Створення акустичного корпусу українського ефірного мовлення

[12]*Н.Б. Васильєва,* [12]*В.В. Пилипенко,* [2]*О.М. Радуцький,* [12]*В.В. Робейко,* [12]*М.М. Сажок*

[1]Відділ розпізнавання та синтезу звукових образів
Міжнародного науково-навчального центру інформаційних технологій та систем;
[2]ТОВ Спеціальні реєструючі системи;
м. Київ, Україна
{ n.vassilleva, sazhok, valya.robeiko } @gmail.com; valery_pylypenko@mail.ru;
alex@srs.kiev.ua

Abstract

Title: The Ukrainian broadcast speech corpus.
In this paper a short description of activities towards building a media-based speech corpus for spoken Ukrainian language is given. Different roles and specific features of text corpus and speech corpus are investigated as well as the most frequent mistakes and misunderstandings of the concept of a speech corpus are mentioned. The concept of a big representative corpus of spoken language and its desired properties are presented. The paper gives an overview of the current state of the art in speech corpora all over the world. It explains the need for a national speech corpus and indicates some of the typical areas of research and applications taking advantage of the existence of such a corpus. The speech databases currently available in Ukraine are listed and the particularities of annotation structures of these databases are pointed out. The authors search for a general annotation structure suitable for the kind of speech corpus envisaged. Some of the basic concepts and technical solutions used in recording and computer aided annotation used for the existing speech corpora are described. The most significant problems standing in the way of building a big speech corpus are pointed out. Furthermore, a pilot version of a speech corpus is presented, containing several recordings and their orthographic transcription.
Index items: speech corpus, database, broadceast speech, spoken language, Ukrainian.

1. Вступ

Мовленнєві корпуси відіграють величезну роль у розробленні мовленнєвих інформаційних технологій. Інформація, яка міститься в корпусах, використовується при побудові акустичних та лінгвістичних моделей для систем диктування, моделей діалогу для систем взаємодії людини з машиною, моделей предметних областей для смислової інтерпретації тощо. Особливі вимоги висуваються до корпусів, що розробляються для сучасних високоякісних систем автоматичного розпізнавання мовлення та озвучення текстів. Кожний корпус створюється з певною метою, враховуючи специфіку наукових досліджень та розроблюваних прикладних систем, у яких цей корпус використовуватиметься.

Створення подібного корпусу в Україні триває вже близько року, і результатом цієї роботи є пілотна версія мовленнєвого ефірного корпусу.

Зараз робота над корпусом продовжується у двох напрямах: поповнення його новими звуковими записами та їх розміткою, а також аналіз наявних даних.

Метою даної роботи є опис засад формування корпусу, його структури, перших конкретних результатів аналізу звукового матеріалу та визначення перспектив подальших досліджень.

2. Мовленнєві корпуси в Україні та світі

Акустичний (мовленнєвий) корпус – структурована множина мовленнєвих фрагментів.

Мовленнєвий фрагмент як базова одиниця корпусу – це оцифрований фрагмент мовленнєвого сигналу, який супроводжується асоційованою інформацією певного типу (типів). Така інформація називається також анотацією до мовленнєвого фрагменту [1].

Перші акустичні корпуси були створені в першій половині 80-х років минулого століття та використовувалися для тестування й оцінки роботи систем розпізнавання мовлення на однаковому стандартному мовленнєвому матеріалі. У другій половині 80-х років минулого століття були доведені переваги систем розпізнавання мовлення на основі великих навчальних мовленнєвих корпусів, що започаткувало формування напряму мовленнєвих технологій, пов’язаного зі створенням акустичних корпусів.

Створення акустичних корпусів – доволі складна технологічна задача, яка потребує значних фінансових і кадрових ресурсів. Для її вирішення у 90-і рр. у світі були створені координаційні центри для збору, зберігання, розповсюдження та створення загальнодоступних і стандартизованих мовних ресурсів, у тому числі мовленнєвих [2]. Сьогодні подібні центри мають сотні мовленнєвих корпусів, сучасні інструментарії для обробки мовлення та створення мовленнєвих баз даних та займаються різноманітними проблемами, пов’язаними з корпусними ресурсами. Створення акустичних корпусів на даний час є самостійним і популярним напрямом мовленнєвих технологій.

В Україні довгий час створенню мовленнєвих корпусів не приділялося уваги з боку офіційних структур, тоді як, наприклад, у Словаччині з її порівняно невеликим населенням вже розроблено ряд серйозних корпусів [3]. Завдяки індивідуальній ініціативі українських наукових співробітників досить скромний матеріал все ж є доступним для підтримання експериментальних досліджень [4].

Серед цього матеріалу відзначимо україномовний багатодикторний корпус „UkReco”, що містить понад 30 тис. реалізацій фонетично збалансованих слів і тисячі речень, записаних від близько 100 дикторів, що мешкають у різних регіонах України. Цей корпус використовується в ряді досліджень по розпізнаванню ізольованих слів, адаптації на голос диктора та при побудові акустичних моделей для усного словника-перекладача [5], [6]. Інший корпус містить виконані через телевізійну мережу записи виступів близько 400 депутатів Верховної Ради України. Обсяг мовлення – біля 40 годин. Цей корпус використовується для розробки експериментальної системи розпізнавання парламентського мовлення [7]. Нещодавно був розроблений україномовний корпус опорного диктора, що містить більше 50 годин мовлення. Він використовується в дослідженнях пофонемного і поскладового розпізнавання [8].

Серед призначених для синтезу українського мовлення корпусів виділяється корпус, створений на основі жіночого голосу професійного диктора [9].

Досвід, використаний у згаданих розробках, став неоціненним при створенні концепції та власне формуванні корпусу на основі величезних масивів записів ефірного мовлення.

3. Структура та склад акустичного корпусу

Акустичний корпус українського ефірного мовлення (АКУЕМ) є загальним за метою використання акустичним корпусом, що містить читане, підготоване та спонтанне мовлення (останнє складає найбільшу частину корпусу). Усі матеріали корпусу є зв’язними та політематичними, за типом мовленнєвого сигналу поділяються на теле- та радіомовлення, також присутні невеликі вкраплення записів публічного мовлення та мовлення в природних обставинах. Основні мови матеріалів корпусу –українська та російська.

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, тому роботи над його створенням ведуться й надалі. 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 (див. Рис. 1).

На даний момент корпус українського ефірного мовлення характеризується такими кількісними показниками: понад 260 анотованих звукових записів, які становлять більше 180 годин звучання (із них для сьогоднішніх потреб розпізнавання використовується лише 50 годин українського мовлення та 55 годин російського мовлення). Словник корпусу містить майже 45 000 слів української мови та майже 50 000 слів російської мови. Проаналізоване мовлення кількох сотень дикторів різного віку, статі, соціального стану та професій. Створено словник суржику (понад 1700 слів), словник територіальних та соціальних діалектів (понад 800 слів).

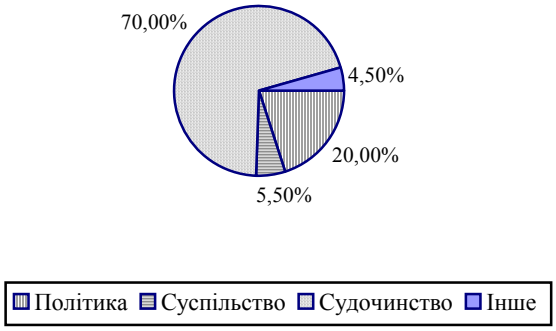

Рисунок 1. Співвідношення записів різних тематик (за тривалістю у хвилинах)

4. Розмітка корпусу

Однією з основних рис, що відрізняють акустичний корпус від звичайної колекції звукових записів чи текстів, є його розмітка (анотування) – особлива додаткова інформація про складові частини корпусу [10].

Розмітка АКУЕМ проводиться вручну на основі попередньої автоматичної розмітки. Фактично експерт переглядає вже готову анотацію, виправляючи помилки та вносячи доповнення, а також додаючи інформацію про дикторів.

Розмітка заключається в сегментуванні мовленнєвого матеріалу та детальному описі певних лінгвістичних та екстралінгвістичних явищ у сегментах. Опис відбувається за допомогою введення в текст, що відповідає звуковому сегменту, термінальних позначень типу *...*. На сьогодні використовується 74 таких позначення.

Усі позначення можна розділити на декілька груп:

• позначення мови;
• позначення нелітературних слів;
• позначення способу вимови слів;
• позначення фону;
• позначення неінформативних слів та звуків, які вимовляє диктор;
• позначення діалогів та хорів;
• позначення шуму.

Позначення мови стосується усіх наступних слів до альтернативного позначення і ставиться перед першим словом відповідної мови. В АКУЕМ на сьогодні зустрічаються записи дев’ятьох мов, хоча основний обсяг (понад 97%) належить українській та російській мовам.

Наступну групу позначень було вирішено ввести для слів, відсутніх у літературних словниках:

• суржику – суміші української та російської мов;
• соціальних діалектів (жаргонів, арго) – мови людей, зв’язаних певною спільністю фахових або соціальних інтересів;
• територіальних діалектів – мови осіб, пов'язаних між собою територіальною спільністю;
• абревіатур і скорочень.

Позначення способу вимови слів включають позначення дефектів мовлення (заїкання, гаркавість та ін.), мовленнєвих збоїв (обриви та обмовки), специфічної вимови слів (наприклад, поскладово, із

редукцією або з розтягуванням). Всі ці позначення стосуються тільки одного слова і ставляться перед відповідним словом.

Позначення неінформативних слів та звуків, які вимовляє диктор включають позначення заповнених пауз, звуків-паразитів та подібних явищ. До цієї ж групи були віднесені неінформативні звуки, на кшталт покашлювання, шмигання носом, сміху, плачу, голосного вдиху або видиху диктора. Такі позначення означають відповідні звуки в записі, які вимовляє диктор і ставляться на місці відповідного звуку. Ця група позначень є найбільшою.

Позначення діалогів та хорів відповідають місцям у звукових записах, де так чи інакше зливається мовлення кількох людей. Діалогом називаються місця, де під час розмови двох дикторів кінець фрази першого диктора накладається на початок фрази іншого диктора. Хор – повне накладання мовлення кількох людей.

Важливою групою позначень, що описують звукові сегменти корпусу, є позначення фону (тла), на якому говорить диктор, та різноманітних шумів, що присутні в сегментах. Такі позначення стосуються цілого сегмента.

Приклади позначень та частота їх використання наведена в Таблиці 1.

Таблиця 1: Деякі позначення, що використовуються під час опису сегментів АКУЕМ

Позначення	Значення	Частота використання
у	українська мова	3299
р	російська мова	3240
р-ак	російська мова з сильним іноземним акцентом	90
с	суржик	6300
ж	жаргон, арго	943
об	обмовка	3388
гарк	гаркавість (неправильна вимова звуків "р" і "л")	1191
см	сміх	369
е	екання	15030
хор	хор	6066
пт	шелест паперу (тло)	4695
мт	музика (тло)	11718
опт	оплески (тло)	1281
стук	стук	979
вул	шум вулиці	2

5. 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. Сучасні статистичні системи розпізнавання мовлення потребують великого обсягу акустичних матеріалів для побудови акустичних моделей та для тестування точності розпізнавання мовлення. На матеріалах корпусу проводяться численні наукові експерименти у галузі 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.

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.

6. 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.

6.1. Засоби стенографування звукозаписів

Стенографування виконується різновидом засобів протоколювання подій SRS-Femida [11].

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.

6.2. Засоби сегментування

Сегментування виконується засобами програмного забезпечення з відкритим кодом *Transcriber 1.5.1* [12] (див. Рис.2), 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 із 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 невпинним розвитком концепції корпусу, появою та вирішенням контроверсій. Програмний модуль *xml2mlf* дає змогу, зокрема, автоматизувати цей процес.

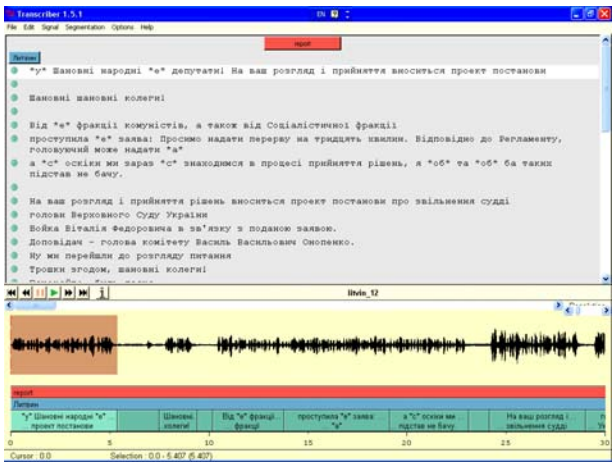

Рисунок 2. Діалогове вікно користувача програми Transcriber 1.5.1

6.3. 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.

6.4. Засоби підготовки до навчання розпізнавання мовлення

Окрім зазначеного вище, утиліта *ParseTrs* готує 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 високу точність їх визначення (біля 80%).

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.

7. Висновки

Розроблена пілотна версія АКУЕМ вже дає змогу створювати акустичні та доповнювати лінгвістичні моделі для досліджень із автоматичного транскрибування звукозаписів, пошуку ключових слів і фраз, розпізнавання диктора.

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.

8. Література

[1]. Кривнова О.Ф. Речевые корпуса на новом технологическом витке // Речевые технологии. – 2008. – № 2. – С. 13-24.

[2]. Кривнова О.Ф., Захаров Л.М., Строкин Г.С. Речевые корпусы (опыт разработки и использование) // Труды семинара Диалог'2001. – Москва, 2001.

[3]. Rusko M., Garabík R. Corpus of Spoken Slovak Language. In: Computer Treatment of Slavic and East European Languages. Bratislava, Slovakia, 2007, pp. 222-236.

[4]. http://speech.com.ua/developers.html

[5]. М. Сажок, Р. Селюх, О. Юхименко. Адаптація акустичних моделей фонем до голосу диктора для пофонемного розпізнавання ізольованих слів української мови // Штучний інтелект. – Донецьк, 2009. – № 4. – с. 230-233

[6]. M. Sazhok, V. Yatsenko. Comparison of word grammars for spoken translation in subject area // Proc. of the 13th International Conference "Speech and Computer: SPECOM'2009". – St. Petersburg, Russia, 2009. – P. 494-497.

[7]. Pylypenko V., Robeiko V. Experimental System of Computerized Stenographer for Ukrainian Speech. // Proc. of the 13th International Conference "Speech and Computer: SPECOM'2009". – St. Petersburg, Russia, 2009. – P. 67-70.

[8]. Н. Васильєва, М. Сажок. Порівняння пофонемного та поскладового розпізнавання мовленнєвого сигналу для української мови // У цьому збірнику.

[9]. Lyudovyk T., Brozinski S., Noner M., Robeiko V., Sazhok M. Speech Synthesis Applied to SMS reading // Proc. of the 13th International Conference "Speech and Computer: SPECOM'2009". – St. Petersburg, Russia, 2009. – P. 300-305.

[10]. http://www.ruscorpora.ru/corpora-intro.html

[11]. О. Радуцький, Л. Богданов. Технічна фіксація судових процесів: системний підхід до розвитку комп'ютерних технологій та інформаційних ресурсів // Юридичний журнал. – 2002. – № 2. http://www.justinian.com.ua/article.php?id=431

[12]. Barras, C., Geoffrois, E., Wu, Z., Liberman, M., 1998. Transcriber: a free Tool for Segmenting, Labeling and Transcribing Speech. In: Proc. First Int. Conf. on Language Resources and Evaluation (LREC 98), Granada, Spain, pp.1373–1376.