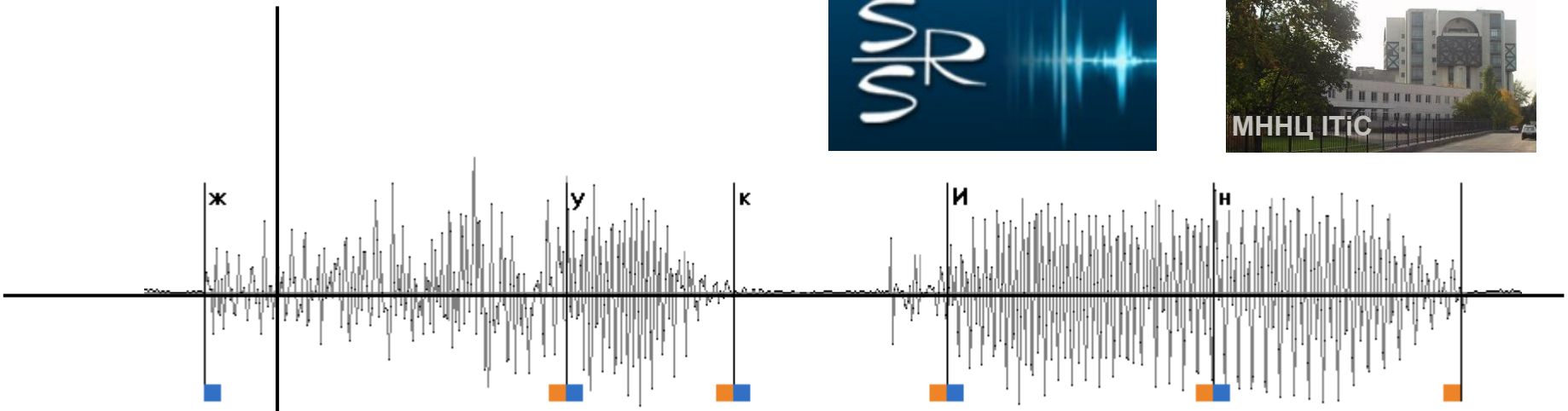
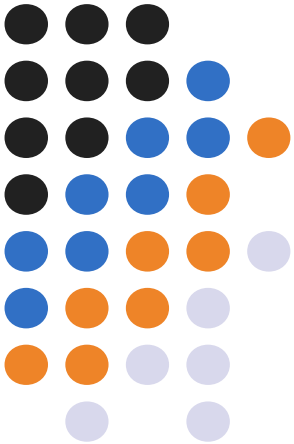




Акустичний корпус українського ефірного мовлення

Валентина Робейко

Акустичний корпус українського ефірного мовлення (АКУЕМ)

АКУЕМ – акустичний корпус, який створюється для розроблення систем розпізнавання мовлення. Робота над корпусом розпочалася у жовтні 2009 року. Участь у розробці корпусу беруть фахівці з відділу розпізнавання та синтезу звукових образів МННЦІТіС та ТОВ Спеціальні реєструючі системи (СРС).

Характеристика АКУЕМ

Характеристика акустичного корпусу:

- за метою використання: загальний;
- за типом мовленнєвого матеріалу: читане мовлення, підготоване мовлення, спонтанне мовлення;
- за типом текстового матеріалу: зв’язні політематичні тексти;
- за типом мовленнєвого сигналу: публічне мовлення, теле-, радіомовлення, мовлення у природних обставинах.
- за типом анотацій: сегментне анотування;
- за кількістю мов: двомовний (українська та російська мови).

Характеристика АКУЕМ

На даний момент анотовано понад **500 звукових файлів**.

Це становить понад **300 годин звукових записів** (разом для української та російської мови).

Корпус містить понад **50 000 слів української мови** та майже **65 000 слів російської мови**.

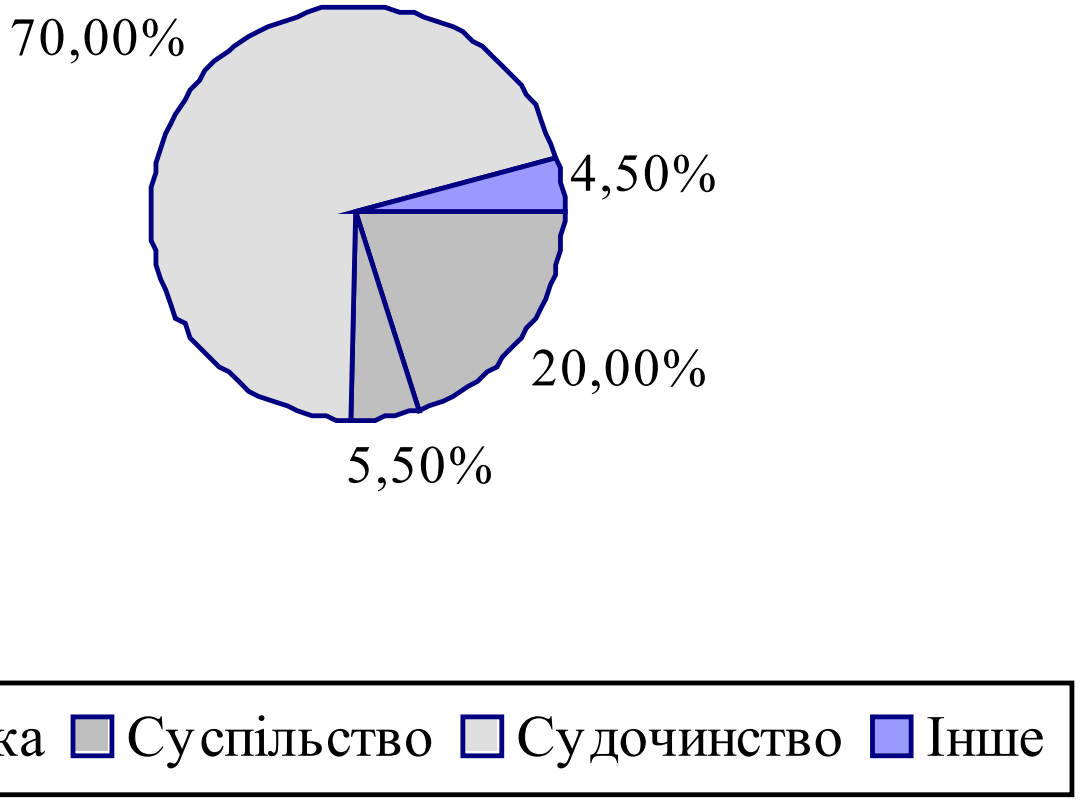
Проаналізоване мовлення **кількох тисяч дикторів**.

Створено **словник суржику** (понад 2500 слів), **словник територіальних та соціальних діалектів** (понад 2000 слів) для української мови.

Структура та склад акустичного корпусу

В АКУЕМ увійшли матеріали різної тематики та жанрів, але основу корпусу складають звукові записи таких рубрик: **новини та інтерв'ю** (політика, культура, освіта, суспільство тощо), **телепередачі та телетрансляції** (судові засідання, політичні дебати, публічні виступи та ін.). У цілому корпус повинен відображати повну картину мовлення українського теле- та радіоефіру, але кількісний розподіл звукових записів по жанрах не є рівномірним, що пов'язано з підбором текстів тієї чи іншої тематики для нагромадження матеріалів, потрібних для робіт по розпізнаванню

Співвідношення записів різних тематик (за тривалістю у хвилинах)



Цільова аудиторія АКУЕМ

Цільовою аудиторією проекту є в першу чергу **розробники систем автоматичного розпізнавання українського та російського мовлення**. АКУЕМ призначений для навчання та тестування систем розпізнавання мовлення. На матеріалах корпусу проводяться численні наукові експерименти у галузі аналізу та розпізнавання мовлення, наприклад, виявлення та класифікація екстралінгвістичних мовленнєвих явищ, дослідження реальних акустичних обставин мовлення, дослідження реальних варіантів вимови дикторів, вивчення специфіки усного спонтанного мовлення на різних рівнях та багато інших.

Цільова аудиторія АКУЕМ

Важливою частиною користувачів корпусу можуть стати **спеціалісти сфери синтезу мовленнєвих сигналів**, оскільки АКУЕМ є джерелом для різноманітних досліджень у галузі.

АКУЕМ відображає сучасну мовну ситуацію в Україні, включає як літературний, так і живий, розмовний стиль мовлення. Тому величезною є роль акустичного корпусу для широкого спектру **досліджень у ділянках лінгвістики, діалектології, мовленнєвої акустики, психоакустики, фонетики, фонології та інших галузей науки**.

Досвід авторів у розробці акустичних корпусів

Україномовний багатодикторний корпус **«UkReco»** містить понад 30 тис. реалізацій фонетично збалансованих слів і тисячі речень близько 100 дикторів, що мешкають в різних областях України.

Акустичний корпус мовлення Верховної Ради України – виконані через телевізор записи виступів близько 400 осіб в Українському парламенті. Обсяг мовлення – біля 40 годин. Використовується для розробки системи розпізнавання злитого мовлення.

Україномовний корпус опорного диктора містить більше 20 тис. речень і 10 тис. реалізацій ізольованих слів. Використовується в пофонемному і поскладовому розпізнаванні.

Для синтезу українського мовлення використовується **корпус, створений на основі жіночого голосу професійного диктора**.

Програмне забезпечення корпусу

Ефективне створення корпусу не можливе без розвиненого інструментарію. До цього інструментарію відносяться програмні засоби для стенографування звукових записів, подальшого їх сегментування та анотування (транскрибування), автоматичного виправлення транскрипцій, статистичного аналізу результатів сегментування, а також підготовки матеріалу до навчання розпізнавання.

Інструментарій. Стенографування

Стенографування звукозаписів засобами SRS Report

Стенографіст створює транскрипцію звукозапису з рівнем деталізації, що включає ознаки мови і мовця. За допомогою ножної педалі відбувається орієнтовний поділ на сегменти, що відповідають зміні мовця, з точністю в 1–2 секунди. Відеоряд, що супроводжує звукозапис, полегшує визначення особи мовця.

Ознака мови вводиться графічним позначенням. Також вказуються ділянки сигналу, де мовлення нерозбірливе, перекривається шумами, або й зовсім відсутнє.

Серед задач стенографіста є забезпечення відсутності орфографічних помилок у текстах, що супроводжують звукозапис. Для цього використовуються стандартні засоби перевірки орфографії, адаптовані до специфіки стенограм: враховуються позначення мовної ознаки та додаються ознаки відхилення від нормативів літературної мови для відповідних слів.

Інструментарій. Сегментування та анотування

Засоби сегментування звукозаписів

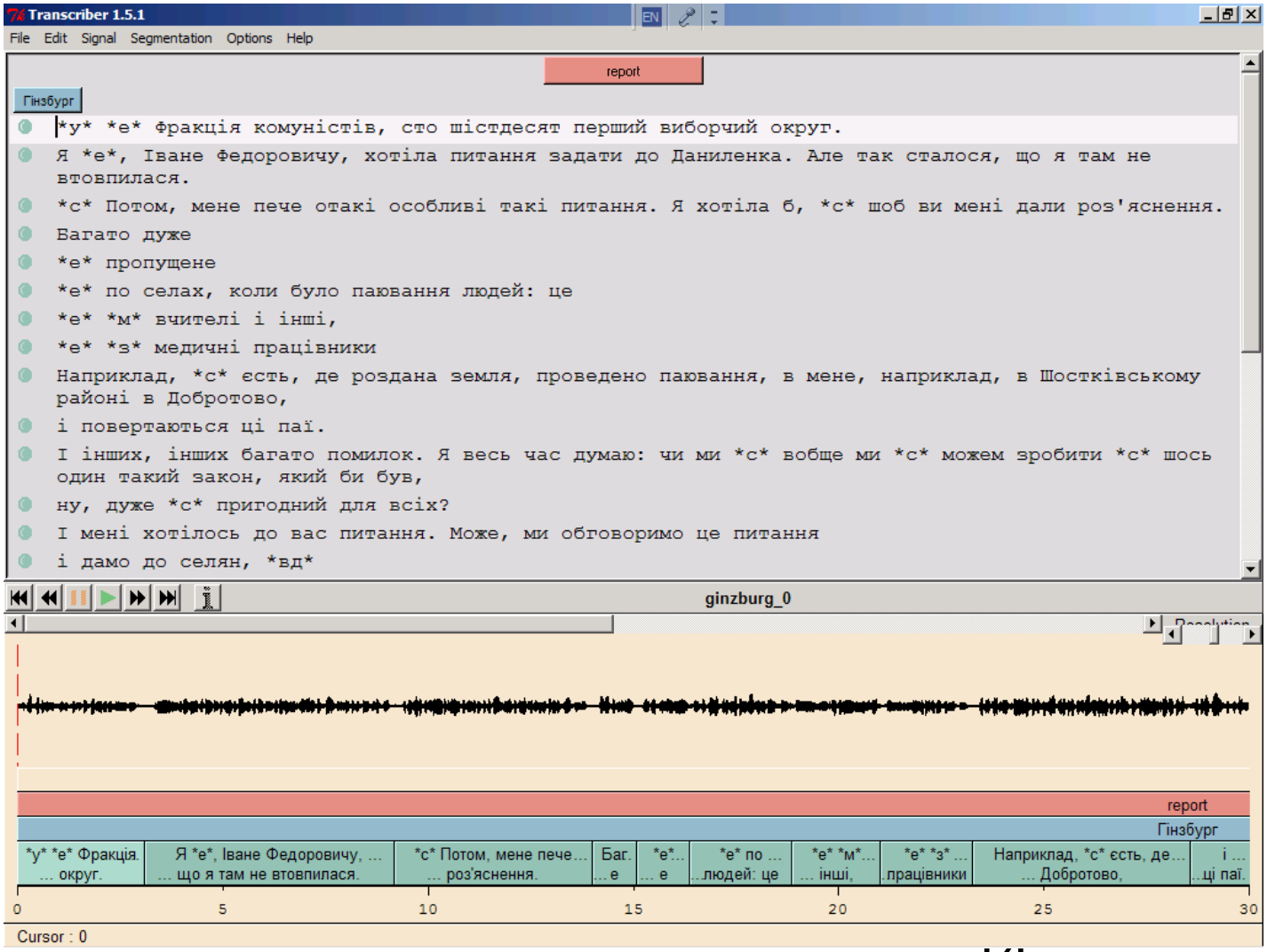
Сегментування виконується засобами програмного забезпечення з відкритим кодом **Transcriber 1.5.**, адаптованого до кирилиці. Підготовлений фахівець із відповідним рівнем лінгвістичної та комп'ютерної освіти поглиблює деталізацію транскрипції, отриманої в результаті стенографування звукозаписів. Проводиться ретельне розбиття за паузою мовленнєвих сегментів, синхронізація їх з відповідним текстом, їх збагачення рядом ознак, що стосуються як окремих слів і звуків, так і в цілому сегмента і конкретного диктора.

Подальший аналіз сегментування полягає у виявленні та виправленні типових помилок та внесенні деяких регулярних змін, що зумовлюються розвитком концепції корпусу, появою та вирішенням контраверсій. Програмний модуль **xml2mlf** дає змогу, зокрема, автоматизувати цей процес.

Інструментарій. Сегментування та анотування

Анотування акустичного корпусу відбувається за допомогою програми **Transcriber** (http://www.etca.fr/CTA/gip/Projets/Transcriber/)

Зразок розміченого звукового запису



Інструментарій. Статистичний аналіз

Засоби статистичного аналізу

Утиліта **ParseTrs** обчислює різні статистики для відсегментованого корпусу, зокрема формуються:

- частотні словники для різних мов, що зустрічаються в корпусі;
- частотні словники суржику, соціального та регіонального діалектів, абревіатур, редукованих слів та ін.;
- статистики довжин мовленнєвих сегментів для кожного звукового файлу та загальна статистика;
- статистики довжин мовленнєвих сегментів для кожного диктора.

Інструментарій. Автоматизація виправлення помилок

Засоби автоматизації виправлення помилок

Утиліта **xml2mlf** дає змогу зійснювати регуляризовані перетворення анотацій аудіозаписів, зокрема:

- проводити заміну фрагментів тексту і позначок в залежності від поточної мови та стилю;
- коригувати характеристики дикторів та здійснювати їх перейменування

Інструментарій. Підготовка до розпізнавання

Засоби підготовки до навчання розпізнавання мовлення

Готуються звукові файли для розбиття на фрази, придатні для навчання та розпізнавання. При цьому кожному звуковому сегменту відповідає текстовий запис та ім’я диктора.

У словник системи розпізнавання занесені слова, які відповідають звукам типу *е*, і, відповідно, під час формування тексту фрази ці звуки розглядаються як окремі слова. (Були проведені експерименти по навчанню на такі звуки-слова, і результати показали надійність їх визначення ~ 80%.)

Інструментарій. Підготовка до розпізнавання

Засоби підготовки до навчання розпізнаванню мовлення

Позначення, які характеризують цілий сегмент, наприклад, *стук*, *муз*, пропонується використовувати для побудови гаусівських сумішей моделей (GMM) для того, щоб система розпізнавання визначала такі сегменти та відносила їх до відповідного класу.

Інформацію про диктора пропонується використовувати для налаштування системи розпізнавання на кластери дикторів. Це дозволить підвищити точність розпізнавання за рахунок попереднього визначення кластеру дикторів і використання індивідуальної акустичної моделі для даного кластера.

Особливості анотування акустичного корпусу

Система спеціальних позначень для анотування АК:

- **позначення мови;**
- **позначення нелітературних слів;**
- **позначення способу вимови слів;**
- **позначення фону;**
- **позначення неінформаційних слів та звуків, які вимовляє диктор;**
- **позначення діалогів та хорів;**
- **позначення шуму.**

Особливості анотування акустичного корпусу

Позначення	Значення	Частота використання
у	українська мова	7490
р	російська мова	7418
р-ак	російська мова з сильним іноземним акцентом	94
с	суржик	12170
ж	жаргон, арго	3044
об	обмовка	7661
гарк	гаркавість (неправильна вимова звуків "р" і "л")	3310
см	сміх	818
е	екання	30764
хор	хор	15085
пт	шелест паперу (тло)	10816
мт	музика (тло)	30266
опт	оплески (тло)	2775
стук	стук	2191
вул.	шум вулиці	3

Перспективи

Подальші дослідження передбачають **створення інформаційно-пошукової системи з веб-інтерфейсом**, що повинна задовольняти будь-які запити користувача щодо групування та пошуку інформації в області даних. Планується розробити **програмні засоби, що автоматизують розбиття на мовленнєві сегменти синхронно з текстом**.
Незважаючи на складність і трудомісткість, ми сподіваємося створити повноцінний ресурс, що ляже в основу багатьох мовленнєвих технологій та систем, які використовуватимуться в різних сферах економіки, освіти, права, управління та в побуті. Матеріал корпусу складається з різноманітних звукових записів із їх розшифровкою і також являє собою певну основу для **створення Національного корпусу українського мовлення**, який би міг стати в один ряд із провідними розробками корпусів у світі.

Спасибі за увагу!

valya.robeiko@gmail.com