OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

**УДК 004.822:514**

# ПРОТОТИП ПЛАТФОРМЫ АНАЛИЗА РЕЧИ НА ТАТАРСКОМ ЯЗЫКЕ

Хусаинов А.Ф., Сулейманов Д.Ш.

*НИИ «Прикладная семиотика» Академии наук Республики Татарстан,*
*Казанский (Приволжский) федеральный университет,*
*г.Казань, Россия*

**khusainov.aidar@gmail.com**

**dvdt.slt@gmail.com**

В работе приводится подход к созданию платформы анализа речи на татарском языке. В контексте данной платформы описываются основные модули, такие как модуль распознавания речевых команд, идентификации языка и распознавания фонем. Основная идея работы заключается в исследовании потенциала улучшения качества работы отдельных систем анализа речи за счет применения кросс-модульного взаимодействия. Предлагаемая модель платформы может быть эффективно использована при создании комплексных систем анализа речи.

**Ключевые слова:** платформа анализа речи, кросс-модульное взаимодействие систем, татарский язык



## Введение

Стремительное развитие современных электронных устройств привело к осознанию ограниченности существующих инструментов человеко-машинного взаимодействия, таких как клавиатура и мышь. Одним из путей преодоления этого ограничения является использование речи при взаимодействии с компьютером и другими устройствами. Данный подход включает в себя работу по решению множества задач, относящихся к таким направлениям анализа речи, как распознавание, синтез речи, идентификация диктора, языка диктора и т.д. Однако большинство исследований ученых сильно специализировано и направлено на разработки отдельных частей речевых технологий. Данный факт может быть объяснен как экономическими причинами, так и технологической сложностью стоящих задач. С одной стороны, такие исследования позволяют глубже изучать предметную область, а с другой, в таком случае зачастую уделяется недостаточно внимания вопросам разработки комплексных архитектурных решений, способных решать весь комплекс задач речевых технологий.

Основываясь на указанных рассуждениях, было решено исследовать возможную архитектуру комплексной системы анализа речи и её использование в контексте татарского языка. Создание такой архитектуры позволит, во-первых, увеличить эффективность работы всех модулей системы за счет синергетического эффекта, а во-вторых, предоставит готовый инструмент разработки речевых технологий для языков, для которых на данный момент не создано качественных систем анализа речи.

Структура данной работы следующая: в разделе 1 работы обсуждаются основные вопросы, касающиеся структуры модулей распознавания речи, идентификации диктора и языка, а также отмечены основные принципы построения комплексной системы анализа речи. В разделе 2 описывается возможность применения предложенной архитектуры системы к анализу татарского языка.

## 1. Архитектура системы

### 1.1. Общие положения

Область речевых технологий включает в себя множество направлений, основные из которых могут быть определены следующим образом:

- распознавание, синтез речи;
- идентификация языка;
- идентификация диктора;
- диаризация;
- определение характеристик диктора;
- идентификация характеристик канала связи.

Синергетический эффект, которого можно достичь при совместной работе различных модулей

анализа речи, можно представить состоящим из 3 уровней:

1. использование общих компонент (вычисление признаков речи, математические преобразования);
2. использование в качестве общих блоков целых модулей анализа речи, таких как модуль определения характеристик шума, распознавания фонем;
3. обмен информацией между различными модулями.

Основываясь на приведенной классификации, отметим, что использование первых двух уровней синергетического эффекта может позволить ускорить процесс разработки программных средств за счет использования общих сущностей в различных ситуациях. В то же время 3 уровень даёт возможность улучшить качество работы отдельных модулей предлагаемой платформы.

Рассмотрим данную точку зрения на примере системы анализа речи. Для простоты будем учитывать только три составных модуля системы: распознавание речи, идентификация языка и диктора. Приведём описание структур модулей в подразделах 1.2-1.4 и описание архитектуры системы в подразделе 1.5.

## 1.2. Распознавание речи

В качестве структуры системы распознавания речи будем использовать общераспространённую схему работы подобных систем, представленную на рисунке 1. Поступающий на вход системы речевой фрагмент подаётся в блок выделения значимых характеристик речи «Feature extraction», вычисленные на первом этапе характеристики используются декодером для получения результата распознавания.

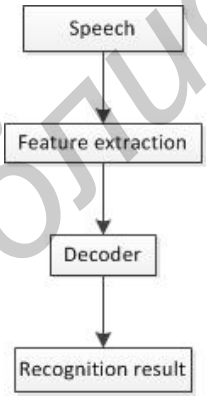


Рисунок 1 – Структура модуля распознавания речи

Для проведения дальнейшего анализа добавим в данную схему основные зависимости системы. Как видно из представленной на рисунке 2 измененной схемы, основными зависимостями блока «Feature extraction» являются подсистема предобработки, включающие в себя различные фильтры, а также подсистема вычисления векторов признаков речи. Декодер использует в своей работе такие подсистемы как модуль обнаружения речевой активности, акустические и языковые особенности.

Здесь и далее используются следующие аббревиатуры:

- MFCC (mel-frequency cepstral coefficients);
- LPC (linear predictive coding);
- PLP (perceptual linear predictive);
- ANN (artificial neural networks);
- HMM (hidden Markov models);
- DTW (dynamic time warping).

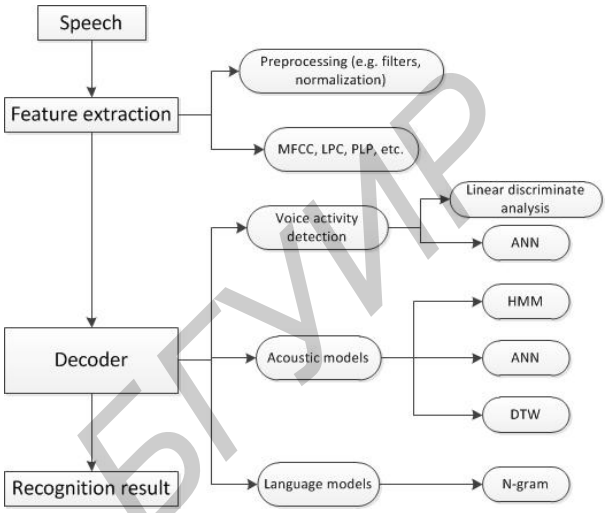


Рисунок 2 –Модуль распознавания речи с зависимостями

## 1.3. Идентификация языка

Исследования в области идентификации языка насчитывают более 30 лет, за которые учеными было разработано множество подходов к решению данной задачи. Однако, для простоты, остановимся лишь на трёх основных подходах: PRLM (phone recognition followed by language modeling), parallel PRLM и PPR (parallel phone recognition).

Как следует из названия подхода, метод PRLM состоит из двух основных этапов: распознавание фонем и моделирования языков (как правило, с помощью моделей n-gram). Структура работы метода PRLM представлена на рисунке 3.

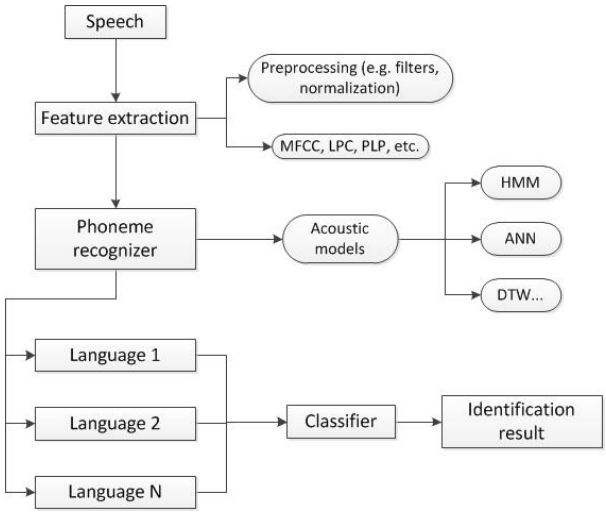


Рисунок 3 – PRLM-подход идентификации языка

При таком подходе сообщения из обучающей части речевой базы подаются на вход системы распознавания фонем, построенной для одного языка (блок «Phoneme recognizer»). На основе полученных последовательностей фонем строятся модели языков, которые в дальнейшем используются в качестве языковых моделей (блоки «Language 1», «Language N»). В процессе распознавания речевой фрагмент также разбивается на последовательность фонем, после чего вычисляются вероятности принадлежности данной последовательности каждой из языковых моделей. С помощью блока классификации «Classifier» определяются вероятности принадлежности произнесённой фразы каждой из доступных языковых моделей. Языковая модель с максимальной вероятностью выбирается в качестве результата идентификации.

Главным и единственным отличием описанного PLRM от parallel PRLM-подхода является использование одновременно нескольких систем распознавания фонем, настроенных на разные языки. Основная идея данной модификации состоит в том, чтобы повысить вероятность того, что максимальное количество произнесенных звуков было корректно распознано на основе языков, использованных для построения систем распознавания фонем.

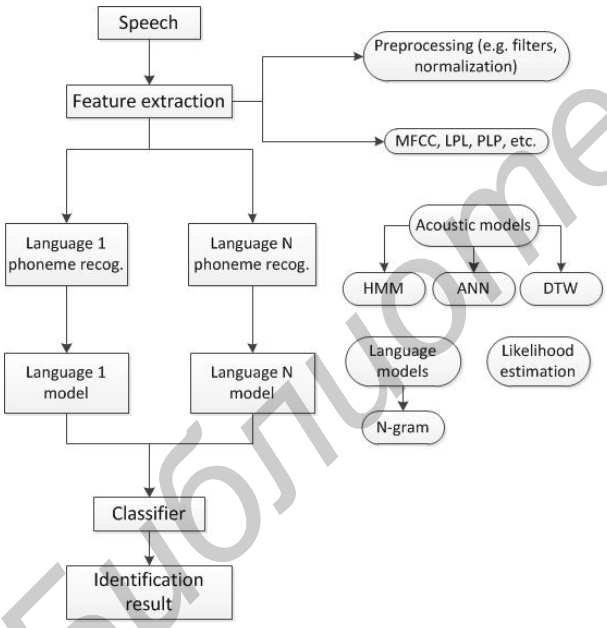


Рисунок 4 – PPR-подход идентификации языка

Если предположить наличие фонетически размеченных речевых корпусов для всех языков, для которых требуется производить идентификацию, то становится возможными использование PPR-подхода. Схема работы данного подхода с учётом имеющихся зависимостей изображена на рисунке 4. Основной особенностью этого метода является использование информации о закономерностях следования фонем в каждом языке не после этапа распознавания, а во время него. Концептуально это позволит отфильтровать не встречающиеся в языке последовательности фонем, что позволит без искажений рассчитывать результаты идентификации. В рамках приведённой схемы работы алгоритма это выражается в использовании элементов «Language models»-«n-gramms» на этапе работы блоков «Language phoneme recognition».

## 1.4. Идентификация диктора

Основная цель данного модуля - определить человека, произносящего фразу, на основе некоторых значимых особенностей речи. Условно задачу можно разбить на два подкласса: текстозависимую (с наличием ключевой фразы) и текстонезависимую. Структура модуля идентификации диктора представлена на рисунке 5.

Здесь и далее используются следующие аббревиатуры: GMM (Gaussian mixture models), VQ (vector quantization).

На этапе обучения моделей «Speaker model 1», «Speaker model N» используются, в зависимости от наличия ключевой фразы, такие методы, как скрытые Марковские модели (HMM), методы динамического программирования (например, DTW), гауссовские смеси (GMM) и т.д. Полученные модели дикторов используются на этапе распознавания для расчета вероятностей произнесения фразы каждым конкретным диктором, после чего блок «Classifier» определяет модель, которой соответствует максимальная вероятность. и которая Диктор, соответствующий данной модели, и будет являться результатом идентификации.

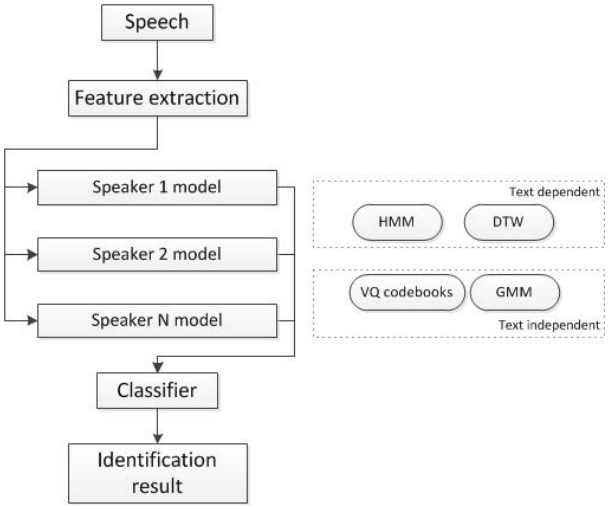


Рисунок 5 – Структура модуля идентификации диктора

## 1.5. Комплексная система анализа речи

Основываясь на предложенной идее о 3 уровнях синергетического эффекта и описанных схемах работы модулей анализа речи, перейдем к построению архитектуры комплексной системы анализа речи. Основными характеристиками разрабатываемой архитектуры должны стать:

- модульный дизайн – возможность независимой разработки различных модулей;

- гибкость – архитектура должна позволять использовать и сравнивать различные варианты реализации отдельных модулей;
- расширяемость – возможность легкого добавления новых модулей, расширяющих функционал системы;
- настраивомость системы на разные языки – каркас системы должен состоять из языко-независимых блоков, однако добавление специфических для конкретного языка элементов должно увеличивать качество работы комплекса.

**1.5.1.** *1 уровень синергетического эффекта*

Как видно из структур модулей, представленных на рисунках 2-5, большинство методов используется больше чем в одном модуле (рисунок 6). Следовательно, выделение данных элементов в отдельный, вспомогательный, модуль («Assistant tools») может позволить ускорить процесс разработки системы.

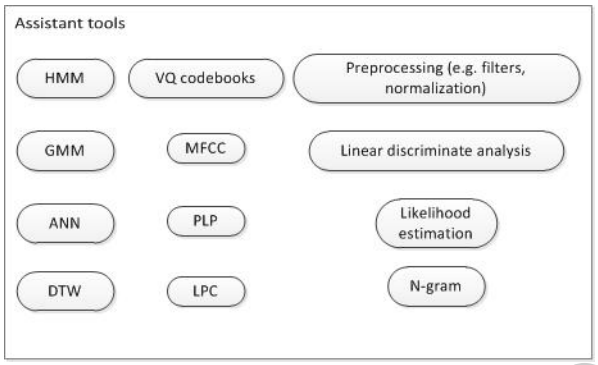


Рисунок 6 – 1 уровень синергетического эффекта

**1.5.2.** *2 уровень синергетического эффекта*

Как было отмечено выше, подход использования общих компонент может быть реализован не только на уровне реализации отдельных инструментов, но и целых модулей, например, модуля распознавания фонем. Стоит заметить, что в случае речевых технологий речь может идти о совместном использовании, в том числе и «неречевых» блоков: модуля анализа движения губ, дополнительной фиксации шумовых компонент.

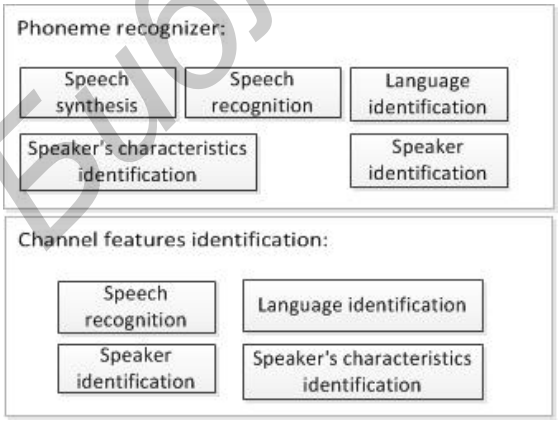


Рисунок 7 – 2 уровень синергетического эффекта в речевых технологиях

Как видно из примеров использования 2 уровня синергетического эффекта, приведенных на рисунке 7, система распознавания фонем «Phoneme recognizer» может быть использована в таких модулях, как синтез речи, распознавания речи, идентификация языка, диктора, а система определения характеристик канала связи – в модулях распознавания речи, определения языка, диктора и характеристик говорящего.

**1.5.3.** *3 уровень синергетического эффекта*

В то время как за счет первых двух уровней синергетического эффекта можно достичь создания системы с лучшим дизайном архитектуры, системы, которую было бы легко расширять и дорабатывать, напрямую на качество работы систем анализа речи они повлиять не в состоянии. Улучшение качества работы может быть достигнуто за счет использования взаимодействий между различными модулями. Возможные варианты взаимодействия и преимущества, к которым может привести их использование, представлены в виде таблицы 1.

Таблица 1 – Примеры взаимодействий между различными модулями анализа речи

| Модуль | Возможные варианты использования | Комментарии |
|---|---|---|
| Идентификация языка | Распознавание фонем, речи, идентификация диктора, синтез речи | Знание языка, на котором произнесена фраза, позволяет использовать специфичные для данного языка алгоритмы и информацию (например, при решении задачи идентификации диктора может быть использован список дикторов, владеющих данным языком) |
| Идентификация диктора | Распознавание фонем, речи, идентификация языка, синтез речи | Наличие информации о дикторе может позволить использовать предопределенные для данного диктора акустические и языковые модели |
| Идентификация пола, возраста диктора | Распознавание фонем, речи, синтез речи | Информация о поле диктора, его возрасте и эмоциональном состоянии может быть использована для корректировки фильтров и других |

| Модуль | Возможные варианты использования | Комментарии |
|---|---|---|
| | | параметров настройки модулей системы |
| Идентификация темы разговора | Распознавание фонем, речи | Тема разговора может быть использована в качестве дополнительного ограничения, накладываемого на используемый при распознавании лексикон |

**1.5.4.** *Аспекты программной реализации*

Основываясь на свойствах платформы, определенных в п.1.5, можно сделать следующие предположения:

- исходная информация, инструменты и базовые алгоритмы должны быть сгруппированы в отдельные модули по их функциональности;
- базовые структуры данных, такие как вектор MFCC, LPC и другие, должны быть предопределены в системе и использоваться едино всеми модулями;
- модули, такие как модуль распознавания фонем, должны быть предопределены, и их функциональность должна быть описана в стандартных для данной платформы интерфейсах;
- должны быть заданы базовые взаимодействия между модулями;
- платформа должна позволять программисту изменять существующие функции, добавлять новые модули, расширять существующие интерфейсы и переопределять взаимодействия между различными модулями.

Предполагается добиться перечисленных характеристик платформы за счет использования следующих сущностей:

- solution (решение) – сущность, содержащая информацию о том, каким образом платформа должна решать конкретную задачу. Данная сущность задается с помощью вспомогательных сущностей solution flow (см. ниже) и необходимых параметров;
- solution flow (цепочка) – сущность, инкапсулирующая отдельные части алгоритма; задаётся набором действий (activity) и правилами перехода между этими действиями;
- solution flow activity (активити) - сущность, представляющая собой реализацию конкретной функции. В данном контексте, справедливо, что каждая функция каждого модуля реализована в виде отдельного элемента активити. Для сущности solution flow данное действие представляется в качестве черного ящика с известными входными параметрами и ожидаемым результатом работы. Сущности solution flow activity могут быть соединены друг с другом, представляя совместно алгоритм решения задачи. При этом каждое activity может иметь несколько исходящих связей, использующееся соединение определяется на основе результатов работы функции, заложенной в активити;
- task (таск) – сущность, хранящая всю текущую информацию в процессе выполнения платформой конкретного задания. Таск позволяет различным элементам решения получать актуальную информацию о работе другого элемента. Фактически, сущность таск представляет собой временный объект, строящейся на основе схемы, заданной соответствующим ему решением.

Пример реализации данного подхода на примере работы модуля идентификации языка (PRLM-подход) представлен на рисунке 8.

Предположим, что у нас есть модули под названием «Corpora», «Phone recognizer» и «Language model», и что их интерфейсы предоставляют все необходимые функции. При данных предположениях, мы можем описать решение задачи идентификации языка с помощью решения (в соответствии с введенной терминологией), состоящего из трех цепочек, по одной для каждого из описанных выше модулей. Каждая цепочка состоит из последовательности элементов-активити, реализующих ту или иную функциональность модуля. Элемент таск в данном случае ответственен за решение задачи согласно описанному алгоритму (создание потоков для каждой из цепочек, вычисление параметров, переход между элементами-активити).

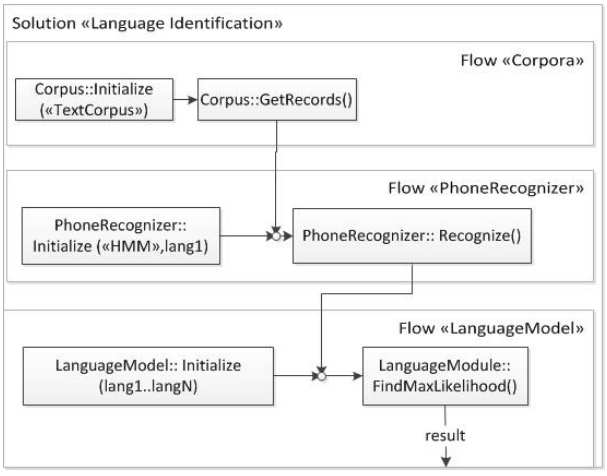


Рисунок 8 – 1 уровень синергетического эффекта в речевых технологиях

Такой тип архитектуры приложения несёт в себе следующие преимущества:

- предоставление инструмента создания сценариев анализа языка, состоящих из простых блоков;
- предоставление простого задаче-ориентированного способа решения конкретных задач. Например, лингвист может изменить

параметры элемента-активити для модуля Corpora и посмотреть, каким образом изменились результаты

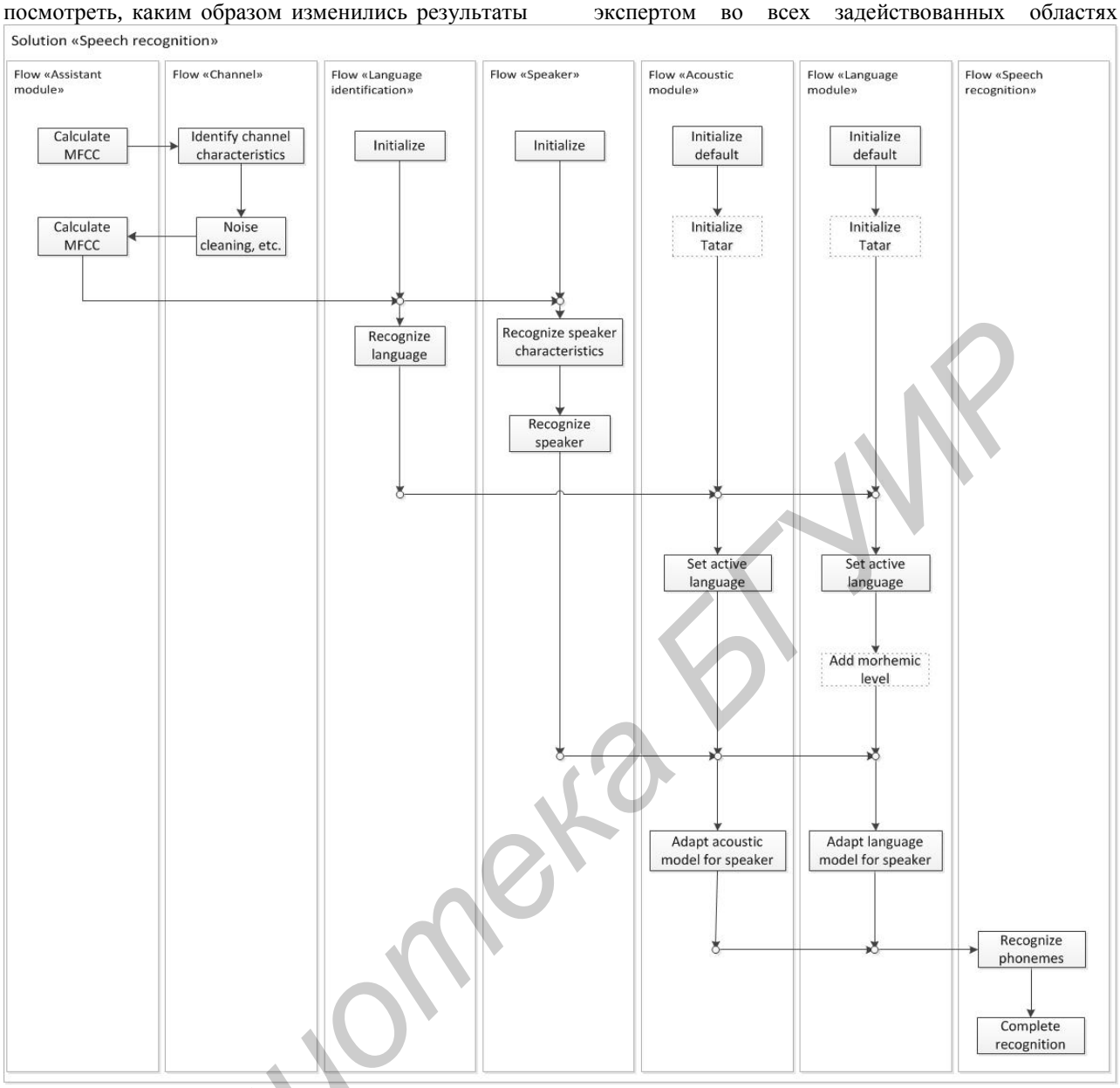


Рисунок 9 – Схема работы системы распознавания речи

анализа речи;

- «параллельная» структура исполнения решений предоставляет возможность уменьшения времени обработки задач.

Кроме того, одним из важнейших преимуществ данной платформы является то, что при таком подходе становится возможным учитывать текущий уровень развития речевых технологий для различных языков. В то время как существует множество ресурсов и алгоритмов, разработанных для английского, французского, испанского языков, существует достаточно большой класс языков с низким уровнем разработок в данной области.

Разработка предлагаемой архитектуры с предопределенными стандартными методами, алгоритмами и речевыми корпусами можно упростить и ускорить процесс создания систем анализа речи для многих языков: разработчики распознавания, при этом от него не требуется быть экспертом во всех задействованных областях подобных систем смогут сконцентрировать свои усилия на значимых особенностях конкретного языка, используя при этом заранее зарекомендовавшие себя алгоритмы. Возможность использования данной платформы для разработки системы анализа речи на татарском языке будет рассмотрена в следующем разделе.

## 2. Платформа анализа речи для татарского языка

Существует несколько основных причин заниматься разработкой речевых технологий для татарского языка. Во-первых, существует социокультурный аспект: язык выступает как инструмент отражения быта и сохранения исторического наследия народа, говорящего на нём. И единственным способом сохранения накопленной информации является помощь языку в

адаптации к современным особенностям жизни, которые включают в себе широкое использование информационных технологий. Во-вторых, создание платформы для анализа татарского языка позволит создавать практические приложения для людей, говорящих на татарском языке, например:

- распознавание и синтез речи – для использования людей с нарушениями зрения;
- распознавание речи – использование для поиска информации в сети Интернет, написание электронных писем, смс, анализа аудио-архивов на татарском языке, для навигации по сайтам, в системах голосового самообслуживания;
- верификация диктора – предоставление безопасного доступа к различной информации;
- распознавание чисел – использование в IVR-меню (interactive voice response menu) в телефонии.

Для демонстрации возможностей использования предложенной архитектуры при проектировании систем анализа речи, приведём схему решения с её помощью задачи распознавания речи (рисунок 9). Кроме того, данный пример демонстрирует возможность адаптации языконезависимого решения к татарскому языку за счет добавления в него специфических для татарского языка элементов. Для наглядности такие элементы выделены в схеме пунктирной линией.

Как видно из рисунка 9, решение «Speech recognition» состоит из 7 отдельных потоков, по одному для каждого из следующих модулей:

- Assistant module – вспомогательный модуль, содержащий инструменты для вычисления стандартных параметров речи (реализующий идею о 1м уровне синергетического эффекта);
- Channel – модуль, определяющий характеристики канала связи и производящий на основе полученных данных фильтрацию возможных искажений исходного сигнала;
- Language identification – модуль идентификации языка диктора;
- Speaker – модуль идентификации диктора;
- Acoustic module – модуль, занимающийся настройкой акустических моделей, используемых при распознавании;
- Language module - модуль, занимающийся настройкой языковых моделей, используемых при распознавании;
- Speech recognition – модуль распознавания речи.

Каждый поток решения состоит из набора элементов-активити, выполняющих определенную процедуру и взаимодействующих согласно установленным связям. Прямые переходы между активити изображены на схеме с помощью стрелок (например, переход между элементом «Calculate MFCC» и элементом «Identify channel characteristics» другого потока). Однако существует ситуации, в которых для выполнения действия необходимо выполнение одновременно нескольких подготовительных процедур. Так, например, до начала процесса распознавания фонем (поток «Speech recognition», активити «Recognize phonemes»), необходимо корректным образом настроить текущие языковую и акустическую модели (активити «Adapt acoustic/language model for speaker»). Для этих целей архитектурой системы предусмотрена возможность синхронизации различных потоков; точка синхронизации в предложенной на рисунке 9 схеме изображается с помощью окружности. Как видно из схемы, результаты работы каждого модуля (потока) используются для корректировки работы других модулей. Так, например, информация о дикторе используется в акустическом и языковом модуле.

## Заключение

В работе представлен подход к созданию платформы анализа речи и её применения для татарского языка. Были определены три уровня синергетического эффекта, которые могут быть использованы в рамках создания систем анализа речи, и рассмотрены основные преимущества предлагаемой архитектуры.

## Библиографический список

**[Huang, 2009]** Huang, X., An overview of modern speech recognition / X. Haung, L. Deng // Handbook of Natural Processing, C. 339-366

**[Hitrov, 2011]** Hitrov, M. Synergetic effect in speech technologies / M. Hitrov // In The 14th International Conference «Speech and Computer», C. 27-32

**[Jurafsky, 2009]** Jurafsy, D. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition / D. Jurafsky, H. James // Prentice Hall, 2009. – 1024 c.

[**Zissman, 1996**] Zissman, A. Comparison of Four Approaches to Automatic Language Identification of telephone Speech / A. Zissman// IEEE transactions on speech and audio processing. 4(1)

[**Rabiner, 1993**] Rabiner, L. Fundamentals of Speech Recognition / L. Rabiner, B.H.Juang // Prentice Hall Signal Processing Series

**[Niesler, 2006]** Niesler, T. Language identification and multilingual speech recognition using discriminatively trained acoustic models / T. Niesler, D. Willett // In ISCA Workshop on Multilingual Speech and Language Processing (MULTILING 2006), Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa

## SPEECH ANALYSIS PLATFORM PROTOTYPE FOR TATAR LANGUAGE

Khusainov A.F., Suleymanov D.S.

*«Applied semiotics» Institute, Tatarstan Academy of Sciences,*

*Kazan (Volga region) federal university, Russia, Kazan*

**khusainov.aidar@gmail.com**

**dvdt.slt@gmail.com**

In this paper, we present a complex speech analysis platform for Tatar language. We explore main modules of this system including speech command recognition,

language identification and phoneme recognition modules. The major idea was to investigate potential of cross-module interactions to increase effectiveness of each module work.

## INTRODUCTION

Nowadays the rapid development of all types of electronic devices led to a problem of limited speed and quality of existing instruments of computer-human communication (keyboard, etc.). One way of solving this problem is to use speech to communicate with computers. This way includes developing speech recognition, synthesis, speaker, language identification systems. But the greater part of scientists' research activities are focused on separate areas of speech technologies. This fact could be explained with economic reasons and complexity of tasks. Such kind of approach makes easier to go deeper in theoretical and practical aspects of specific area of speech technologies. In other hand, it gives too little attention to explore and develop ways of building complex architecture which could provide functionality of all tasks of speech analysis.

Based on the above observation we have focused our research on studying the possible structure of complex platform for speech analysis and its application to Tatar language. Creating such kind of architecture will make possible, first of all, to increase effectiveness of speech analysis systems via synergetic effect and, secondly, will provide tool for developing speech technologies for languages which don't yet have well-designed speech analysis systems.

### MAIN PART

Speech analysis technologies include various areas of tasks; for today following main areas could be distinguish as:

- speech recognition;
- speech synthesis;
- language identification;
- speaker identification;
- speaker verification;
- diarisation;
- speaker's characteristics identification (sex, emotional status, prosodic features);
- signal channel's characteristics identification (channel type, existence and nature of noise, etc.).

Synergetic effect in speech analysis systems could be described consisting of three levels:

- lower level – using common tools;.
- middle level – using whole modules (like phoneme recognition) as common blocks;
- top level – exchanging information between different modules.

According to this classification we can see that lower and middle levels give researchers possibility to accelerate algorithms developing speed and increase simplicity of that process by using same entities in different situations. At the same time top level gives us opportunity to increase effectiveness of separate modules' work.

Based on idea of three-level synergetic effect and speech analysis modules' structure, we now can enumerate main features of proposed system:

- modular design – possibility to separate developing of different modules;
- flexibility - system must provide possibility of using and comparing different algorithms;
- expandability – easy to expand system by adding new modules or new type of realizations;
- language-adjustable – skeleton of the system consists of language-independent blocks, but adding language-specific items and algorithm realizations could increase speed and quality of the system.

Found on given features of platform we suggest that it can be realized by using following internal entities:

- solution – entity which contains information about how concrete task will be solved in system. It consists of algorithm parts represented by solution flows (see below) and required parameters;
- solution flow – entity which encapsulates separate part of algorithm; it's represented by actions, called activities (see below), and transition rules;
- solution flow activity – entity which contains realization of specific function. It can be assumed that all functions of all speech analysis modules are wrapped into this entity;
- task – consolidates all information during execution program.

One of the biggest advantages of this approach is the fact that proposed platform takes into account current development level of speech technologies for different languages. While there are various corpora and algorithms for such languages as English, French, etc., there is still a wide range of less developed languages. And creating proposed platform with pre-defined standard methods, algorithms and existing corpora can simplify process of developing comprehensive speech technologies for another language; developers can concentrate on distinctive features of this language and use already well-designed tools.

### CONCLUSION

We have presented a speech platform and its application for Tatar language. We have also discussed possible benefits of creating such kind of complex system and demonstrated three levels of synergetic effect that can be used. We conclude that proposed programming model is well suited to speech technologies due to their computational complexity, etc.

In the future we intend to develop proposed speech platform for Tatar language, implement existing language tools (e.g. Tatar speech synthesis, command recognition system); and to create application-oriented systems based to this platform.