметрами голосового тракта.

Переходя в частотную область, получим:

$$F(\omega) = S(\omega) \cdot W(\omega). \tag{8}$$

Путём логарифмирования этого выражения перейдем к сумме:

$$\ln[S^2(\omega) \cdot W^2(\omega)] = \ln S^2(\omega) + \ln W^2(\omega). \tag{9}$$

Теперь, используя преобразование Фурье, получим окончательное выражение для кепстра:

$$C(q) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} \ln[F(\omega)]^2 \cdot e^{i\omega q} d\omega, \tag{10}$$

где q имеет размерность времени, однако в силу проведенных преобразований это не простое время, а кепстральное. Иногда q называют «сачтота» или «кьюфренси».

5. Алгоритм получения MFCC

Мел-частотные кепстральные коэффициенты (англ. Mel-Frequency Cepstrum Coefficients) позволяют выделить акустические параметры и признаки речи. Используя шкалу перевода частоты сигнала в высоту звука в мелах, необходимо получить многомерные векторы признаков, которые в свою очередь передаются на алгоритмы сравнения [5].

Вычисление MFCC начинается с разбиения входящего речевого сигнала на отдельные перекрывающиеся интервалы — фреймы или кадры. Разбиение на фреймы совершается для уменьшения вычислительной сложности задачи, а перекрытия необходимы для сбора информации на границах фреймов, которую можно было бы потерять при последовательном их следовании. Как правило, длину фреймов выбирают от 20 до 40 миллисекунд, а перекрытие составляет 50 %.

Для дальнейшего анализа данных необходимо произвести преобразование Фурье. Прежде чем это сделать, необходимо использовать оконную функцию для каждого фрейма, чтобы уменьшить утечку спектра и

свести к минимуму разрывы сигнала. После применения оконной функции можно переходить к ДПФ, вычисляя при этом спектр мощности.

Полученные на предыдущем этапе спектральные коэффициенты накладываются на мел-частотные окна (рис. 3), которые располагаются плотнее в диапазоне низких частот и увеличивают своё расхождение при переходе к высоким частотам. Такое расположение окон в наибольшей степени соответствует восприятию звука: чем ниже частота, тем меньше различий между соседними частотами.

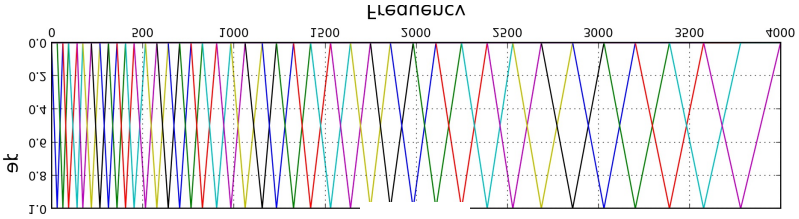

Рис. 3. Окна на мел-шкале

Полученные мел-частотные спектральные коэффициенты говорят о том, сколько энергии сигнала попало в каждое треугольное окно. Теперь осталось только получить кепстральные коэффициенты путём взятия «спектра спектра». Обычно для этого используют дискретное косинусное преобразование:

$$a_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right], \qquad k \in [0, N-1]. \tag{11}$$

6. Детектор речевой активности

Детектор речевой активности (англ. Voice Activity Detector, VAD) представляет собой программный алгоритм, позволяющий обнаружить голосовую активность во входном акустическом сигнале, отделяя при этом активную речь от фонового шума или тишины [6]. Использование такой системы позволит существенно сократить количество вычисли-

тельных операций, а также исключить возможность записи тишины в словарь.

Реализовать VAD возможно множеством различных способов. В ходе реализации настоящего проекта был разработан собственный алгоритм выделения речевых данных из общего аудиопотока. Для этого были использованы три различные характеристики, которые позволяют чётко обнаружить периоды речевой активности. Рассмотрим их подробнее.

6.1. Краткосрочная энергия Краткосрочная энергия (англ. Short-Term Energy, STE) — это обычная энергия, которая вычисляется для каждого фрейма отдельно. Как ранее было сказано, длина фреймов подбирается порядка 10−40 мс, так что такое название вполне оправданно. Вычислить краткосрочную энергию можно следующим образом [6]:

$$E = \sum_{m=0}^{N-1} s^2(m), \tag{12}$$

где $s(m)$ — мощность m-го фрейма, N — количество фреймов.

График зависимости краткосрочной энергии от номера фрейма приведен на рис. 4.

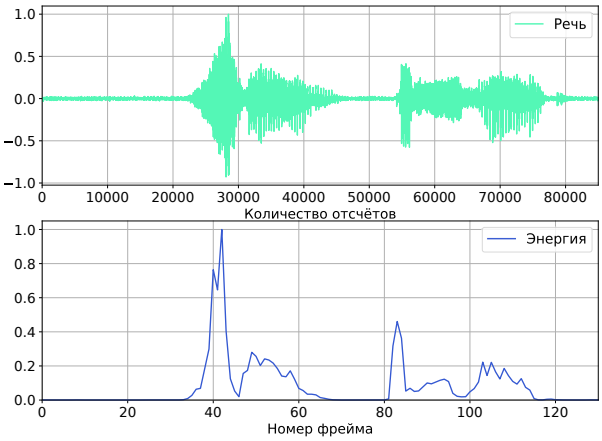

Рис. 4. Краткосрочная энергия речевого сигнала

6.2. Частота пересечения нуля Частота пересечения нуля (англ. Zero Crossing Rate, ZCR). Эта величина показывает, сколько раз в течение фрейма сигнал меняет свой знак. Количество пересечений для функции речевого сигнала будет заметно меньше, чем у функции, описывающей шум, так как речевой диапазон характеризуется преимущественно низкими частотами по сравнению с шумом [7].

Определить ZCR можно по формуле [8]:

$$Z(n) = \sum_{m=1}^{N} |\mathrm{sign}[s(m)] - \mathrm{sign}[s(m-1)]|. \tag{13}$$

На рис. 5 можно увидеть график частоты пересечения нуля. Из рисунка видно, что, начиная примерно с 40 фрейма, значение частоты пересечения нуля падает, что говорит о наличии речевой активности.

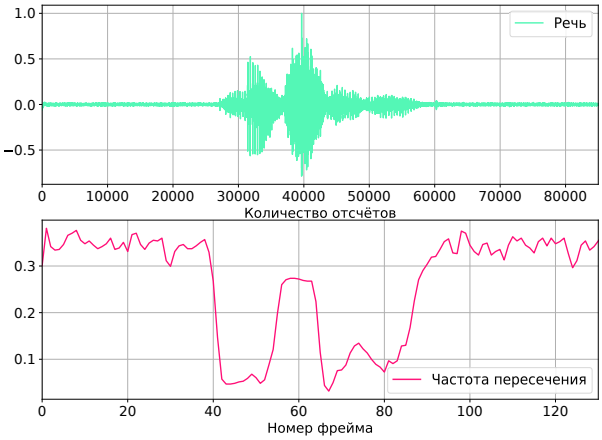

Рис. 5. Частота пересечения нуля

7. Спектральная плоскотность

Мера спектральной плоскостности (англ. Spectral Flatness Measure, SFM) — это спектральная характеристика, используемая в цифровой обработке сигналов, которая даёт хорошую количественную оценку наличия речевых фрагментов в потоке аудиоданных [6]. Вычислить эту

характеристику можно по формуле

$$F = 10\lg \cdot \frac{G}{A} = 10\lg \cdot \frac{\left(\prod_{m=0}^{N-1} s(m)\right)^{\frac{1}{N}}}{\frac{1}{N}\sum_{m=0}^{N-1} s(m)}, \tag{14}$$

где G и A есть среднее геометрическое и арифметическое соответственно.

Высокая спектральная плоскостность (приближающая к 1 без учёта шкалы Дб) говорит о том, что спектр имеет одинаковое количество мощности во всём диапазоне частот, что схоже с белым шумом. График спектра в таком случае будет выглядеть относительно плоско и гладко. Низкая спектральная плоскостность (близкая к 0 без учёта шкалы Дб) указывает на то, что мощность сигнала сосредоточена в относительно небольшой полосе частот. В таком случае график спектра будет иметь чётко выраженные пики активности. Таким образом, вычисляя спектральную плоскостность для каждого фрейма можно наблюдать хорошо заметные периоды речевой активности. График спектральной плоскостности тестового речевого сигнала показан на рис. 6.

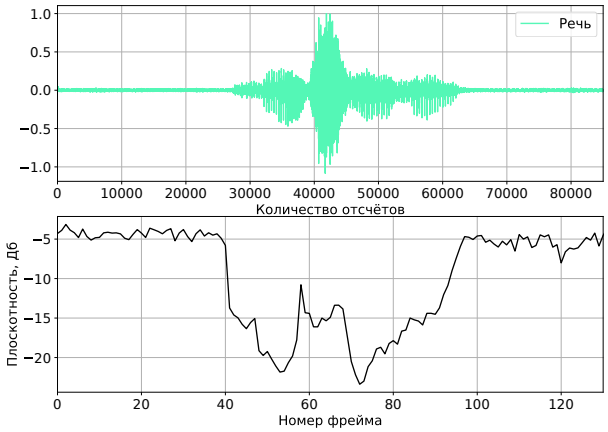

Рис. 6. Спектральная плоскостность речевого сигнала

Как видно из графиков, каждая величина помогает оценить наличие речи в потоке аудиоданных.

8. Разработанный алгоритм распознавания речи

Прежде всего для упрощения вычислительной задачи и более точного наблюдения изменений сигнала производится разбиение аудиосигнала на фреймы. Разбиение длительностью 30 мс с перекрытием 50 %. Для увеличения производительности можно распределить задачу между различными процессами, каждый из которых обрабатывает свой собственный фрейм.

Следующим шагом необходимо вычислить основные характеристики для каждого фрейма: краткосрочную энергию, частоту пересечения нуля и спектральную плоскостность. Для расчёта плоскостности следует рассчитать спектр для каждого кадра. Далее отбираются первые 30 фреймов, которые, как правило, являются тишиной. Для работы алгоритма появляется необходимость установить пороги срабатывания, которые подбираются эмпирическим путём. Это единственные значения, которые устанавливаются извне. Если значения всех трёх характеристик для какого-либо фрейма превышают пороговое значение, то фрейм считается речевым и записывается в массив. В противном случае фрейм отмечается как невокализованный и отсеивается. Полученные речевые фреймы объединяются обратно в общий сигнал, который записывается в файл формата WAV.

После всех манипуляций с сигналом необходимо составить словарь речевых образцов, с которыми будет проводиться сравнение входящего сигнала. Для проверки системы были записаны различные речевые фрагменты. Запись производилась 2 раза для каждого слова, чтобы добавить некоторую энтропию в словарь, так как невозможно повторить одно и то же слово с одинаковой интонацией, скоростью и т. д. В конечном счете длина словаря составила 20 слов. В качестве входного сигнала использовался сигнал непосредственно с микрофона, подключённого к звуковой карте персонального компьютера.

Для сравнения двух временных рядов используется алгоритм дина-

мической трансформации временной шкалы. На выходе этот алгоритм выдаёт степень подобия двух последовательностей в виде специальной переменной, значение которой записывается в массив. Чем меньше значение переменной, тем больше похожи друг на друга последовательности. Очевидно, что при сравнении двух одинаковых последовательностей, значение указанной переменной будет равно нулю. Сравнение производится со всеми доступными элементами словаря. Следующим этапом является нахождение минимального элемента массива и его индекса. Таким образом мы получим наиболее релевантное словарное слово.

9. Заключение

Итак, в ходе работы над проектом были получены следующие результаты:

- точность распознавания речи — 86 %;
- точность срабатывания детектора речевой активности — 97 %;
- время выполнения распознавания — 2–4 секунды.

Реализованный на языке программирования Python программный комплекс пригоден для запуска на микрокомпьютерах класса Raspberry Pi, широко использующихся для построения систем домашней автоматизации.

Список литературы

1. **Линдсей П., Норман Д.** Переработка информации у человека. М.: Мир, 1974. 546 с.
2. **Huang X., Acero A.** Spoken Language Processing: A Guide to Theory Algorithm, and System Development. Prentice Hall, 2001. 965 p.

3. **Lyons R. G.** Understanding Digital Signal Processing. Addison Wesley Pub. Co, 2006. 656 p.

4. **Bracewell R. N.** The Fourier Transform and its Applications. McGraw Hill, 2000. 620 p.

5. **Ganchev T., Fakotakis N.** Comparative evaluation of various MFCC implementations on the speaker verification task // *10th International Conference on Speech and Computer. Patras, Greece, 2005.*

6. **Moattar M. H., Homayounpour M. M.** A efficient real-time voice activity detection algorithm // *Laboratory for Intelligent Sound and Speech Processing (LISSP), Computer Engineering and Information Technology Dept., Amirkabir University of Technology, Tehran, Iran. 24 августа 2009.*

7. **Nandhini S., Shenbagavalli A.** Voiced/Unvoiced Detection using Short Term Processing // *International Journal of Computer Applications (0975–8887). 2014.*

8. **Bachu R., Kopparthi S., Adapa B., Barkana B.** Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy // *AdvancedTechniques in Computing Sciences and Software Engineering, 2010. Pp. 279–282.*

Summary

Gorev A. V., Ustyugov V. A. Development of the speech recognition systems for home automation

The article describes the mathematical foundations necessary for the speech recognition systems development. An embodiment of a speech recognition algorithm based on a comparison of the mel-frequency cepstral coefficients of audio signal samples is described. The implementation of the

software speech activity detector is presented.
Keywords: speech recognition, mel-frequency coefficients, kepstrum.

References

1. **Lindsei P., Norman D.** *Pererabotka informatsii u cheloveka* (Humans information processing), Mir, 1974, 546 p.

2. **Huang X., Acero A.** *Spoken Language Processing: A Guide to Theory Algorithm, and System Development*, Prentice Hall, 2001, 965 p.

3. **Lyons R. G.** *Understanding Digital Signal Processing*, Addison Wesley Pub. Co, 2006, 656 p.

4. **Bracewell R. N.** *The Fourier Transform and its Applications*, McGraw Hill, 2000, 620 p.

5. **Ganchev T., Fakotakis N.** Comparative evaluation of various MFCC implementations on the speaker verification task, *10th International Conference on Speech and Computer*, Patras, Greece, 2005.

6. **Moattar M. H., Homayounpour M. M.** A efficient real-time voice activity detection algorithm, *Laboratory for Intelligent Sound and Speech Processing (LISSP), Computer Engineering and Information Technology Dept., Amirkabir University of Technology,* Tehran, Iran, 24.10.2009.

7. **Nandhini S., Shenbagavalli A.** Voiced/Unvoiced Detection using Short Term Processing, *International Journal of Computer Applications*, 0975–8887, 2014.

8. **Bachu R., Kopparthi S., Adapa B., Barkana B.** Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy,

AdvancedTechniques in Computing Sciences and Software Engineering, 2010, pp. 279–282.

Для цитирования: Горьев А. В., Устюгов В. А. Разработка системы распознавания речи для домашней автоматизации // *Вестник Сыктывкарского университета. Сер. 1: Математика. Механика. Информатика. 2019. Вып. 2 (31). С. 26–41.*

For citation: Gorev A. V., Ustyugov V. A. Development of the speech recognition systems for home automation, *Bulletin of Syktyvkar University. Series 1: Mathematics. Mechanics. Informatics*, 2019, 2 (31), pp. 26–41.

СГУ им. Питирима Сорокина *Поступила 09.08.2019*