Лабораторная работа №15

«Фонетика»

Уровень 1

Ответы на вопросы по теме «Задачи автоматического анализа речи»:

1. Какие задачи речевой обработки являются наиболее актуальными?

Наиболее актуальными задачами речевой обработки являются: автоматическое создание транскрипции, синхронное преобразование живой речи в текстовый формат, управление голосовыми помощниками, выявление особенностей речи говорящего, выделение наиболее значимых слов в речи, идентификация и верификация личности, создание и разработка технологий для людей с ограниченными возможностями, создание диалоговых систем и чат-ботов

2. В каких областях решение этих задач находит свое применение?

Преобразование звучащей речи в текст применяется в:

- Голосовых помощниках (Siri, Alexa, Google Assistant)
- Системах голосового ввода текста
- Транскрипция аудиозаписей (интервью, лекций, совещаний)
- Автоматизация колл-центров

Определение, кто говорит (по биометрическим характеристикам голоса) используется в биометрической защите и авторизации

Определение эмоционального состояния говорящего применяется в:

- В call-центрах для анализа удовлетворённости клиента
- В психологии и медицине (мониторинг депрессии, стресса)
- В интерактивных системах (роботы, персонализированные ИИ-ассистенты)

Синтез речи применим в:

- В навигаторах (например, Яндекс.Навигатор)
- В книгах с озвучиванием (аудиокниги)
- В образовательных приложениях для детей и людей с ограниченными возможностями

Анализ речи используется для диагностики и мониторинга заболеваний:

- Неврологические расстройства (например, болезнь Паркинсона)
- Психические заболевания (шизофрения, депрессия)
- Логопедия и восстановление после инсультов

3. Как происходит процесс верификации говорящего?

Под верификацией понимают определение системой говорящего как «своего» или «чужого». В качестве примера верификации можно рассмотреть следующую ситуацию: говорящий производит речевой сигнал – сообщает данные о своей личности (инициалы или специальный код), а система автоматического распознавания индивидуальных особенностей голоса и речи должна подтвердить личность говорящего, исходя из эталонной модели, т.е. имеющейся записи голоса говорящего

4. В чем заключается отличие верификации от идентификации говорящего по голосу?

Процесс верификации подтверждает личность говорящего, а процесс идентификации распознает, исходя из поступившего речевого сигнала, кто из пользователей вступает в контакт.

5. Чем обусловлена актуальность решения такой задачи как преобразование речи в текст?

Данная технология активно применяется: при автоматическом создании субтитров; при разработке приложений и технических средств для людей с ограниченными возможностями, у которых возникают сложности при наборе текста вручную; для распознавания команд голосовыми помощниками (Google Assistant, Яндекс Алиса, Siri, Alexa).

6. В чем заключается отличие первых систем Speech-to-Text от тех, которые разрабатываются в настоящее время?

Первые образцы систем STT работали исключительно на основе эталонных моделей, при этом речь членилась системой на отдельные крупные блоки, после чего происходил поиск наиболее близкого акустического сигнала из уже записанных в базе данных, что приводило к большому количеству ошибок.

В настоящее время, благодаря использованию баз данных коллокаций (устойчивых сочетаний) и N-gramm (последовательностей из N символов), системы STT способны предугадывать целые фразы, исходя из правильно распознанного фрагмента речи.

7. За счет чего системы распознавания речи стали допускать меньше ошибок?

Немаловажную роль в сокращении количества ошибок при распознавании играет внедрение машинного обучения, которое позволяет системам STT постоянно совершенствоваться.

8. Перечислите этапы работы систем распознавания говорящего и преобразования речи в текст.

Можно выделить три этапа преобразования речи в текст:

- Анализ сигнала. Полученный на вход речевой сигнал очищается от лишних шумов и делится на минимальные фрагменты.
- Распознавание сигнала. Фрагменты проходят через эталонную модель, определяющую, какие звуки были произнесены; после чего система пытается объединить их в значимые единицы.
- Преобразование в текст. При помощи баз данных и модели языка происходит подбор нераспознанных единиц, исходя из контекста, и преобразование полученных результатов в графический вид.

Практическое задание:

Оцените качество работы систем распознавания речи на русском и английском языках. Проверьте, ухудшится ли результат, если на вход системы STT сообщить: несвязный речевой сигнал (случайные наборы слов), несогласованные друг с другом слова (например, по роду) или речевой сигнал с параллельной речью нескольких людей (или посторонним шумом). Можете использовать системы распознавания речи, внедренные в онлайн переводчики (Яндекс Переводчик, Google Translation) или в голосовые помощники (Google Assistant, Яндекс Алиса, Siri, Alexa).

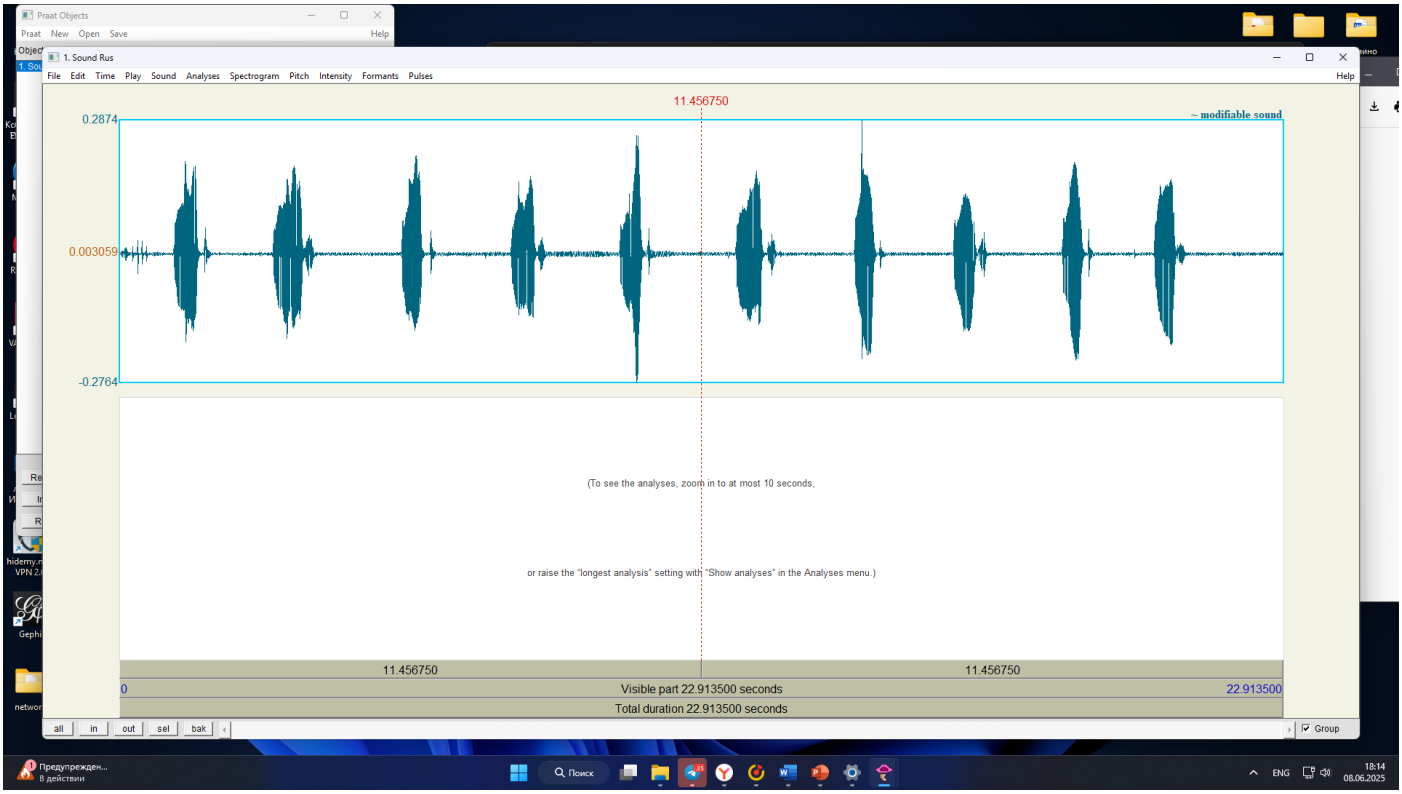
Выполнение практического задания:

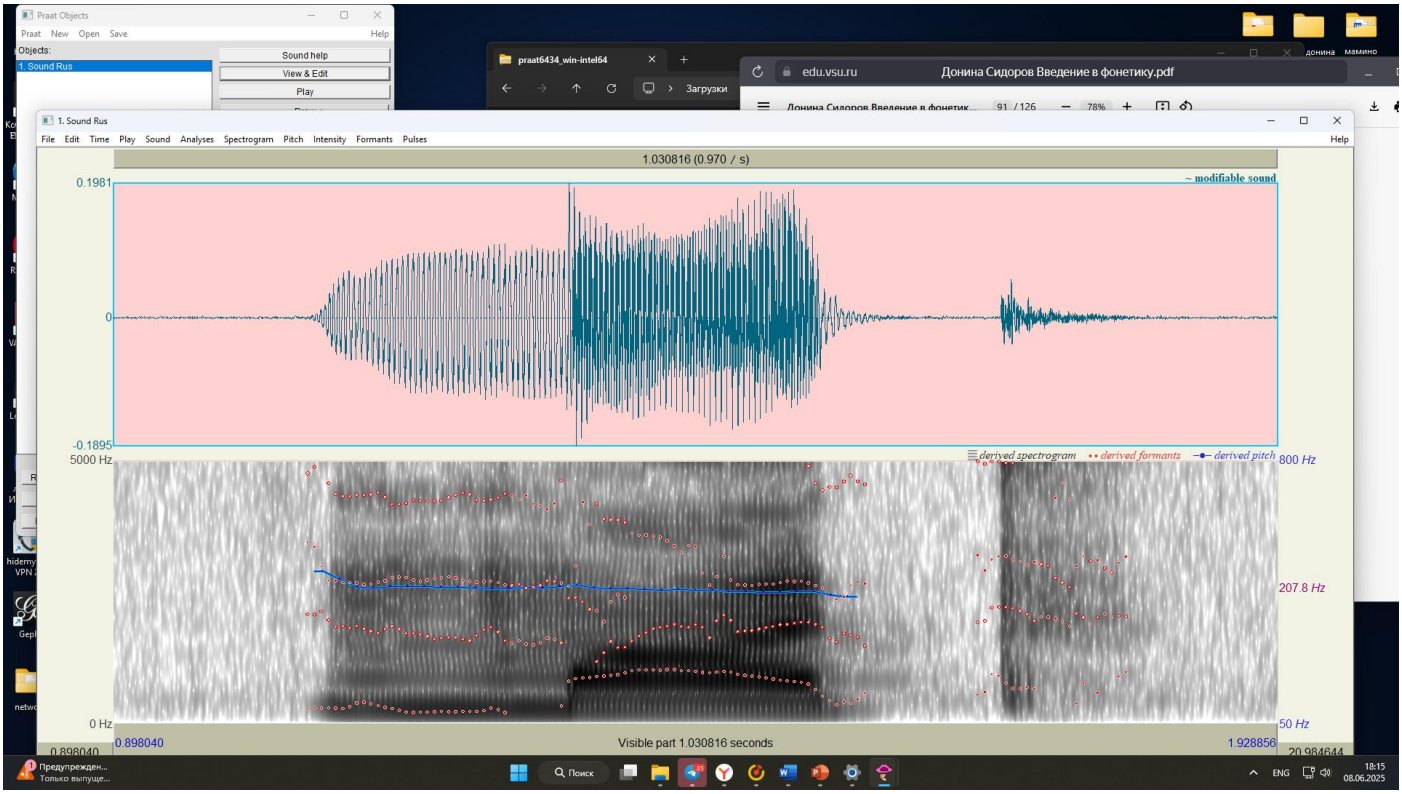
Для оценки качества работы систем распознавания речи я выбрала голосовые помощники Siri и Яндекс Алиса:

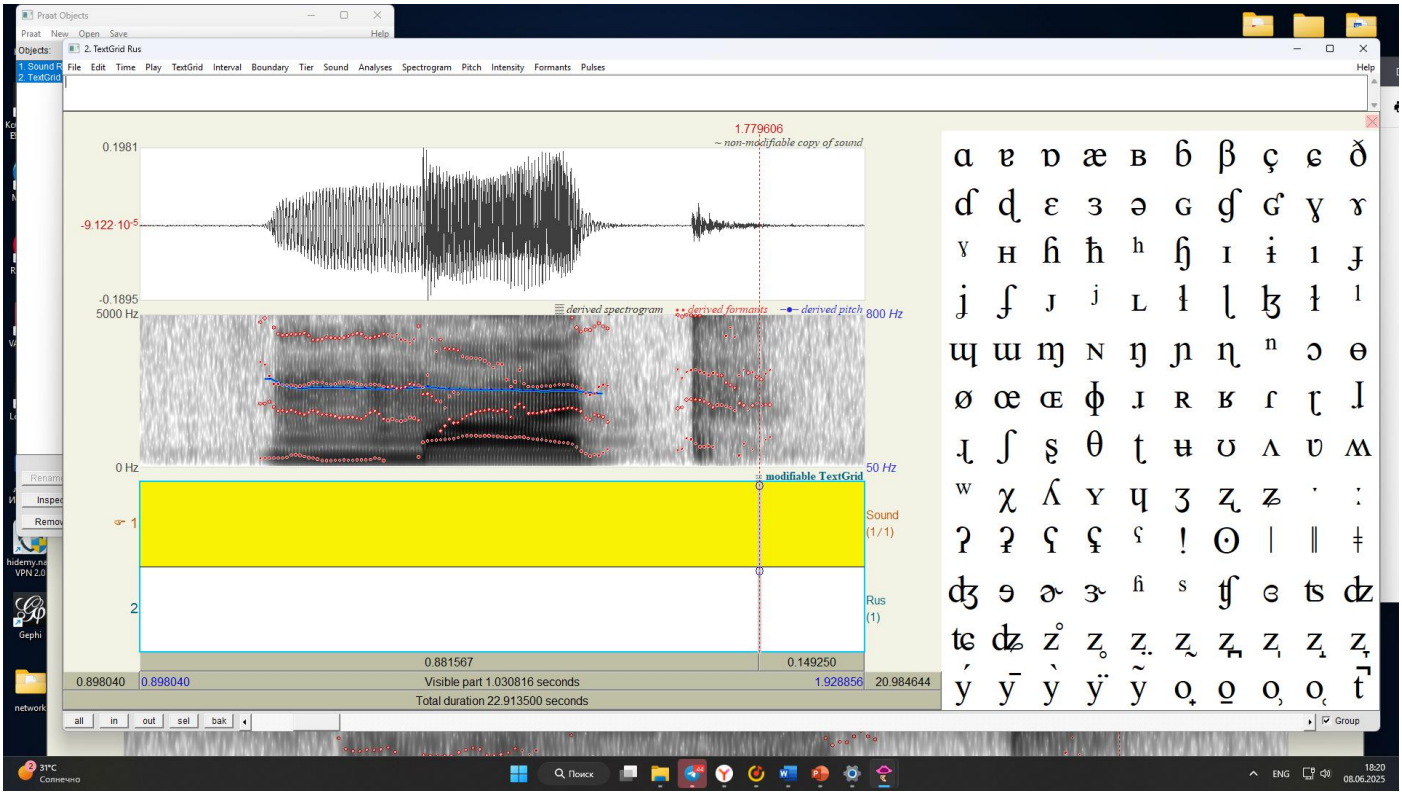
условие	google assistant	яндекс алиса
говорю связно и четко на русском языке	точно распознает речь	точно распознает речь
говорю связно и четко на английском языке	неплохо распознает английскую речь, но иногда делает ошибки	не очень хорошо распознает речь, делает ошибки в словах
говорю случайные наборы слов	неплохо распознает речь	неплохо распознает речь, но делает ошибки
говорю несогласованные друг с другом слова	система может переформулировать	передает как есть, но иногда может переформулировать
говорю с посторонним шумом	смешивает речь	смешивает речь

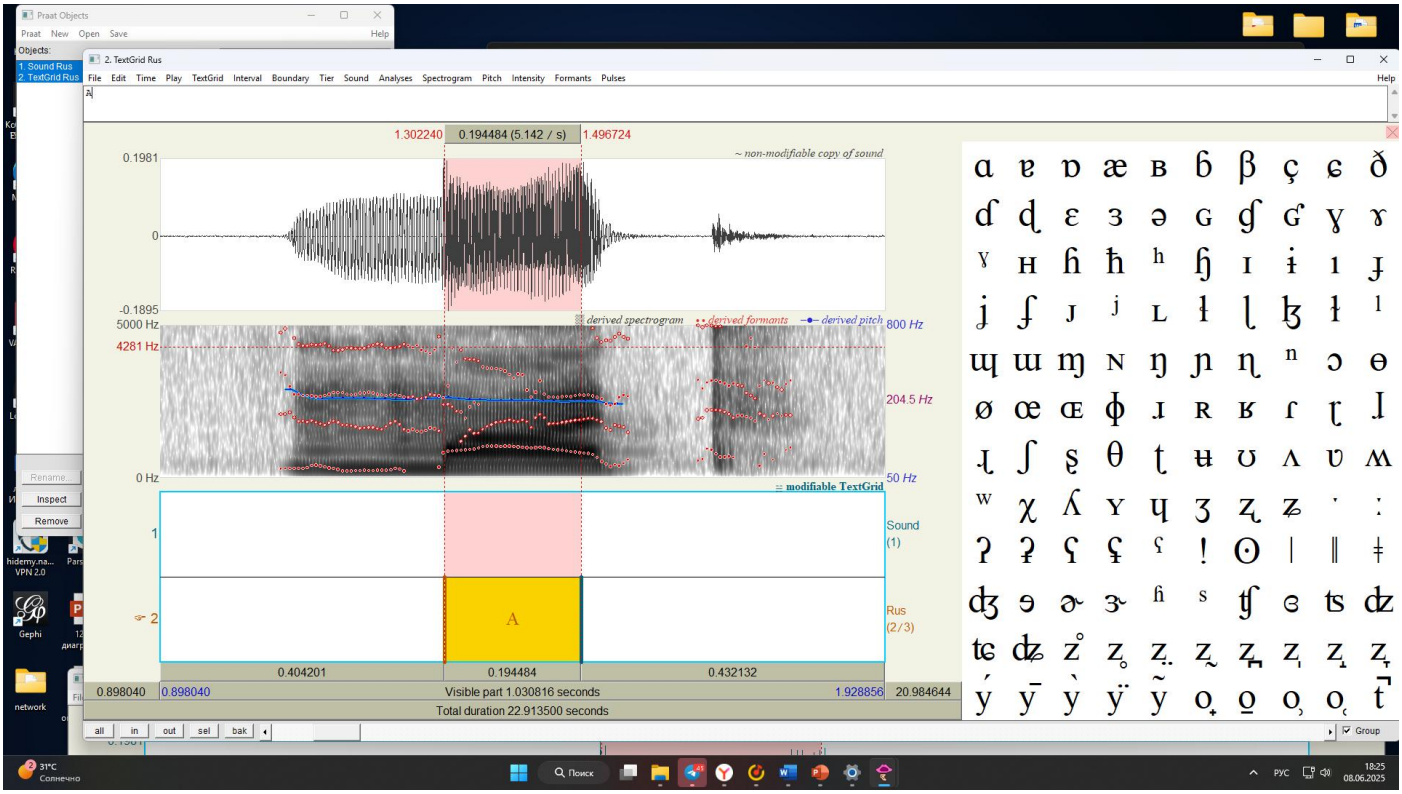
Лабораторная работа

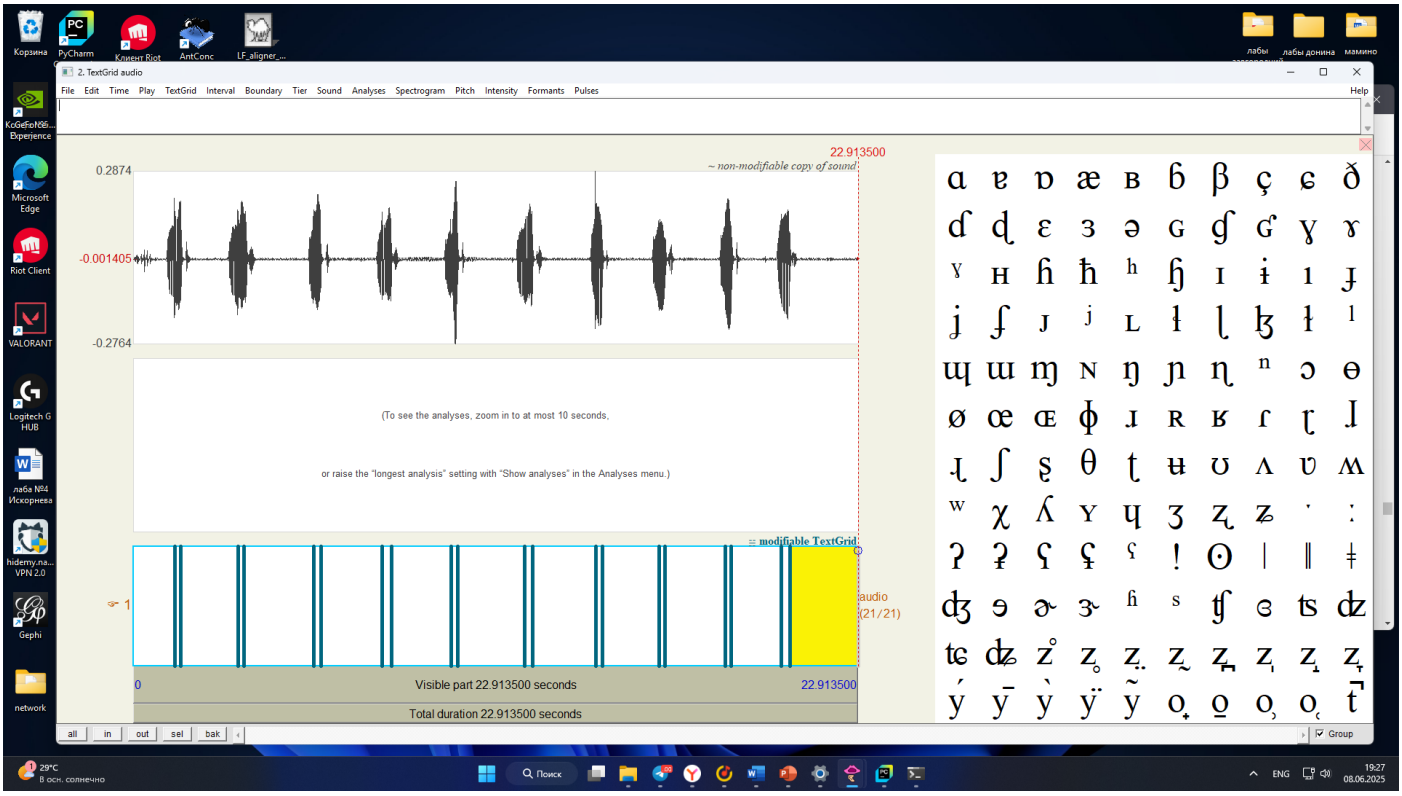
1. Использовались словоформы: «мат»-«мять», «мот»-«мёть», «мэт»-«меть», «мыт»-«мить», «мут»-«мють»




Praat Objects
Praat New Open Save Help
Objects:
1. Sound Rus
Sound help
View & Edit
Play
praat6434_win-intel64
Загрузки
edu.vsu.ru
Донина Сидоров Введение в фонетику.pdf
1. Sound Rus
File Edit Time Play Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
1.030816 (0.970 / s)
0.1981
0
-0.1895
~ modifiable sound
derived spectrogram
derived formants
derived pitch
5000 Hz
800 Hz
207.8 Hz
0 Hz
50 Hz
0.898040
Visible part 1.030816 seconds
1.928856
20.984644


Praat Objects
Praat New Open Save Help
Objects:
2. TextGrid Rus
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
1.779606
~ non-modifiable copy of sound
0.1981
-9.122·10⁻⁵
-0.1895
derived spectrogram
derived formants
derived pitch
5000 Hz
800 Hz
0 Hz
50 Hz
modifiable TextGrid
Sound
(1 / 1)
Rus
(1)
0.881567
0.149250
0.898040
Visible part 1.030816 seconds
1.928856
20.984644
Total duration 22.913500 seconds
all
in
out
sel
bak
Group


2. TextGrid Rus
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses
1.302240
0.194484 (5.142 / s)
1.496724
~ non-modifiable copy of sound
0.1981
0
-0.1895
derived spectrogram
derived formants
derived pitch
5000 Hz
4281 Hz
800 Hz
204.5 Hz
0 Hz
50 Hz
modifiable TextGrid
Sound (1)
Rus (2/3)
A
0.404201
0.194484
0.432132
0.898040
0.898040
Visible part 1.030816 seconds
1.928856
20.984644
Total duration 22.913500 seconds
all in out sel bak
Group


2. TextGrid audio
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses
22.913500
~ non-modifiable copy of sound
0.2874
-0.001405
-0.2764
(To see the analyses, zoom in to at most 10 seconds,
or raise the "longest analysis" setting with "Show analyses" in the Analyses menu.)
modifiable TextGrid
audio (21/21)
0
Visible part 22.913500 seconds
22.913500
Total duration 22.913500 seconds
all in out sel bak
Group

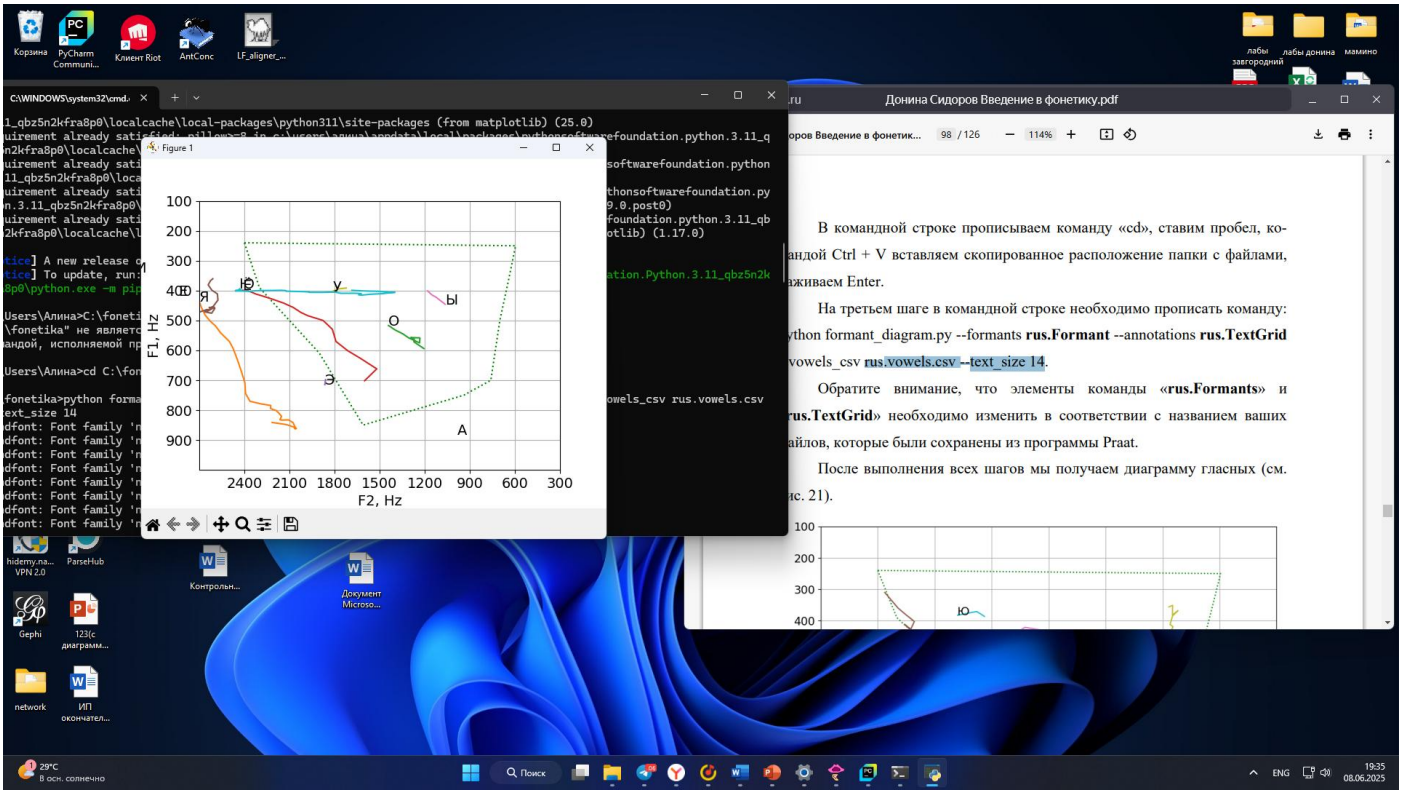
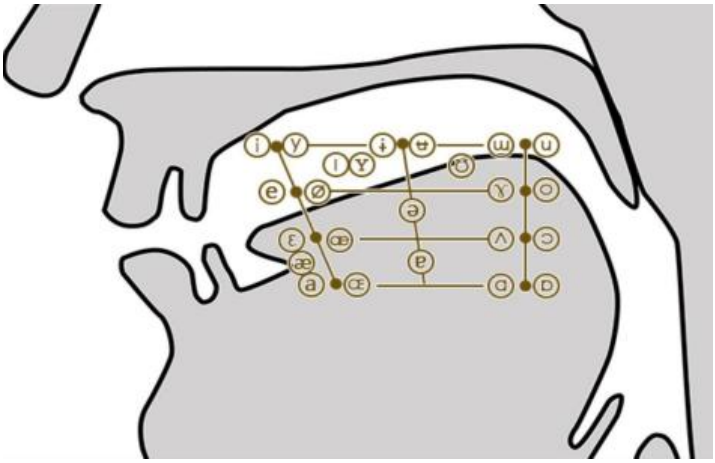


Figure 1
F1, Hz
F2, Hz
100
200
300
400
500
600
700
800
900
2400 2100 1800 1500 1200 900 600 300
Я
Ю
У
Ы
О
Э
А
Донина Сидоров Введение в фонетику.pdf
98 / 126
114%
В командной строке прописываем команду «cd», ставим пробел, ко-
андой Ctrl + V вставляем скопированное расположение папки с файлами,
аживаем Enter.
На третьем шаге в командной строке необходимо прописать команду:
ython formant_diagram.py --formants rus.Formant --annotations rus.TextGrid
vowels_csv rus.vowels.csv --text_size 14.
Обратите внимание, что элементы команды «rus.Formants» и
rus.TextGrid» необходимо изменить в соответствии с названием ваших
айлов, которые были сохранены из программы Praat.
После выполнения всех шагов мы получаем диаграмму гласных (см.
ис. 21).



Ответы на вопросы по теме «Основы работы систем преобразования текста в речь»

1. Что из себя представляют системы Text-To-Speech и что является их основной задачей?

Основная задача систем преобразования текста в речь (TTS – Text-ToSpeech) – создание голосового сообщения без участия человека напрямую

1. По каким критериям оценивают сгенерированную речь?

Основные критерии для оценки качества синтеза речи: разборчивость, т.е. возможность человеком декодировать речевой сигнал, и естественность – автоматически созданное высказывание не должно значительно отличаться от речи человека

2. Какой способ синтеза речи подразумевает использование записанных фрагментов человеческой речи? Опишите принцип его работы.

В основе компилятивного синтеза речи заложены заранее записанные сегментные единицы (от звуков речи до целых высказываний), т.е. на вход системы подаются подготовленные человеком образцы речевых сигналов, которые система объединяет и озвучивает в требуемой последовательности.

3. Какие два способа синтеза речи основываются на акустических свойствах речи? В чем их отличие?

Формантный (синтез речи происходит путём моделирования формант — частотных полос, характеризующих речевые звуки (особенно гласные). Форманты определяют тембр и тип звука, и они напрямую связаны с акустическими характеристиками сигнала.) и статистический (речь строится на основе статистических моделей , обученных на реальных речевых данных. Эти модели предсказывают акустические параметры (спектр, длительность, просодию и т. д.) на основе текста.)

4. В чем заключается особенность артикуляционного способа синтеза речи?

Артикуляционный синтез основывается на создании модели человеческого речевого аппарата и ее реализации с помощью некоторого механизма – «говорящей» машины, которая порождает речевой сигнал путем артикуляции

5. Перечислите основные сложности для синтеза речи каждым способом. Какие существуют пути их решения?

- Компилятивный: Если записанные человеком фрагменты речи достаточно велики (например, целые высказывания), подобный синтезатор не сможет воспроизвести множество разнообразных высказываний; в противном случае, при записи минимальных сегментных единиц (звуков речи), возможности системы гораздо выше, но возникают проблемы с разборчивостью и естественностью речевого сигнала на выходе. Для разрешения проблемы использования минимальных сегментных единиц, исследователи пришли к использованию не отдельных фонов, а дифонов – речевых участков, которые берут свое начало в середине первого звука, а заканчиваются в середине второго звука. Использование дифонов позволяет удачно синтезировать переходные участки между звуками, но при этом возникают помехи (шумы) при дифонных соединениях вследствие перепадов частот разных сегментов, что вызывает необходимость в дополнительной обработке. Распространено и использование интонационных модификаций для определенных участков синтезируемой речи, например – конец повествовательно высказывания произносится с понижающейся тональностью, а вопросительные – с восходящей.
- Формантный: Трудности с передачей естественной интонации и сложных согласных.
- Статистический: Менее естественная речь.

- Артикуляционный: Моделирование физических процессов требует мощных вычислений: решение уравнений аэродинамики, упругости тканей, взаимодействия воздуха и мягких тканей. Нет полной и точной базы знаний о том, как именно двигаются артикуляторы при произнесении каждого звука.

Практическое задание

Проанализируйте как минимум четыре синтезатора речи. Оцените разборчивость и естественность сгенерированного речевого сигнала. Проверьте, учитывают синтезаторы речи суперсегментные характеристики речи – интонацию (паузы при перечислении, интонирование в восклицательных и вопросительных предложениях) и ударение (способен ли отличить одинаковые словоформы с разными ударными слогами (пример: Все зависит от того, к чему у тебя лежит душа′; После принятия ду′ша не стоит сразу выходить на улицу)). Для анализа можете использовать синтезаторы речи, внедренные в онлайн переводчики (Яндекс Переводчик, Google Translation), речь голосовых помощников (Google Assistant, Яндекс Алиса, Siri, Alexa), а также следующие онлайн синтезаторы речи:

• Acapela: https://www.acapela-group.com/

• Oddcast: https://ttsdemo.com/

• Text-to-speech: https://text-to-speech.imtranslator.net/speech.asp

• PiliApp: https://ru.piliapp.com/text-to-speech/

• iSpeech: https://www.ispeech.org/text.to.speech

условие	Acapela	Oddcast	PiliApp	Яндекс Переводчик
Восклицательное предложение	Делает акцент на восклицание, но не сильный	Нет акцента на восклицание	Нет акцента на восклицание	Нет акцента на восклицание
Вопросительное предложение	Произносит с вопросительной интонацией	Произносит с вопросительной интонацией	Произносит с вопросительной интонацией	Не соблюдает вопросительную интонацию
Пунктуация	Соблюдает паузы	Соблюдает паузы	Соблюдает паузы	Соблюдает паузы
ударение	Не отличает одинаковые словоформы с разными ударными слогами	Различает одинаковые словоформы с разными ударными слогами	Не отличает одинаковые словоформы с разными ударными слогами	Различает одинаковые словоформы с разными ударными слогами
естественность	7/10	3/10	2/10	2/10
разборчивость	10/10	10/10	10/10	9/10