# АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ СПОНТАННОЙ УКРАИНСКОЙ РЕЧИ (НА МАТЕРИАЛЕ АКУСТИЧЕСКОГО КОРПУСА УКРАИНСКОЙ ЭФИРНОЙ РЕЧИ)

***Людовик Т.В.*** *(tetyana.lyudovyk@gmail.com),*
***Пилипенко В.В.*** *(valeriy.pylypenko@gmail.com),*
***Робейко В.В.*** *(valya.robeiko@gmail.com)*

Международный научно-учебный центр информационных технологий и систем,
Киев, Украина

# Цель исследования

- эксперимент по распознаванию украинской речи с использованием базовой системы распознавания;
- анализ ошибок;
- меры по повышению точности распознавания с учетом спонтанного характера речи и сужения ее тематики;
- сравнение результатов распознавания спонтанной речи участников телешоу судебной тематики и речи реального судьи, выступающего в ходе судебных заседаний.


**Frederick Jelinek:**

**"Every time I fire a linguist, the performance of our speech recognizer goes up."**

("a linguist leaves the group", "we fire a phonetician/linguist")

# Исследованный материал

- Речевой материал:

аудиозаписи телепередач «Судові справи» («Судебные дела») из корпуса АКУЕМ; ~ 52 часа, ~ 2 000 дикторов

- Текстовый материал:

корпус АКУЕМ и тексты из Интернета (400 Мбайт)

- Контрольная выборка:

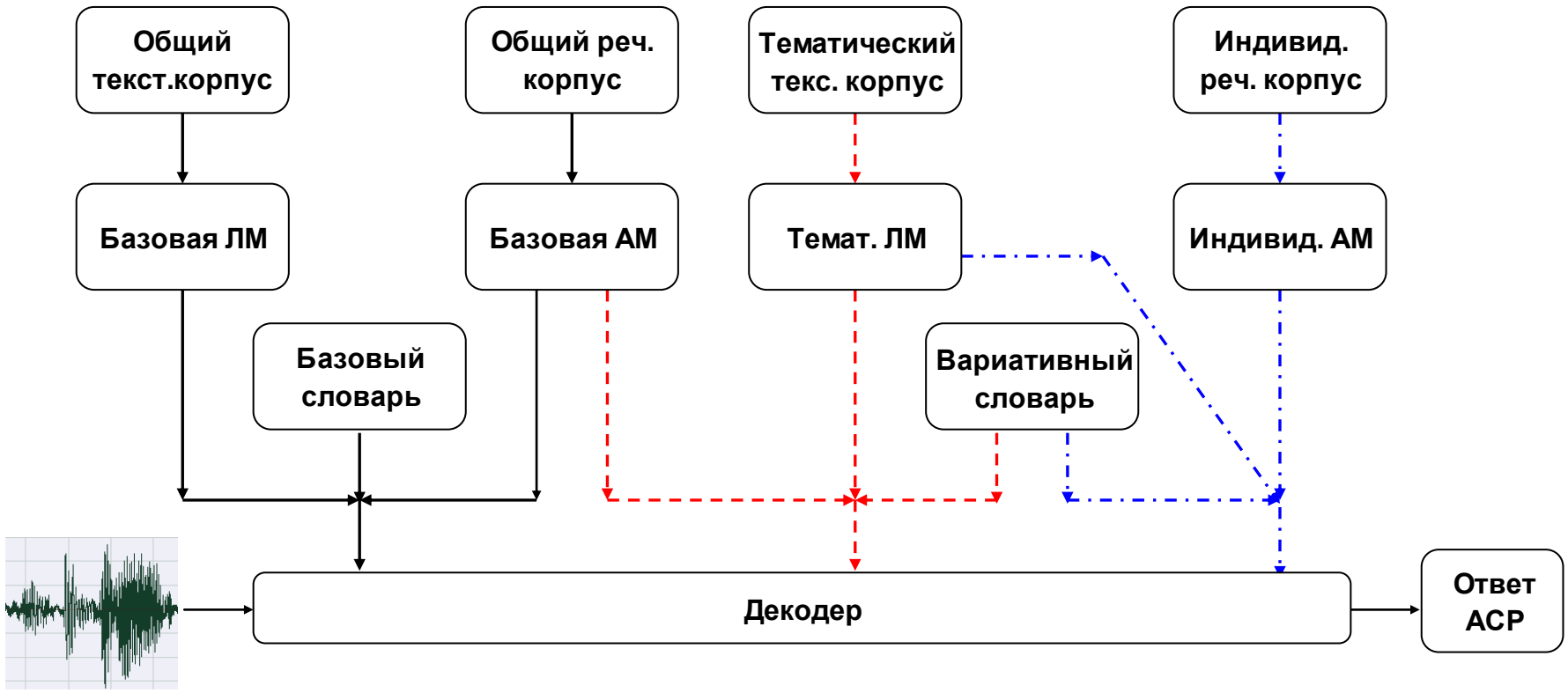
аудиозаписи ~ 3,74 часа; 29 500 словоформ; 32 диктора.

# Корпус АКУЕМ

Акустический корпус украинской и русской речи, записанной из украинского телеэфира (Акустичний корпус українського ефірного мовлення).

Пример. Русская речь (ток-шоу “Свобода слова”, диалог ведущего и депутата):

# Конфигурации системы распознавания речи



# HTK

Универсальный инструментарий, позволяющий создавать системы распознавания речи

**http://htk.eng.cam.ac.uk/**

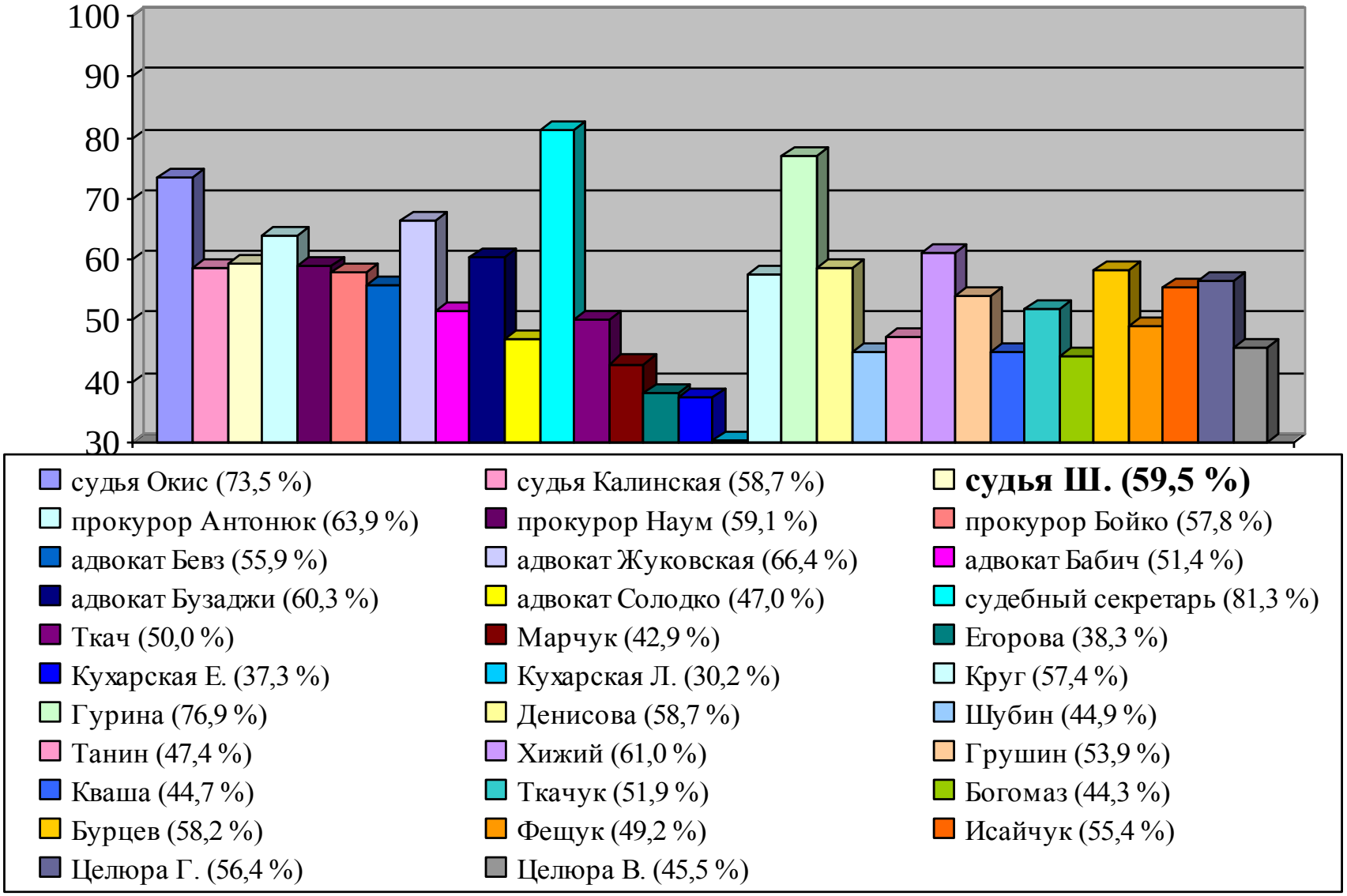
На основе HTK создана система распознавания речи.

Использовались результаты обучения, полученные инструментарием HTK.

# Результаты экспериментов. Базовый вариант

| Дикторы | Источник аудиозаписей | Точность распознавания (%) | Образцы речи |
|---|---|---|---|
| Судья Окис | телепередача | 73,47 | 100% |
| Судья Калинская | телепередача | 58,65 | |
| Судья Ш. | судебное заседание | 59,47 | |
| Прокурор Антонюк | телепередача | 63,90 | 90% |
| Прокурор Наум | телепередача | 59,10 | |
| Прокурор Бойко | телепередача | 57,76 | |
| Адвокат Бевз | телепередача | 55,93 | 92% |
| Адвокат Жуковская | телепередача | 66,38 | 75% |
| Адвокат Бабич | телепередача | 51,64 | |
| Адвокат Бузаджи | телепередача | 60,28 | |
| Адвокат Солодко | телепередача | 46,95 | |
| Судебный секретарь Сологуб | телепередача | 81,26 | 100% |

# Результаты экспериментов. Базовый вариант



# Анализ ошибок. Речь судьи на судебном заседании

| Причина | Количество ошибок | % от всего количества ошибок |
|---|---|---|
| Отсутствие в словаре (OOV) | 230 | 21 |
| Редукция | 164 | 15 |
| Суржик | 88 | 8 |
| Лингвистическая модель | 59 | 5 |
| Быстрый темп | 50 | 5 |
| Вдохи и хезитация | 46 | 4 |
| Нет пометок в корпусе | 30 | 3 |
| Числительные | 16 | 1 |
| Игнорирование пометки | 15 | 1 |
| «Отрезано» начало | 12 | 1 |
| Слова с дефисом | 11 | 1 |
| Неправильная транскрипция в словаре | 9 | 1 |
| Омофоны | 8 | 1 |
| Индивидуальная особенность | 6 | 1 |
| Неизвестная причина | | 32 |

# Анализ ошибок. OOV

| В корпусе | Распозналось как |
|---|---|
| винності батальщікова у | в області на те що його |
| відбігли | він біг у |
| відбуватися | від двадцять |
| галайчук | На гачок |
| галайчук | Коли чую |
| галайчук | Але щоб |
| галайчук | але чому |
| галайчук все | але часом |
| діти є утриманці | в дію тринадцять |
| зговір | з г о в |
| зговір | згоди |
| каюсь каюсь | київська русь |
| класним | власним |
| компрометуючими | компромат у чому |
| прокурора колнишеву | прокурор у у нашому |

# Анализ ошибок. Редукция

| *Словоформа* | *Литературная фонемная транскрипция* | *«Спонтанные» фонемные транскрипции* |
|---|---|---|
| виявлено | в И й а в л е н о | в И й а л е н о<br>в И й л е н о<br>в И й л и н и |
| п'ятнадцять | п й а т н А дз' ц' а т' | п й а т н А ц' а т'<br>п' а т н А ц' а т'<br>п' а т н А ц' |
| в'ячеславовича | в й а ч е с л А в о в и ч а | в' а ч е с л А в о в и ч а<br>в' а ч е с л А в и ч а |

# Анализ ошибок. Лингвистическая модель

| **В корпусе** | **Распозналось как** |
|---|---|
| тридцять вам | тридцять два |
| разом цим | разом з цим |
| дев'яте лютого | дев'ятого лютого |
| сто сто сто сто сто | до списку сто сорок |
| Президентский вертикаль власти (рус.) | |
| Теперь нас такая возможность есть (рус.) | |

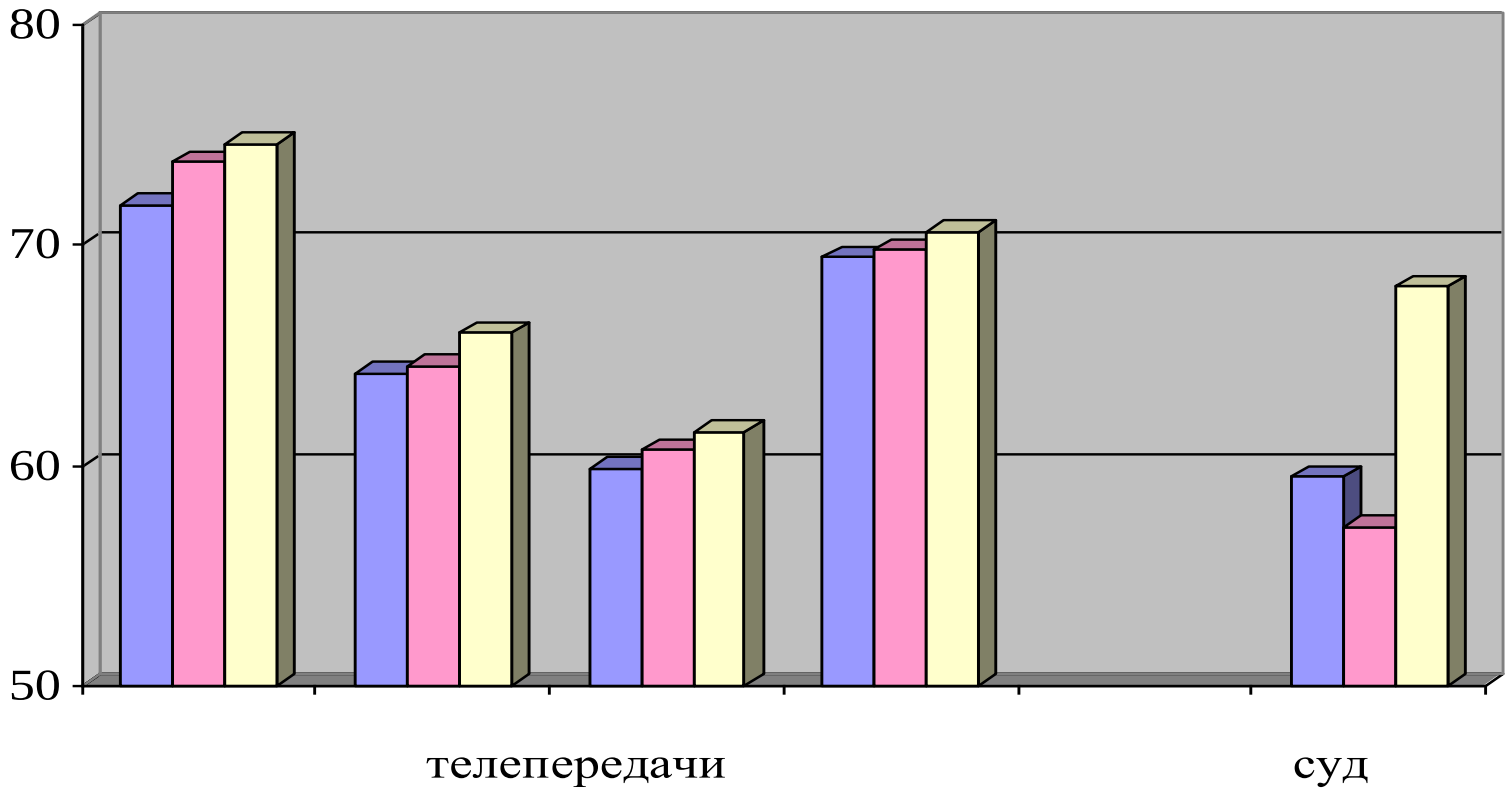
# Анализ ошибок. Вдохи и хезитация

| В корпусе | Распозналось как |
|---|---|
| *в* документи | два документи |
| дома. *вд* | до москви |
| потребуєте такої *вд* | потребує тих хто скоїв |
| справі *е* | справі у |
| *е* захисту | і захисту |
| вас *е* влаштовує | вас не влаштовує |
| а *е* вашого | вашого |
| *е* підсудні | не підсудні |
| ви *е* сказали | ви самі сказали |
| *е* сімейний | один сімейний |
| показання *м* | показаннями |
| це *м* | це вам |
| *м* не соромно | ми не соромно |
| то *ок* що | то що |
| *пл* сідайте | і сідайте |

# Анализ ошибок. Омофоны

| В корпусе | Распозналось как |
|---|---|
| про те | проте |
| які спільні | якісь спільні |
| чия | чи я |
| немає | не має |
| не відомо | невідомо |
| і з | із |
| щодо | що до |

# Результаты экспериментов



# Выводы

- Базовая система распознавания речи обеспечивает 59,61 % точности распознавания спонтанной украинской речи (данные конца января 2011 года).
- Близкая к нормативной речь с незначительной склонностью к хезитации и редукции распознается в среднем с 81,26 % точности.
- Низкая степень нормативности и разборчивости снижает точность распознавания до 58–59 %.
- Склонность диктора одновременно к хезитации и редукции приводит к точности распознавания 47–60 %.
- Наиболее существенное повышение точности достигнуто за счет включения тематического частотного словаря и учета вариативности, связанного со спонтанным характером произнесения.

**Спасибо за внимание!**