

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.

Sign in

In Everlasting Memory

Frederick Jelinek
1932 - 2010

Story Video Add Photo Sign Guestbook

Add It

Frederick Jelinek

Frederick Jelinek (born as Bedřich Jelínek 18 November 1932 – 14 September 2010) was a researcher in information theory, automatic speech recognition, and natural language processing. Jelinek's early career produced fundamental contributions to information theory and coding. He later became a pioneer in applying statistical modeling to speech recognition and natural language processing. He and his colleagues were the first to apply hidden Markov models to these tasks and also the first to build statistical models for machine translation. His special interest was language modeling, and much of his recent work had to do with moving beyond n-gram models to take advantage of long distance and syntactic regularities. Jelinek was born in Kladno in what was then Czechoslovakia. He taught at the Massachusetts Institute of Technology from 1959 to 1962, at Harvard University

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.

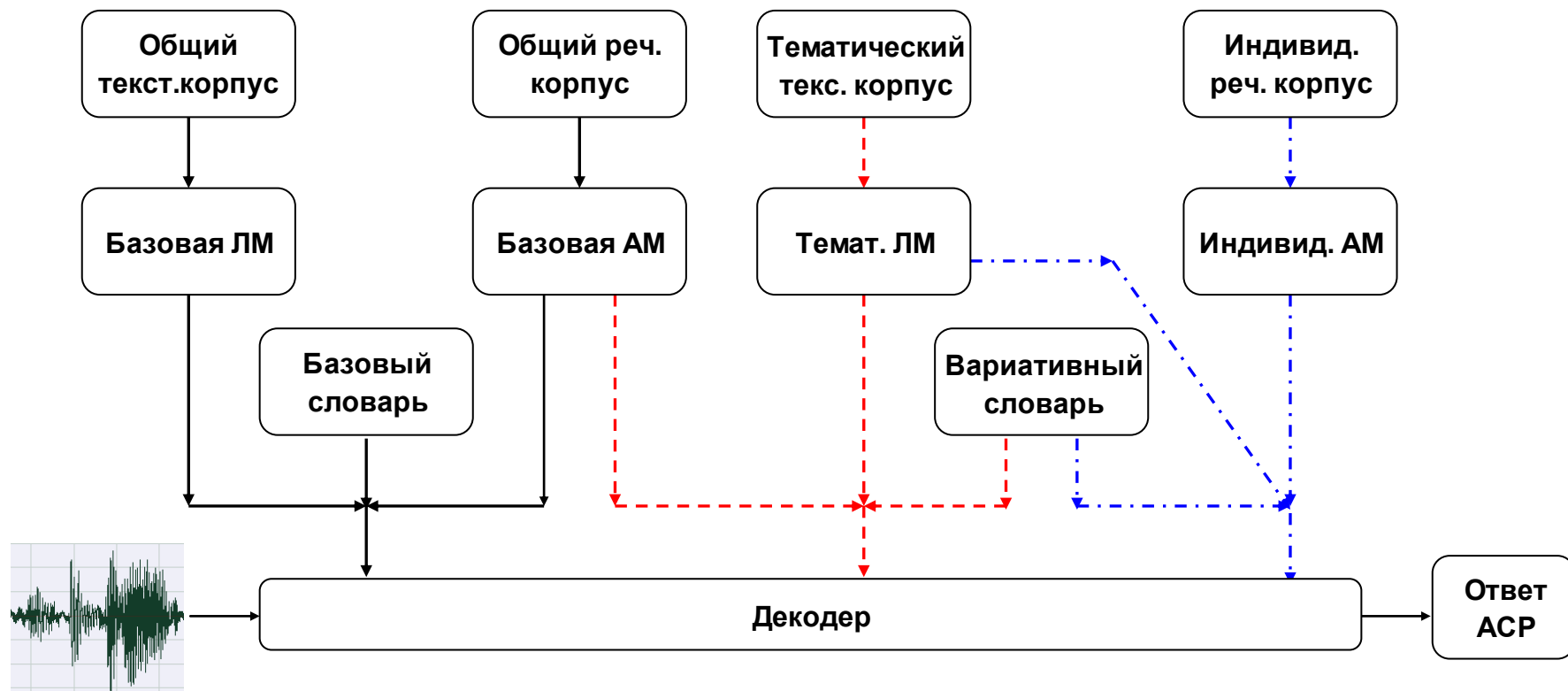
Корпус АКУЕМ

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



НТК












Универсальный инструментарий, позволяющий
создавать системы распознавания речи

<http://htk.eng.cam.ac.uk/>

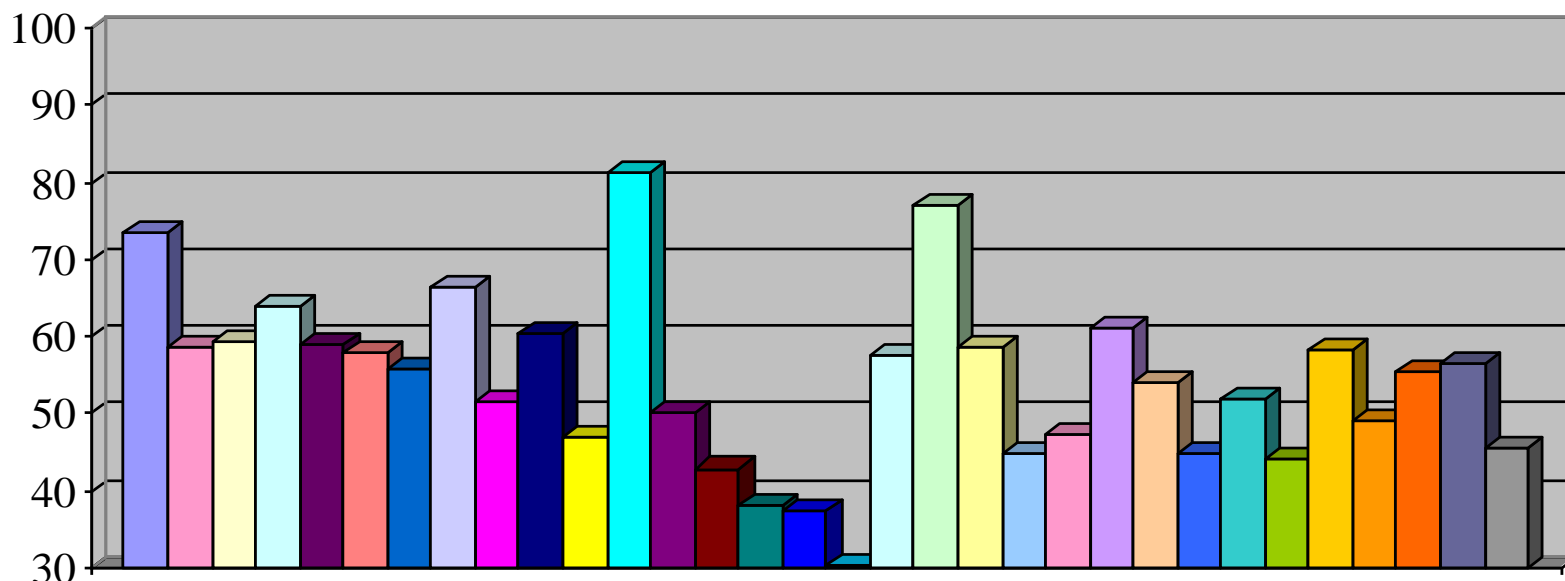
На основе НТК создана система распознавания
речи.

Использовались результаты обучения, полученные
инструментарием НТК.

Результаты экспериментов. Базовый вариант

Дикторы	Источник аудиозаписей	Точность распознавания (%)	Образцы речи
Судья Окис	телепередача	73,47	 100%
Судья Калининская	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



- | | | |
|-----------------------------|------------------------------|-------------------------------|
| ■ судья Окис (73,5 %) | ■ судья Калинская (58,7 %) | ■ судья Ш. (59,5 %) |
| ■ прокурор Антонюк (63,9 %) | ■ прокурор Наум (59,1 %) | ■ прокурор Бойко (57,8 %) |
| ■ адвокат Бевз (55,9 %) | ■ адвокат Жуковская (66,4 %) | ■ адвокат Бабич (51,4 %) |
| ■ адвокат Бузаджи (60,3 %) | ■ адвокат Солодко (47,0 %) | ■ судебный секретарь (81,3 %) |
| ■ Ткач (50,0 %) | ■ Марчук (42,9 %) | ■ Егорова (38,3 %) |
| ■ Кухарская Е. (37,3 %) | ■ Кухарская Л. (30,2 %) | ■ Круг (57,4 %) |
| ■ Гурина (76,9 %) | ■ Денисова (58,7 %) | ■ Шубин (44,9 %) |
| ■ Танин (47,4 %) | ■ Хижий (61,0 %) | ■ Грушин (53,9 %) |
| ■ Кваша (44,7 %) | ■ Ткачук (51,9 %) | ■ Богомаз (44,3 %) |
| ■ Бурцев (58,2 %) | ■ Фешук (49,2 %) | ■ Исайчук (55,4 %) |
| ■ Целюра Г. (56,4 %) | ■ Целюра В. (45,5 %) | |

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

<i>Словоформа</i>	<i>Литературная фонемная транскрипция</i>	<i>«Спонтанные» фонемные транскрипции</i>
виявлено	в И й а в л е н о	в И й а л е н о в И й л е н о 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' п' а т н А ц' а т' 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 в' а ч е с л А в и ч а

Анализ ошибок. Лингвистическая модель

В корпусе	Распозналось как
тридцать вам	тридцать 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.)	

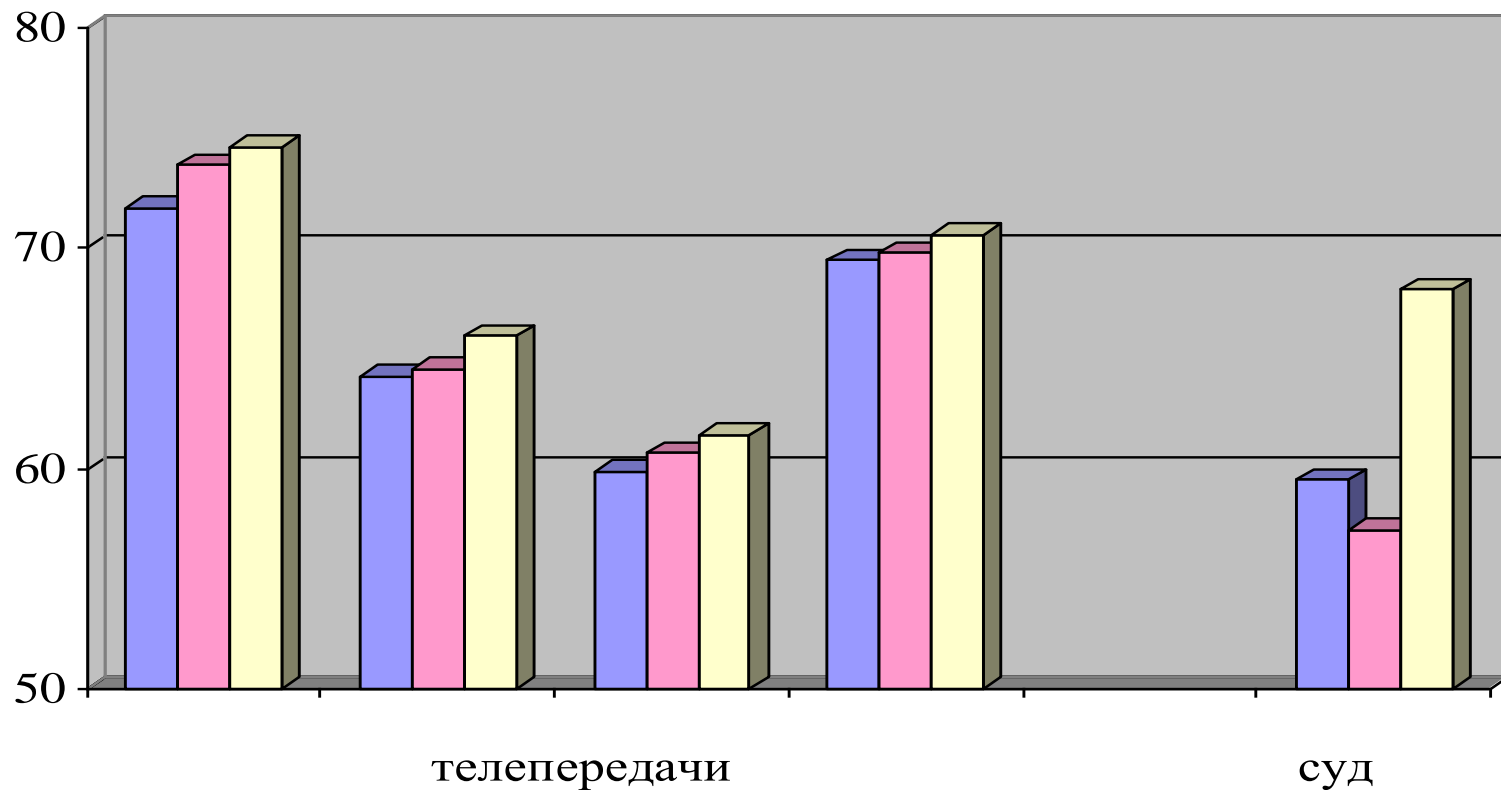
Анализ ошибок. Вдохи и хезитация

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



■ базовая модель

■ новая лингвистическая модель

■ новая лингвистическая модель + новый лексикон

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!