Експериментальна система автоматизованого стенографування українського мовлення

Пилипенко В.В., Робейко В.В.

Міжнародний науково-навчальний центр
інформаційних технологій та систем, Київ
valery_pylypenko@mail.ru, valya.robeiko@gmail.com

This paper presents a research system of computerized stenographer. It makes the text from sound records based on the speech recognition system aided by human. Large vocabulary (more than 10K words) continuous speech recognition system for a number of speakers is used to process recorded files. Human introduces out-of-vocabulary words and repairs errors to produce the final text. Personal phonetic rules are listed and used to individualize transcriptions for speakers. To improve system performance the retraining process is running to take into account repairs. Experimental recognition results are presented.

1. Вступ

Стенографування широко використовується для документування матеріалів засідань і нарад різного рівня, для роботи секретарів і журналістів тощо. Комп'ютери значно розширили можливості й дозволили збільшити гнучкість застосування систем стенографування. Сьогодні актуальною задачею є зменшення частки ручної праці в таких системах. Для цього пропонується використовувати автоматичне розпізнавання мовлення для перетворення звуку в текст.

Мовлення кожної людини – індивідуальне. Тому перетворити звук на текст, натиснувши одну кнопку, – завдання доволі складне для системи стенографування. Така система повинна максимально спростити роботу оператора й пришвидшити перетворення звукового файлу в текстовий, а також врахувати всі особливості мовлення диктора. Існує багато програмно-апаратних комплексів автоматизованого стенографування з різними можливостями, але навіть найпростіший дозволяє збільшити швидкість перетворення звуку в текст у кілька разів.

Автоматичне розпізнавання злитого мовлення багатьох дикторів із великих словників значно спрощує роботу оператора, дозволяючи лише виправляти помилки системи стенографування. Донавчання системи скорочує кількість помилок у процесі експлуатації.

2. Автоматизована vs автоматична

Системи стенографування можна умовно поділити на три категорії залежно від співвідношення участі людини й комп'ютера у процесі створення стенограм:

- **автоматичні** (без участі людини в процесі розпізнавання мовлення);
- **автоматизовані** (людина так чи інакше бере участь у процесі розпізнавання мовлення комп'ютером);
- **стенографування** за допомогою комп'ютера (людина набирає текст, а комп'ютер використовується як магнітофон і друкарська машинка).

Різниця між **автоматичною** й **автоматизованою** системами полягає в надійності автоматичного розпізнавання мовлення.

Досвід експлуатації говорить, що початкова стенограма, створена людиною, містить певну кількість помилок, які виправляються у процесі редагування тексту. У середньому кількість помилок – 5 на одну сторінку тексту, отже надійність стенографування – 98%, оскільки на одній сторінці міститься приблизно 2 тис. знаків або 250 слів. Таким чином система стенографування стає **автоматичною** при надійності розпізнавання мовлення більше 98%.

Така надійність сьогодні може бути досягнена для автоматичного розпізнавання мовлення за деяких обмежень. При цьому розпізнається мовлення тільки одного диктора. Для ізольовано вимовлених слів словник досягає 15 тис. слів, а для злитого мовлення така надійність досягається за словника в 1 тис. слів.

Тому на даний момент актуальним є створення програм розпізнавання мовлення, вільних від таких обмежень. Для стенографування необхідно досягнути об'єму словника від 10 тис. слів до декількох мільйонів. Кількість задіяних дикторів – від 100 до 1 тис. При цьому повинне розпізнаватися злите мовлення в реальному часі для сучасних комп'ютерів.

Автоматизованою систему є сенс називати за надійності 80% і вище. Оператор повинен буде виправляти не більше, як кожне п'яте слово в тексті, що можна робити під час прослуховування звукової доріжки в процесі її відтворення.

3. Система розпізнавання злитого мовлення

У даній роботі як базова система використовується інструментарій HTK [1] на основі прихованих Марківських моделей. Інструментарій HTK використовується для побудови акустичних і лінгвістичних моделей. Для розпізнавання мовлення був розроблений програмний комплекс, сумісний із акустичними й лінгвістичними моделями HTK.

3.1. Інтерфейс користувача програми

Інтерфейс користувача програми зображений на рисунку 1. У верхньому вікні схематично зображається осцилограма звукової доріжки з автоматично виділеними сегментами мовлення (фразами або синтагмами). Оператор виділяє потрібний йому сегмент і прослуховує його. Оператор також має можливість подивитися відповідь розпізнавання, яку можна виправити у випадку помилки. Після редагування відповідь додається до стенограми й автоматично відбувається перехід до

наступного сегмента. Користувач може перейти до записів потрібного диктора, прослухати необхідний уривок стенограми.

Розпізнавання проводиться автоматично в фоновому режимі роботи програми. Усі помилки розпізнавання фіксуються, і після того, як закінчилося формування стенограми, відбувається донавчання системи стенографування. При цьому в навчальну вибірку додаються нові слова й інформація про нових дикторів. Таким чином надійність розпізнавання зростає у процесі експлуатації системи стенографування.

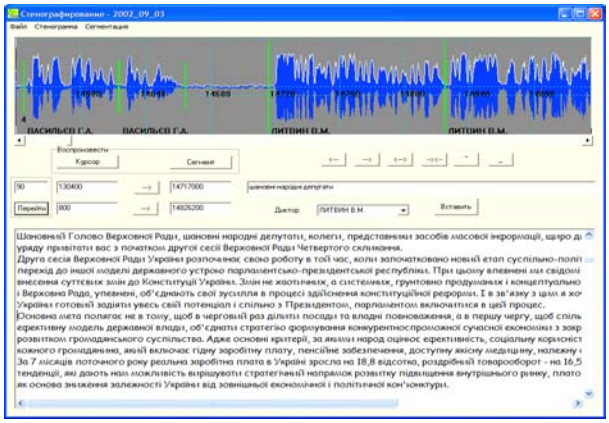

Рис. 1: Загальний вигляд програми стенографування

3.2. 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.

3.3. Акустична модель

У якості акустичних моделей використовуються приховані Марківські моделі. 56 українських контекстно-незалежних фонем (включно з фонемою-паузою) моделюються трьома станами Марківського ланцюга без пропусків. Використовується діагональний вигляд Гауссівських функцій щільності ймовірності.

Фонеми, що зустрічаються рідко, моделюються 64 сумішами Гауссівських функцій щільності ймовірності. Фонеми, які зустрічаються частіше, моделюються більшою кількістю сумішей, для найуживаніших фонем використовується 1024 суміші.

Словник транскрипцій створюється автоматично на основі орфографічного словника з використанням контекстно-залежних правил.

3.4. Багатодикторна система

Розпізнавання мовлення незалежно від диктора – не вирішена задача в розпізнаванні мовлення. У [2] використовувалася модель кооперативного розпізнавання, у якій у процесі навчання змішувалося мовлення різних дикторів. Крім того, мовлення різних дикторів розглядалося як різні реалізації мовлення одного диктора. Було доведено, що надійність розпізнавання покращилась не тільки для дикторів із кооперативу, але й для дикторів, незнайомих системі. Це може бути пов'язане з тим, що мовлення знайомих системі дикторів схоже на мовлення інших дикторів.

Застосування такого підходу довело, що у процесі використання більше 100 дикторів у кооперативі надійність розпізнавання мовлення стає дуже подібною до системи, незалежної від диктора.

Методи роботи з багатьма дикторами, закладені в інструментарій HTK, такі як нормалізація довжини мовленнєвого тракту (Vocal Tract Length Normalization) й адаптація моделі за допомогою лінійного перетворення максимальної подібності (Maximum Likelihood Linear Regression) дозволяють покращити надійність розпізнавання мовлення для окремих дикторів за умови, що якимось незалежним способом заздалегідь визначається диктор. Передбачається використовувати методи ідентифікації дикторів для автоматичного визначення мовця.

4. Акустичне та текстове наповнення

4.1. Навчальна вибірка

Навчання проводилося на мовленні депутатів Верховної Ради України, записаного через телевізійну мережу. Парламентське мовлення характеризується деякими особливостями:

- Це спонтанне мовлення. Зустрічаються окремі доповіді, прочитані з підготованого заздалегідь тексту, але лише дехто з дикторів дотримується цього тексту;
- Через обмеження в часі мовлення багатьох дикторів має дуже швидкий темп;
- Часто мовлення емоційно забарвлене;
- В основному записи складаються з неперервних монологів дикторів, але в них можуть зустрічатися репліки голови засідань або інших депутатів;
- Якість записів досить висока, оскільки кожне депутатське місце забезпечене мікрофоном.

Для навчання використовувалися записи тривалістю 99 тис. сек., у яких зустрілося 211 224 слова. Всього було записано 208 дикторів. Дикторів із тривалістю запису понад 300 сек. виявилося 87.

Навчання проводилося на попередньо розміченій вибірці. Для цього записи автоматично розбивалися на фрази з кількох слів, обмежені паузами більшими за 400 мс. Для кожної фрази створювалася мітка у вигляді тексту зі стенограми. Потім текст автоматично перетворювався на послідовність фонем відповідно до

контекстно-залежних правил. Вибірка, розмічена в такий спосіб, використовувалася для побудови акустичної моделі.

4.2. Контрольна вибірка

Розпізнавання проводилося на записах мовлення депутатів, зроблених у відмінні від навчальної вибірки дні. Для розпізнавання використовувалися записи тривалістю 30 тис. сек., у яких зустрілося 68 819 слів. Всього використовувалися записи 118 дикторів. Дикторів із тривалістю запису понад 300 сек. виявилося 37. Записів 36 дикторів не було в навчальній вибірці. Отже, ці диктори були невідомі для системи розпізнавання.

4.3. Текстовий корпус

Словник був створений із текстів стенограм засідань Верховної Ради України. Із офіційного сайту Верховної Ради були завантажені всі стенограми засідань, починаючи з 1991 р., що становить понад 100 МБ тексту. Усі тексти стенограм були модифіковані для того, щоб уникнути службової інформації (наприклад, повідомлення про аплодисменти чи вигуки), записати числа в текстовому вигляді, а також відокремити український текст від російського.

Потім матеріал був поділений на дві частини – перша містить усі тексти, крім 2002-2003 рр. (складається з 14 629 111 слів), друга включає стенограми 2002-2003 рр.(409 244). 99,6% слів із нового тексту (друга частина) вже наявні у першому словнику.

Для першої частини був створений словник із 156 108 слів і обчислена частота вживаності кожного слова зі словника. Майже 95% із них мають частоту вживання 50 і більше (такі слова утворюють словник на 15 тис. слів.).

Досліджувалася надійність розпізнавання залежно від обсягу частотного словника із використанням біграмної моделі мови. Словника обсягом 15 тис. слів достатньо для розпізнавання мовлення з невеликим (2%) зменшенням надійності в порівнянні з максимально можливим розміром словника.

5. Біграмна модель мови

У процесі розпізнавання мовлення використовувалася біграмна модель мови, яка задавалася ймовірностями появи пар слів. Оскільки в текстах, на основі яких обчислювалися статистики, зустрілися не всі пари слів, можливі для даного словника, то для апроксимації *неспостережених* пар слів використовувалися зворотні (back off) коефіцієнти [1].

Таблиця 1: Приклади виправлення помилок розпізнавання за допомогою біграмної моделі мови

Фраза	Вільний порядок слів	Біграмна граматика
доброго ранку	до ***в в о*** ранку	доброго ранку
прошу займати робочі місця	прошу ***з е мав те й*** **робоче *й*** місця	прошу займати робочі місця
прошу підготуватися до реєстрації	прошу **б й** готуватися до реєстрації **б**	прошу підготуватися до реєстрації

Біграмна модель мови дозволила виправити багато помилок розпізнавання (у таблиці 1 продемонстровані приклади таких виправлень).

6. Індивідуалізація транскрипцій

Для перетворення орфографічного тексту у фонемний був розроблений режим розбору орфографічного тексту й сформований набір контекстно-залежних правил, за якими орфографічне слово перетворюється на послідовність фонетичних символів (шляхом перетворення однієї послідовності символів на іншу). При цьому генерується відразу декілька варіантів транскрипцій для випадків неоднозначностей, заданих у правилах.

Для всіх дикторів був створений загальний варіант транскрибування. Крім цього всі диктори були розподілені на групи, для яких розроблені свої правила індивідуалізованого транскрибування, які доповнюють або заміняють основний варіант.

Результати вивчення мовлення багатьох дикторів свідчить, що ніхто з них не притримується орфоепічних правил вимови у повному обсязі. У першу чергу це стосується заборонених літературною нормою регресивної асиміляції за глухістю в парі фонем „дзвінка+глуха" й оглушення приголосних перед паузою (**тобто → т о п т о**; **підтримати → п' і т т р И м а т и**; **робив → р о б И ф**). Диктори з такими особливостями вимови були виділені в окрему групу.

Були виокремлені і багато інших характерних рис вимови різних дикторів: редукція закінчень деяких слів (прикметників, дієслів) у злитому мовленні (**шановний → ш а н О в н и; доброго → д О б р о**), „акання" (**робити → р а б И т и**), тверда вимова м'яких приголосних (**синього → с И н о г о**) та ін.

Для деяких слів (службових частин мови, слів із різними наголосами, наприклад) задається декілька варіантів транскрипцій – із наголосом на різних складах (якщо в мові можливі різні варіанти прочитання таких слів) або взагалі без наголосу: **коли → к о л И ; к О л и ; к о л и**.

Такі тенденції моделюються шляхом зміни правил переходу від одних послідовностей символів до інших і розширенням існуючих правил.

Усі правила індивідуалізованої модифікації транскрипцій можна розділити на кілька груп.

До позиційних змін звуків у потоці мовлення (змін, що залежать від загальних фонетичних умов – позиції в складі/слові, наголошеності/ненаголошеності тощо [3]) зараховуємо:

1) окрім редукції ненаголошених *е*, *и* та *о* до e^{u}, u^{e}, o^{y}, також ослаблену вимову *о* як *а* в ненаголошеній позиції, рідше трапляється редукція ненаголошених голосних до повного зникнення (**тепер → т и п Е р, зозуля → з у з У л' а, боротьба → б а р а д' б А** або **б р а д' б А**);
2) оглушення дзвінких приголосних перед паузою (**брід → б р' І т, зараз → з А р а с**);
3) редукцію у термінальних частинах слів у процесі мовлення – зникнення приголосного звука в закінченнях *-ого*, *-их*, *-ий*, *-іх*, *-ій*, -її, *-ої*, *-еї*, *-ою*, *-єю*, *-ими* та подібних (**коротший → к о р О ч ш и, синіх → с И н' і, безпекою → б е с п Е к о у**); зникнення

кінцевого голосного звука в закінченнях *–ою*, *-ею*, *-єю* та подібних (**доброю → д О б р о й, землею → з е м л Е й**) та ін.

До комбінаторних змін (якісні та кількісні зміни сусідніх звуків [3]) відносимо:

1) повну регресивну асиміляцію за глухістю у сполуці „дзвінкий+глухий" на межі будь-яких морфем у слові та на межі слів (**без причини → б е с п р и ч И н и, розсунути → р о с с У н у т и , книжка → к н И ш к а, сядьте → с' А т' т' е**);
2) асиміляцію за м'якістю свистячих та шиплячих приголосних, губних та задньоязикових приголосних (**злі → з' л' І, шлях → ш' л' А х, квітка → к' в' І т к а**);
3) вимову подовжених і подвоєних приголосних звуків як одного звука, вимову двох голосних як одного звука (**віддати → в' і д А т и, знання → з н а н' А, зоопарк → з о п А р к, аеропорт → а р о п О р т**);
4) неповне спрощення в групах приголосних (**чесний → ч Е с т н и й**) та ін.

7. Результати розпізнавання злитого мовлення

Експерименти проводились на описаній контрольній вибірці у вигляді записів засідань протягом одного дня. У таблиці 2 наведені результати розпізнавання записів для різних днів. Надійність розпізнавання значно відрізняється для різних дикторів. Наприклад, у вибірку DAY2 потрапила велика доповідь диктора SHL, промовлена у занадто швидкому темпі. Застосування індивідуалізованих транскрипцій дозволило покращити надійність розпізнавання майже на 1%.

Таблиця 2: Надійність розпізнавання для деяких вибірок

Вибірка	Тривалість контрольної вибірки (сек.)	Надійність для загальної транскрипції (%)	Надійність для індивідуалізованої транскрипції (%)	Покращення (%)
DAY1	1 849	71,23	72,66	1,43
DAY2	5 374	61,97	63,05	1,08
DAY3	10 032	68,16	68,73	0,57
DAY3a	5 990	68,69	69,75	1,06
DAY4	7 260	76,66	77,13	0,47
Всього	**30 505**	**69,28**	**70,06**	**0,78**

Таблиця 3 наводить результати розпізнавання вибірки DAY4 для деяких дикторів. Аналіз результатів демонструє, що на надійність суттєво впливають такі фактори, як тривалість навчальної вибірки й темп мовлення кожного диктора.

Надійність розпізнавання для різних дикторів варіюється від 50% до 90%.

Остання колонка показує зміну надійності за умови застосування індивідуалізованих транскрипцій. Для деяких дикторів надійність погіршується, отже для них потрібно створювати інші транскрипції.

Час розпізнавання для комп'ютера Pentium 2GHz становить близько 10 сек. для однієї секунди мовлення. Застосування алгоритмів прискорення прийняття рішень [4] дозволить досягнути реального часу розпізнавання мовлення.

8. Висновки

У статті описується експериментальна система автоматизованого стенографування. Продемонстрована можливість побудови таких систем за умови покращення розпізнавання мовлення до рівня, необхідного для його практичного застосування. Запропоновано використовувати індивідуальну інформацію про дикторів для покращення надійності розпізнавання.

9. Література

[1] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, P. Woodland. *The HTK Book,* Cambridge University Engineering Department, 2002.

[2] Винцюк Т.К. *Анализ, распознавание и смысловая интерпретация речевых сигналов*. – Киев, Наукова думка, 1987. – 264 с.

[3] *Сучасна українська літературна мова. Фонетика: Навч. посібник для студентів-філологів*. – К.: Видавничо-поліграфічний центр „Київський університет", 2002. – С 60.

[4] Пилипенко В.В. *Распознавание дискретной и слитной речи из сверхбольших словарей на основе выборки информации из баз данных*. // Искусственный интеллект, 2006. – № 3. – С. 548-557.

Таблиця 3: Надійність розпізнавання для деяких дикторів

Диктор	Тривалість навчальної вибірки (сек.)	Тривалість контрольної вибірки (сек.)	Кількість слів у контрольній вибірці	Темп мовлення (слів/сек.)	Надійність для загальної транс крипції (%)	Надійність для індивідуалізованої транс крипції (%)	Покращення (%)
LIT	15 805	2 336	5 721	2,45	79,85	80,56	0,71
POR	3 715	411	853	2,08	80,30	80,54	0,24
MOR	1 728	362	950	2,62	70,74	71,47	0,73
SIM	1 484	125	255	2,04	80,00	80,78	0,78
MAT	1 305	174	292	1,68	80,14	77,05	-3,09
KLU	998	107	209	1,95	86,60	89,0	2,40
KIN	585	223	417	1,87	64,27	66,43	2,16
ONI	483	100	209	2,09	79,90	80,38	0,48
MIS	195	148	312	2,11	69,87	69,23	-0,64