УДК 004.934

АВТОМАТИЗИРОВАННАЯ СИСТЕМА ДИКТОРОНЕЗАВИСИМОГО ГОЛОСОВОГО РУССКОЯЗЫЧНОГО УПРАВЛЕНИЯ ОПЕРАЦИОННОЙ СИСТЕМОЙ WINDOWS

Кравченко К. В., Дьяченко Р. А.

ФГБОУ ВПО «Кубанский государственный технологический университет», Краснодар, Россия (350072, Россия, Краснодарский край, г. Краснодар, ул. Московская, д. 2), e-mail:adm@kgtu.kuban.ru

В данном исследовании рассматриваются методы и способы организации распознавания русскоязычных речевых команд. Целью исследования является создание дикторонезависимой системы для распознавания русскоязычных голосовых команд. Среди методов были рассмотрены такие алгоритмы как – MFCC (Mel-Frequency Cepstral Coefficients), LPCC (Linear Predictive Cepstral Coefficients), PLP (Perceptual Linear Predictive). Также не отвергаются от рассмотрения готовые программные решения и модули от ведущих компаний, имеющих опыт работы в сфере речевых технологий. Предполагалось использовать один из алгоритмов, который себя наиболее зарекомендует по качеству получения признаков голоса, в синтезе с нейронной сетью для обучения системы. В ходе анализа и тестирования не один из алгоритмов не зарекомендовал себя для создания дикторонезависимой системы. В итоге был выбран модуль распознавания SpeechAPI от корпорации Microsoft. Единственным недостатком модуля является отсутствие поддержки русского языка, но данная проблема решена созданием программного класса, преобразующего русские машинописные слова в англоязычные транскрипции, поддерживаемые модулем распознавания. В качестве демонстрации возможной системы, использующий данный метод, была разработана автоматизированная система русскоязычного дикторонезависимого голосового управления операционной системой Microsoft Windows.

Ключевые слова: автоматизированная система, распознавание, диктор, преобразование Фурье, операционная система, речь, голос, команда, детектор голоса, модуль распознавания, русскоязычная.

AUTOMATED VOICE SYSTEM INDEPENDENT OF THE SPEAKER RUSSIAN-LANGUAGE CONTROL FOR WINDOWS

Kravchenko K. V., Dyachenko R. A.

Kuban State Technological University, Krasnodar, Russia (350072, Krasnodar, street. Moscow, 2), e-mail:adm@kgtu.kuban.ru

This study examines the methods and ways of organizing the recognition of Russian speech commands. The aim of the study is to establish a system for recognizing speaker independent Russian voice commands. Among the methods considered algorithms such as MFCC (Mel-Frequency Cepstral Coefficients), LPCC (Linear Predictive Cepstral Coefficients), PLP (Perceptual Linear Predictive). It is also not rejected from the consideration of ready-made software solutions and modules from leading companies with expertise in the field of speech technology. It was supposed to use one of the algorithms, which itself most prove obtain evidence on the quality of voice in the synthesis of a neural network learning system. During the analysis and testing is not one of the algorithms is not proven itself to create speaker independent system. In the end, was selected recognition module Speech API from Microsoft. The only drawback is the lack of module support for the Russian language, but the problem is solved by creating a class of software that converts typed words in Russian English-language transcription of the module supports recognition. To demonstrate the potential of using this method has been developed automated system for Russian-speaking speaker independent voice control of the operating system Microsoft Windows.

key words: automated system, recognition, speaker, Fourier transform, operating system, speech, voice, command, detection voice, recognition engine, russian.

В настоящее время вопросы проектирования и создания систем машинного синтеза и распознавания речи являются актуальной проблемой. Совместное использование таких систем является фундаментом полноценного голосового интерфейса, спектр применения которого на практике чрезвычайно широк. Исследованиями в области речевого интерфейса занимаются многие ученые, а разработки ведут крупнейшие компьютерные организации, в том числе, Microsoft, Intel и IBM.

Применение системы дикторонезависимого распознавания речевых русскоязычных команд может иметь широкий спектр и реализовываться во многих сферах общества и промышленности.

В этих условиях проблема создания отечественной дикторонезависимой системы голосового управления операционной системой является актуальной. Актуальность работы также подтверждает тот факт, что при помощи описанной выше информационной системы возможно голосовое управление персональным компьютером, оснащенным современной операционной системой, людьми с ограниченными возможностями (ограниченные возможности движения), инвалидами.

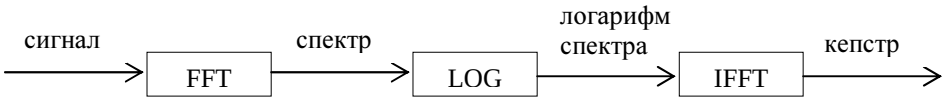
Основой создания дикторонезависимых голосовых систем управления являются алгоритмы и методы распознавания речи. Наиболее распространенными и эффективными являются следующие методы:

– MFCC (Mel-FrequencyCepstralCoefficients) – метод кепстральных коэффициентов на шкале мел. Этот метод не уступает методу PLP, но при этом является более простым в реализации. Заключается в вычислении коэффициентов спектра Фурье, накладывания на полученный спектр набора фильтров шкалы мел, выполнения логарифмирования измененного спектра и реализации дискретного косинусного преобразования.

– LPCC (LinearPredictiveCepstralCoefficients) – метод кепстральных коэффициентов линейного предсказания. Основывается на вычислении коэффициентов авторегрессионной модели для каждого фрейма аудио сигнала. После получения всех параметров модели вычисляются кепстральные LPCC – коэффициенты по рекурсивной функции.

– PLP (PerceptualLinearPredictive) – метод коэффициентов перцептивного линейного предсказания. Метод отличается от метода LPCC тем, что учитываются особенности восприятия различных частот человеком – перед вычислением параметров авторегрессионной модели сигнал проходит определённую предобработку. Вычисленный мгновенный спектр Фурье преобразуется в спектр на шкале барков, после чего выполняется операция свертки маскирующих кривых критических полос с полученным спектром для получения эффекта маскировки частоты. Далее производится аппроксимация кривой громкости и кепстральная обработка.

Анализ рассмотренных методов показал, что большинство из них сосредотачивают усилия на извлечении частотной характеристики речевого тракта человека, отбрасывая при этом характеристики сигнала возбуждения. Это объяснено тем, что коэффициенты первой модели обеспечивают лучшую разделимость звуков. Для отделения сигнала возбуждения от сигнала речевого тракта прибегают к кепстральному анализу. Схематически этот метод представлен на схеме [5]:



где FFT – блок быстрого преобразования Фурье сигнала (БПФ), LOG – блок логарифмирования спектра, IFFT – блок обратного быстрого преобразования Фурье (ОБПФ).

Для описанных выше методов первые два этапа цифровой обработки сигнала одинаковы (предварительное усиление и сегментация на фреймы).

На первом этапе к сигналу применяется БИХ-фильтр, который вычисляется по формуле:

$$H_{pre}(z) = 1 + a_{pre} z^{-1}.$$

Данный фильтр позволяет «усилить» высокочастотную область спектра сигнала (нужно для выравнивания спектра, т.к. вокализованные участки речи характеризуются резко спадающим спектром, а также по причине того, что человеком лучше воспринимаются частоты выше 1кГц). Значение коэффициента обычно выбирается из промежутка [-1.0, -0.4].

На втором этапе речевой сигнал разбивается во времени на перекрывающиеся короткие промежутки (фреймы), в которых проводится «мгновенный» кепстральный анализ. Обычно продолжительность фрейма составляет от 20 мс до 40 мс. Полагается, что на этих участках речевой сигнал можно считать квазистационарным. К фрейму применяется оконная функция Хемминга, которая имеет следующий вид:

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N}).$$

Алгоритм LPCC [1] начинается с вычисления p коэффициентов $\{a_k\}_{k=1}^{p}$ авторегрессионной модели для каждого фрейма на основе модели S, которая имеет вид:

$$\hat{S}(z) = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}}.$$

После нахождения всех параметров модели вычисляются кепстральные LPCC-коэффициенты по рекурсивной функции, которая выглядит следующим образом:

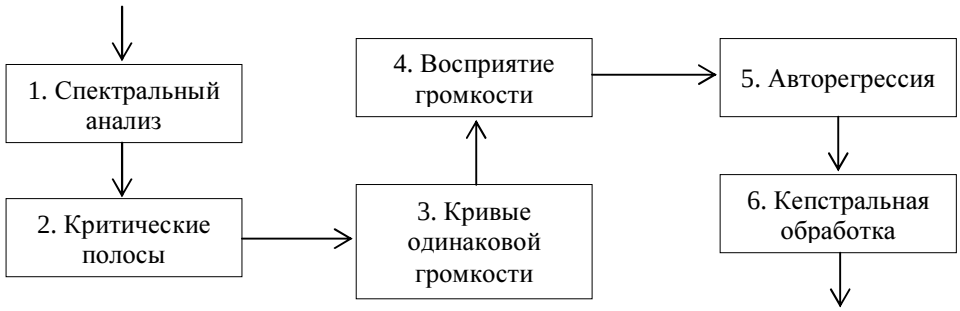
$$c(n) = \begin{cases} 0 & \\ \log_e(A) & n < 0 \\ a_n + \sum_{k=1}^{n-1} (\frac{k}{n}) c(k) a_{n-k} & n = 0 \\ \sum_{k=n-p}^{n-1} (\frac{k}{n}) c(k) a_{n-k} & 0 < n < p \\ & n > p \end{cases}.$$

На основе конечного числа коэффициентов линейного предсказания может быть получено бесконечное число LPCC-коэффициентов. Экспериментально установлено, что 12–

20 коэффициентов достаточно для формирования оптимального для данного метода вектора признаков.

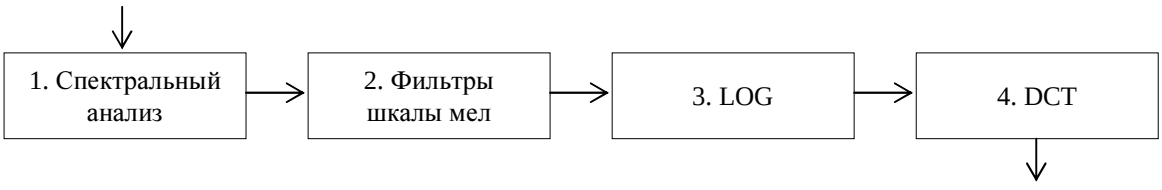
Алгоритм PLP [3] отличается от предыдущего тем, что учитывает особенности восприятия различных частот человеком – перед вычислением параметров авторегрессионной модели сигнал проходит определенную предобработку. Алгоритм схематически представлен на схеме ниже.

В блоке 1 вычисляется мгновенный спектр Фурье в текущем фрейме. В блоке 2 спектр Фурье преобразуется в спектр на шкале барков, после чего выполняется операция свертки маскирующих кривых критических полос, с полученным спектром для получения эффекта маскировки частоты. В блоке 3 к данным применяется функция кривой одинаковой громкости для аппроксимации уровня чувствительности человека к слышимому звуку 40 дБ. В блоке 4, исходя из закона восприятия громкости звука человеком, из спектральных коэффициентов извлекается кубический корень. Блоки 5 и 6 вычисляют значение выражений и соответственно.



Преимуществом метода PLP по сравнению с LPCC является то, что он позволяет подавить информацию, связанную с индивидуальными характеристиками диктора, путем выбора подходящего порядка модели. Тем не менее, данный метод более чувствителен к частоте основного тона.

Алгоритм MFCC [4] не уступает в эффективности алгоритму PLP, при этом является гораздо более простым в реализации. Алгоритм схематически представлен ниже:



В блоке 1 вычисляются коэффициенты спектра Фурье.

В блоке 2 на вычисленный спектр накладывается набор из M фильтров шкалы мел (обычно M=20 или M=24) по формуле:

$$x_i = \sum_{k=0}^{N-1} \left| X_k \right| H_i(f_k),\ i = 1..M\ .$$

Фильтр шкалы мел H имеет треугольный вид:

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

В формуле (рис. 10) значения f_{c_i} рассчитываются, исходя из центральных мел-частот, по формуле: $f_{c_i}=700(e^{f_{c_i}/1127}-1)$.

В блоке 3 выполняется логарифмирование измененного спектра по формуле:

$$x_i=\log(x_i),\ i=1..M\ .$$

Благодаря логарифмированию достигается эффективное сжатие пространства признаков и преимущества гомоморфной обработки. Однако логарифм малых чисел стремится к минусу бесконечности. Чтобы обойти этот эффект, можно применить метод маскировки (логарифм от значения и его смещения) либо заменить логарифм кубическим корнем (и то, и другое приводит к снижению качества распознавания).

В блоке 4 производится дискретное косинусное преобразование по формуле:

$$c_j=\sum_{i=1}^{M}x_i\cos(j\cdot(i-0.5)\cdot\frac{\pi}{M}),\ j=1..J\ .$$

Обычно число MFCC-коэффициентов J для формирования вектора признаков выбирают равным 12. Наиболее релевантная информация содержится в первых 6 коэффициентах. Важность включения остальных коэффициентов определяется конкретным случаем и диктором.

Сравнительный анализ показал, что качественно методы, основанные на кепстральной обработке сигнала, практически не отличаются (небольшие расхождения в показателях вызваны, скорее, спецификой речевой базы). Поэтому выбор метода остается за предпочтениями исследователя. В таблице 1 приведен также эффективный метод, не основанный на кепстральном анализе, – анализ спектра модуляции (MSG). Видно, что данный метод уступает любому из «кепстральных» методов.

Анализ существующих методов для обработки полученных данных параметров речи (кепстра) показал, что наиболее эффективными являются следующие методы:

– DTW (DynamicTimeWarping) – алгоритм динамического искажения времени. Представляет собой технику эффективного выравнивания временных рядов [2].

– Нейросеть типа персептрон с одним скрытым слоем – необходима для обработки параметров аудио сигнала (речевой команды) и обучения системы автоматического распознавания.

– СММ (Скрытой Марковской модели) – статистическая модель, имитирующая работу процесса, похожего на марковский процесс с неизвестными параметрами, и задачей ставится разгадывание неизвестных параметров на основе наблюдаемых.

Все методы по обработке аудио сигналов могут быть реализованы в системах распознавания речевых команд совокупно с методами обработки полученных данных, однако уровень распознавания русскоязычных речевых команд не соответствует уровню распознавания англоязычных речевых команд и находился ниже. Данный факт объясняется особенностями произношения фонем в русском языке.

Эффективным решением проблемы русскоязычного дикторонезависимого распознавания является разработанный метод, который автоматически перестраивает русские слова, вводимые пользователем в формат английских транскрипций, которые могут быть использованы в речевой базе распознавания команд.

В качестве модуля распознавания можно выбрать любой из доступных англоязычных модулей распознавания (например,MicrosoftSpeechAPI, который не имеет поддержки русского языка).

Для сравнения русских слов в информационной системе был создан объектно-ориентированный класс для анализа и сравнения слов по определённому алгоритму, не зависящему от длины слов.

Разработанная на этой основе система автоматического дикторонезависимого распознавания русскоязычных речевых команд была реализована в виде программного обеспечения «VoiceActivated», с возможностью управления основными WinAPI функциями операционной системы Windows.

Система автоматически создаёт все необходимые базы для работы и предоставляет возможность пользователю задавать речевые команды без произношения слова вслух и обучения системы.

Программное обеспечение не имеет основного интерфейса, но имеет поддержку «виджетов» для отображения действий распознавания в реальном времени.

Виджеты можно разрабатывать на любом языке программирования, но программное обеспечение виджета должно иметь несколько свойств:

– способность принимать параметры без перезагрузки;

– способность отправлять параметры основной программе.

Настройки распознавания дают возможность регулировать следующие параметры:

– автозапуск вместе с Windows;

– введение лога событий и ошибок и сохранение его в файл;

– определение пути хранения лога ошибок и событий;

– управление виджетами;

Тонкие настройки системы подразумевают возможность управления шумодавлением, временем задержки распознавания и снимания результатов распознавания и процент совпадения нескольких результатов, при котором вызывается окно выбора пользователя.

Настройки распознавания позволяют пользователю редактировать речевую базу без особых знаний системы, выбор API функции для выполнения и регулировки работы модуля распознавания.

Настройки обучения позволяют отключать или включать подсистему обучения и выбрать уровень интенсивности. Вызов подсистемы обучения осуществляется только в те моменты, когда модуль распознавания получил несколько вариантов распознавания и они схожи по числу процентов.

Программное обеспечение и виджеты выполнены в интегрированной среде разработки программного обеспечения Microsoft Visual Studio, с поддержкой технологии Windows Forms и WPF, с помощью Visual C#, являющейся реализацией языка C#, и предназначены для демонстрации основных возможностей речевого модуля MicrosoftSpeechAPI.

В папке исполняемого файла программы находится исполняемый файл, конечный файл XML речевой базы команд, папка Widgets, где хранятся для работы системы виджеты.

Основные функции и методы, используемые в программном продукте, выполнены в отдельном классе по правилам объектно-ориентированного программирования.

В результате исследования были исследованы и проанализированы, опробованы на практике основные, являющиеся предположительно на момент исследования успешными методы распознавания русскоязычных речевых команд.

Была разработана система автоматического дикторонезависимого распознавания русскоязычных речевых команд для управления операционной системой.

Также были разработаны: подсистема обучения, класс преобразования русских слов в английские фонемы с учётом особенностей произношения русскоязычных пользователей, метод детекции голоса (VAD), класс для поддержки виджетов и класс для сравнения в процентном соотношении русских слов.

Данное программное решение может эффективно применяться при необходимости бесконтактного управления операционной системой, для подачи необходимых пользователю команд посредством голоса на русском языке.

Список литературы

1. Huang X, Acero A., Hon H. Spoken Language Processing: A guide to theory, algorithm, and system development.Prentice Hall, 2001.

2. Зубань Ю. А. Система распознавания голосовых команд / Ю. А. Зубань, И. В. Скляров // Вісник Сумського державного університету. Серія Технічнінауки. – 2007. – № 1. – С. 178-182.

3. Hermansky H. Perceptual Linear Predictive (PLP) Analysis of Speech. The Journal of the Acoustical Society of America, 87 (4): 1738-1752, 1990.

4. Zheng F., Zhang G., Song Z.. Comparison Of Different Implementations Of MFCC. J. Computer Science & Technology, 16(6): 582-589, 2001.

5. Вісник донецького національного університету. Сер. А: Природничі науки, 2008. Вип. 2, С. 536-540.

Рецензенты:

Ключко Владимир Игнатьевич, доктор технических наук, профессор, заведующий кафедрой ВТиАСУ, ФГБОУ Кубанский государственный технологический университет (Министерство образования и науки РФ), г. Краснодар.

Степанов Владимир Васильевич, доктор технических наук, профессор кафедры Информатики, ФГБОУ Кубанский государственный технологический университет (Министерство образования и науки РФ), г. Краснодар.