Речевой корпус данных TeCoRus.

Речевой корпус данных предназначен для того, чтобы предоставить достаточное количество контролируемого речевого материала для проведения исследовательских и прикладных работ по распознаванию русской речи, в частности, в области телекоммуникационных приложений.

Материал корпуса записан в 1999-2000 годах, дикторы – сотрудники ВЦ РАН и студенты МГЛУ.

Корпус может быть использован для оценивания параметров контекстно-зависимых Марковских моделей звуков речи, создания и тестирования алгоритмов и программного обеспечения для распознавания слитной речи и цифровой обработки речевых сигналов, например компенсации искажений канала передачи, распознавания речи в условиях ограничений полосы частот, помех и т.п.

Речевой корпус данных TeCoRus включает две части.

1. *Акустико-фонетическая часть*.

Акустико-фонетическая часть корпуса данных предназначена для создания базовых (загрузочных) моделей звуков речи, например, Марковских моделей. Она предоставляет массив основных вариантов звуков русской речи, записанных от дикторов, обладающих нормативным произношением, без явно выраженных дефектов речи. Тексты акустико-фонетической части содержат 3050 предложений, прочитанных 6 дикторами. Каждое предложение записано в отдельный аудио файл и сопровождается отдельным текстовым файлом-аннотацией, который содержит текст высказывания, его фонетическую транскрипцию и сегментацию речевого файла.

Весь речевой материал записан в двухканальном варианте:

- узкополосный телефонный сигнал (полоса до 4 кГц, проводные линии связи - 135 и 137 станции МГТС)
- широкополосный микрофонный сигнал (полоса до 11кГц, микрофон Koss SB35).

Размер акустико-фонетической части – 600 Мбайт.

2. *Речевая часть* корпуса данных предназначена для обучения моделей звуков речи, подгонки параметров и тестирования алгоритмов и программ распознавания и обработки речевых сигналов. Речевая часть состоит из системы интерактивных тестов - интервью и включает в себя как читаемый, так и спонтанно произносимый материал (в свободной форме, и в форме контролируемых ответов на вопросы). В частности, речевая часть содержит образцы произнесения:

- простых утвердительных и отрицательных ответов на вопросы
- цифр (последовательностей, длиной от 1 до 16 цифр)
- чисел
- названий дней недели, месяцев
- имен
- букв русского алфавита

- дат и времени (часов, минут)
- фонетически представительных предложений
- коротких рассказов на заданную и свободные темы

Каждый ответ в ходе интервью записан в отдельный аудио файл и сопровождается аннотацией.

Речевая часть включает записи 100 дикторов.

Система аннотирования речевых файлов выбрана совместимой с системой аннотирования для баз данных TIMIT, за исключением расширенной системы меток фонем для русского языка, а также описания неречевых акустических событий.

60% речевой части корпуса данных записано в двухканальном варианте:

- телефонный канал (полоса до 4 кГц, проводные (135 и 137 станции) и сотовые каналы связи.
- микрофонный сигнал (полоса до 11кГц, микрофоны Shure SM10 и Koss SB35).

Размер речевой части – 2,1Гбайт.

Речевой корпус входят:

- сам корпус данных (аудио файлы и файлы с аннотацией)
- произносительный словарь корпуса данных
- сопутствующее программное обеспечение.

Объем речевого корпуса данных TeCoRus составляет 2.7 Гбайт