Технологія прискореної підготовки корпусів для навчання систем розпізнавання мовлення

[1]*В.В. Пилипенко,* [2]*О.М.Радуцький,* [1]*Т.В.Людовик*

[1]Міжнародний науково-навчальний центр інформаційних технологій та систем
40 просп. Академіка Глушкова, Київ 03680
[2]ТОВ Спеціальні реєструючи системи
valeriy.pylypenko@gmail.com, alex@srs.kiev.ua, tetyana.lyudovyk@gmail.com

Анотація

В роботі розглядається нова технологія створення корпусів мовлення для навчання автоматичних систем розпізнавання. Пропонується застосувати існуючу систему розпізнавання для формування стенограми з автоматичним сегментуванням на слова. Експерт порівнює аудіозапис з отриманою стенограмою та відмічає помилково розпізнані слова. Для подальшої роботи використовується виключно правильно розпізнаний мовленнєвий сигнал. Запропонований метод сегментації дозволяє отримати корпус без безпосереднього стенографування, що значно зменшує обсяг роботи експертів та їх кваліфікацію, що дозволяє ефективно отримати великі корпусі без залучення висококваліфікованих експертів-лінгвістів. Проведені експерименти підтвердили працездатність розробленої технології.

1. Вступ

Сучасні методи розпізнавання з використанням Прихованих Марковских Моделей (ПММ) потребують для навчання величезні корпуса мовлення. Створення акустичних корпусів – складна задача, яка потребує значних ресурсів.

Акустичний корпус відрізняється від звичайної колекції мовленнєвих аудіозаписів та їх стенограм особливою розміткою, яка включає прив'язку тексту до звуку (сегментація) та додаткову інформацію про складові частини (анотування) .

Наприклад, розмітка Акустичного корпусу АКУЕМ [1] потребувала значної ручної роботи експертів, які декілька разів прослуховували аудіо запис, переглядали відповідну стенограму, виправляли помилки в тексті. Крім того, додавалася анотація яка включала детальний опис певних лінгвістичних та екстралінгвістичних явищ у сегментах, інформацію про дикторів та інше.

Слід зауважити що створення корпусу для навчання систем розпізнавання не потребує такої детальної анотації, достатньо вказати які слова промовлялися.

Це можна зробити застенографував мовленнєві записи. Звичайно стенографіст витрачає в 4-5 разів більше часу ніж звучить аудіо запис. Крім того, людина вчиняє від 1 до 5 процентів помилок що неприпустимо для навчання розпізнаванню мовлення. Виправлення помилок досить трудомісткий процес набагато повільний ніж безпосереднє стенографування.

В [2,3,4] пропонована та розвинута технологія Lightly Supervised Training (LST) де система розпізнавання мовлення використовується для виявлення помилок стенографування. Ця технологія дозволяє задіяти для навчання стенограми, не повністю співпадаючі з промовлянням [5].

2. Технологія автоматизованого формування корпусів мовлення.

В роботі розглядається технологія створення корпусів для навчання де існуюча система розпізнавання задіяна для попереднього стенографування аудіо записів для отримання стенограми чернетки.

Після отримання стенограми чернетки експерт прослуховує аудіозапис та відмічає помилки розпізнавання, наприклад, іншим кольором в звичайному редакторі текстів. Як доводять експерименти цей процес потребує ненабагато більш часу (приблизно на 20-30%) ніж відтворення звуку записи.

Потім всі слова відмічені як помилки автоматично видаляються з навчального

корпусу. Отримані дані цілком придатні для навчання систем розпізнавання мовлення.

Наступні розділи присвячені опису експериментального комплексу для навчання та розпізнавання мовлення в якому використовується пропонована технологія.

3. Експериментальний комплекс для навчання та розпізнавання мовлення

Експериментальний комплекс побудовано на базі відкритого ПЗ НТК [6] якій був модифікований для роботи с кириличними алфавітами. Комплекс включає сервер автоматичного розпізнавання для формування стенограм, робочі міста експертів для прослуховування записів та корекції стенограм (SRS-Femida), а також спеціалізоване ПЗ для автоматичного навчання.

4. Препроцесинг мовленнєвого сигналу

Мовленнєвий сигнал перетворюється у послідовність векторів ознак із інтервалом аналізу 25 мс і кроком аналізу 10 мс. Спочатку мовленнєвий сигнал фільтрується фільтром високих частот із характеристикою $P(z) = 1 - 0.97z^{-1}$. Потім застосовується вікно Хеммінга і обчислюється швидке перетворення Фур'є. Спектральні коефіцієнти усереднюються з використанням 26 трикутних вікон, розташованих на мел-шкалі, і обчислюються 12 кепстральних коефіцієнтів.

Логарифм енергії додається у якості 13-го коефіцієнта. Ці 13 коефіцієнтів розширюються до 39-вимірного вектора параметрів через додавання першої та другої різниць сусідніх у часі коефіцієнтів. Для врахування впливу каналу застосовується віднімання середнього значення кепстра.

5. Акустична модель

У якості акустичних моделей застосовуються приховані Марківські моделі. 51 російська контекстно-залежна фонема (включно з фонемою-паузою) моделюються трьома станами Марківського ланцюга без пропусків. Використовується діагональний вигляд Гауссівських функцій щільності ймовірності.

Для моделювання варіантів фонем в залежності від правого та лівого контексту в злитому мовленні застосовуються трифонні моделі фонем. Усього в навчальної вибірці та в словнику розпізнавання зустрілося 18114 трифонів.

Трифони, що зустрічаються рідко, моделюються 16 сумішами Гауссівських функцій щільності ймовірності. Фонеми, які зустрічаються частіше, моделюються більшою кількістю сумішей, для найуживаніших фонем застосовується 128 суміші.

6. Текстовий корпус

Словник був створений із текстів стенограм засідань Вищого Арбітражного суду Російської Федерації (ВАСРФ). З офіційного сайту ВАСРФ були завантажені стенограми засідань, що становить понад 50 МБ тексту. Усі тексти стенограм були модифіковані для того, щоб уникнути службової інформації, записати числа в текстовому вигляді.

Для лінгвістичного корпусу був створений словник із 220 тис слів і обчислена частота вживаності кожного слова зі словника. Майже 98% із них мають частоту вживання 3 і більше (такі слова утворюють словник на 59 тис. слів.).

7. Лінгвістична модель

Біграмна та триграмна моделі мови, які описуються ймовірностями появи пар та трійок слів, дозволяють значно поліпшити точність розпізнавання злитого мовлення. Оскільки в текстах, на основі яких обчислювалися статистики, зустрілися не всі сполучення слів, можливі для даного словника, то для апроксимації неспостережених пар слів застосуються зворотні (back off) коефіцієнти.

8. Перетворення буква - фонема

Словник транскрипцій був створений автоматично на основі орфографічного словника з використанням контекстно-залежних правил Для перетворення орфографічного тексту у фонемний був сформований набір контекстно-залежних правил, за якими орфографічне слово перетворюється на послідовність фонетичних символів (шляхом перетворення однієї послідовності символів на іншу).

При цьому було породжено декілька (у середньому 1.5) варіантів транскрипцій. Найбільш споживані слова мали до 10 варіантів транскрипцій що відображає варіативність промовляння слів різними дикторами [7].

9. Навчальна вибірка

Навчання проводилося з використанням аудіозаписів суддів Вищого Арбітражного суду Російської Федерації які на сайті ВАСРФ.

Якість записів висока, оскільки застосується конференц-система, яка забезпечувала гарантований запис тільки одного виступаючого.

Суддівське мовлення характеризується певними особливостями, а саме:

- Спонтанне мовлення (зустрічаються окремі доповіді, прочитані з підготованого заздалегідь тексту);
- Використовується специфічна лексика;
- Монологи. Записи в основному складаються з монологів дикторів, але в них можуть зустрічатися репліки голови судового засідання.

Для навчання базової системи задіяні записи тривалістю 120 годин мовлення. Всього в цих записах зустрілося понад 359 дикторів. Дикторів із тривалістю запису понад 300 сек. виявилося 204.

Навчання проводилося на відрізках звукового сигналу з кількох слів, обмежених паузами більшими за 400 мс. Для кожного відрізку спів зіставлялася автоматично розпізнана стенограма, в якої експерти відмічали помилково розпізнані слова. Для усіх помилково розпізнаних слів застосувалася єдина модель “слово-помилка”. Потім текст автоматично перетворювався на послідовність фонем. Вибірка, розмічена в такий спосіб, використовувалася для побудови акустичної моделі

10. Контрольна вибірка

Розпізнавання проводилося на записах мовлення суддів, зроблених у відмінні від навчальної вибірки дні. Стенограма записів була ретельно перевірена експертами. Обсяг звукового сигналу контрольної вибірці сягав 11 годин мовлення, у яких зустрілося близько 80 тис. слів.

Всього в КВ зустрілися записи 103 дикторів. Виявилося 49 дикторів із тривалістю запису понад 300 сек.

11. Результати експериментів

Для порівняння використовується корпус мовлення побудований за LST технологією [5] тривалістю в 120 годин мовлення. Для стенографування корпусу знадобилося приблизно 500 годин роботи експертів.

Показники розпізнавання наведені в НТК форматі:

*WORD: %Corr=**70.79**, Acc=**60.8**4*

[H=56089, D=3412, S=19728, I=7883, N=79229],

де N – кількість розпізнаних слів, I – кількість помилкових вставок слів, S – кількість замін слів, D – кількість віддалених слів, H – кількість слів, співпадаючих з еталоном, Acc = 100*(H-I) /N – точність розпізнавання, Corr = 100* H/N – “правильність” розпізнавання.

Обсяг корпусу мовлення збудованого за пропонованою прискореною технологією сягає 180 годин мовлення. Експерти витратили близько 230 годин на прослухування та на відмітку помилкових слів. Таким чином, затрати часу роботи експертів понад втричі менш ніж за попередньою технологією LST.

Показники розпізнавання наведені в НТК форматі:

*WORD: %Corr=**72.23**, Acc=**61.89***

[H=57254, D=2674, S=19335, I=8201, N=79263]

Поліпшення точності розпізнавання більшою мірою сталося за рахунок збільшення

Таблиця 1: Точність розпізнавання для деяких найкращих дикторів

Диктор	Кількість слів у контрольній вибірці	Точність (%)
Докладчик Петрова С.М	636	81.29
Председатель	6884	80.36
Должник	895	74.64
Новоселова Л.А	860	72.56
Гвоздилина О.Ю	859	72.29
Балахничева Р.Г	1530	71.11
Кочеткова И.В	873	70.90
Куликова В.Б	1538	70.68
Заявитель	2537	70.08
Разумов И.В	1916	70.04
Ответчик	1272	68.08

Таблиця 2: Точність розпізнавання для деяких найгірших дикторів

Диктор	Кількість слів у контрольній вибірці	Точність (%)
ООО «Лесное озеро»	1714	52.04
Истец	996	52.01
Горячева Ю.Ю.	989	50.86
Чекулаев Дмитрий Петрович	1041	48.70
ОАО «Коммерческий банк Петрокоммерц»	971	47.99
Шичанин А.В	1336	47.83
Участница Нешатаева Т. Н.	511	46.58
Сергунин А.К.	1708	43.15
Водовозов А.А.	678	40.12
Общество с ограниченной ответственностью «Сабмиллер рус»	555	32.07

обсягу навчальної вибірці.

Наведені таблиці показують точність розпізнавання для деяких найкращих (таб. 1) та найгірших дикторів (таб. 2).

12. Висновки

Запропонована технологія підготовки аудіозаписів з використанням автоматичних систем розпізнавання дозволила отримувати корпус мовлення в 3-4 разі швидше ніж традиційні технології сегментування без зниження точності.

Таким чином, була підтверджена працездатність розробленої технології.

Безумовно в навчальний корпус не потрапляють помилково розпізнанні слова, серед яких зустрічаються корисні для побудови моделей фонем.

В майбутньому планується побудувати звуковий редактор для стенографування тільки помилково розпізнаних слів.

Література

1. Pylypenko V., Robeiko V., Sazhok M., Vasylieva N., Radoutsky O. Ukrainian Broadcast Speech Corpus Development. Proceedings of SPECOM 2011, 14-th International Conference “Speech and Computer”. Kazan, Russia, 2011, pp. 435–440.
2. L. Lamel, J.L. Gauvain, and G. Adda. Lightly Supervised Acoustic Model Training. In ISCA ITRW Workshop on Automatic Speech Recognition: Challenges for the new Millenium, pages 150–154, Paris, 2000.
3. L. Lamel, J.L. Gauvain, and G. Adda. Lightly supervised and unsupervised acoustic model training. Computer Speech and Language, 16(1):115–229, 2002.
4. Matthias Paulik, Alex Waibel: Lightly supervised acoustic model training on EPPS recordings. INTERSPEECH 2008: 224-227.
5. Пилипенко В.В., Технология разметки звуковых файлов с использованием неточного текстового сопровождения, Кибернетика и вычислительная техника, 2012, вып. 169.
6. Young S. et al. The HTK Book (for HTK Version 3.4). Cambridge, UK, 2009
7. Tetyana Lyudovyk, Valeriy Pylypenko. Code-switching speech recognition for closely related languages. The 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14), St. Petersburg, Russia, 2014, pp.188-193