

Моделювання пофонемного розпізнавання мовленнєвого сигналу для ПК та мікропроцесорів ЦОС

Микола Сажок, Ніна Васильєва, Руслан Селюх, Дмитро Федорин

Міжнародний науково-навчальний центр інформаційних технологій та систем
40 просп. Академіка Глушкова, Київ 03680

Електронна пошта: { mykola, ninel, selyukh, fedoryn }@uasoiro.org.ua

Mykola Sazhok, Nina Vasyliieva, Ruslan Selyukh, Dmytro Fedoryn. Phoneme-based speech recognition modeling for PC and DSP. The aspects of phoneme-based recognition are analysed for both PC and DSP. The benefits of context-dependent models are spotted. The description of knowledge and data base speech for speech recognition system evaluation is given. Experimental results are discussed.

1 Вступ

В рамках генеративної моделі розпізнавання та розуміння мовленнєвого сигналу пофонемне розпізнавання, на відміну від послівного, дає змогу розширювати робочий словник без навчання на нові слова [1]. Крім того, з'являється можливість досліджувати дрібніші елементи ніж слово, такі як склади та морфеми. З використанням словника із цих елементів охоплюється ширший лексикон без збільшення обсягів словника [4]. Також ми впритул підходимо до реалізації підходу багатозначного багатозначного розуміння мовленнєвого сигналу, що є найбільш продуктивним при створенні систем диктування та систем усного діалогу для ряду надзвичайно флективних мов з відносно вільним порядком слідування слів, до яких відносяться і слов'янські мови [5]. Цей підхід ще є можливістю розподілити науково-дослідну роботу між експертами в акустиці, фонетиці, лінгвістиці та інформатиці.

Для пофонемного розпізнавання базовим є вибір акустичної моделі фонемного рівня. Потрібно визначитися з топологією моделі та способом врахування коартикуляції. Для акустичної моделі кожної фонемі з алфавіту фонем фіксуються кількість станів, допустимі переходи між станами, спосіб опису області, яку стан апроксимує, в умовах обраного простору ознак. При виборі контекстно-залежної моделі типово зупиняються на фонематрифонах [2], які враховують коартикуляцію безпосередньо прилеглих звуків, хоч можливим є і більш широкий контекст.

Наступним суттєвим кроком є моделювання допустимих послідовностей фонем шляхом побудови грамастик на основі словника, в якому задаються транскрипції слів або дрібніших елементів слова аж до фонемі. В останньому випадку отримаємо модель автоматичного фонемного стенографа. Вільна граматика не задає жодних обмежень на слідування елементів словника. Клас обмежених грамастик дуже широкий, і чим більше обмежень накладаємо, тим складніший граф обробляється розпізнавачем.

Окрім високої надійності (не менше 95%), до систем розпізнавання часто висувається вимога надання відповіді розпізнавання в реальному часі. Цей аспект теж повинен враховуватися при реалізації пофонемного розпізнавання як на персональному комп'ютері (ПК), так і в портативних пристроях, що розробляються на основі мікропроцесорів цифрового оброблення сигналів (ЦОС). Останні значно поступаються ПК у швидкодії та обсягах оперативної пам'яті. Також повинні враховуватися особливості архітектури мікропроцесорів ЦОС, такі як доступ до оперативної пам'яті та фіксована кома.

У цій статті ми зробимо підсумок проведених в рамках ДНТП "Образний комп'ютер" досліджень моделей пофонемного розпізнавання на ПК та на прототипах портативних пристроїв.

У Розділі 2 ми характеризуємо базу даних і знань, яка використовується при оцінюванні параметрів акустичних і лінгвістичних моделей. У Розділі 3 розглядаються моделі пофонемного розпізнавання. Розділ 4 присвячено лінгвістичним моделям або граматикам. У Розділі 5 описуються особливості реалізації пофонемного розпізнавання на мікропроцесорах ЦОС.

2 База даних і знань

База даних і знань включає україномовний мовленнєвий корпус для оцінки параметрів акустичних і лінгвістичних моделей та для формування лексикону.

Ми використовували україномовний багатодикторний мовленнєвий корпус, який містить понад 30 000 реалізацій слів і тисячі речень близько 100 дикторів, що мешкають у різних областях України. Реалізації зберігають частотні пропорції фонем і є фонетично збалансованими [4].

Лексикон містить близько 2 мільйонів словоформ, яким відповідають 151 000 основних форм (лем). Фактично, цим лемам відповідає більше 3 мільйонів словоформ, але у багатьох з них однакова орфографія і вимова [4].

На основі лексикону та текстового корпусу обсягом 250 МБ було згенеровано частотний словник на 157 000 слів, який використовувався для формування навчальної та контрольної вибірок.

До бази знань також входять правила відображення узагальнених символічних послідовностей для перетворень між фонемним і орфографічним текстами. Загалом таких правил не більше 30.

3 Акустичні моделі фонем

Слова, що вимовляються диктором, є послідовністю реалізацій фонем – найменших смислороздільних звуків мови. Вважаємо заданим алфавіт фонем у вигляді скінченної множини K , куди входять фонем $k \in K$, які спостерігаються в природній мові [2]. До алфавіту включено також фонему-паузу #.

Так у множині K для української мови розрізняємо наголошені та ненаголошені голосні, м'які та тверді приголосні: $k \in \{a, o, y, e, и, i, A, O, Y, E, И, I, б, б', в, в', г, г', г', д, д', ж, ж', з, з', й, к, к', л, л', м, м', н, н', п, п', р, р', с, с', т, т', ф, ф', х, х', ц, ц', ч, ч', ш, ш', дз, дз', дж, дж', \#\} \equiv K$ – загалом 57 фонем.

В багатьох випадках неможливо встановити точний початок і закінчення фонем. Це ми можемо спостерігати на рис. 1 а) і б). Тому часто говорять про ймовірність перебування в тій чи іншій фонемі $k \in K$ в дискретний момент часу t_i $P(k/t_i)$, причому

$$\sum_{k \in K} P(k/t_i) = 1.$$

Рис. 1 а) ілюструє автоматичну сегментацію на фонем з кроком аналізу 10 мс. Початок інтервалу аналізу, що є найближчим до t_j , такого що $P(k_1/t_j) = P(k_2/t_j)$, оголошується границею між фонемами k_1 і k_2 .

При акустичному моделюванні фонем неодмінно стикаємося з проблемою врахування мінливості фонем, що головним чином пов'язано з такими чинниками як темп вимовлення, тембр голосу, інтонація та взаємовплив звуків у потоці мовлення.

Рис. 1 а) – в) ілюструє наскільки є різними осцилограми реалізацій фонем о та л для одного і того ж диктора (жіночий голос). Спостерігаємо суттєво відмінні форми одноквазіперіодичних

мікросегментів, що спричинено коартикуляцією та різною довжиною мікросегментів, яка залежить від інтонації (поточної частоти основного тону).

Нелінійні зміни темпу апроксимуються повторюванням або пропуском станів еталонів фонем. Допустимі повторювання або пропуски задаються матрицею переходів між станами.

Стани наповнюються середнім та дисперсією множини відповідних векторів з обраного простору первинних ознак сигналу. Для більш детального опису простору, в якому спостерігаються вектори, що належать станам, вводяться додаткові моди через суміш нормальних законів (гаусіанів). На рис. 2 зображено проекції розподілів для кожного компонента вектора ознак – свого роду відбиток еталону фонемі.

Розглядалися способи врахування коартикуляцію шляхом використання контекстно-залежної моделі розпізнавання та за рахунок збільшення гаусіанів.

В якості контекстно-залежної моделі нами використовувалися фонемно-трифонна модель розпізнавання та синтезу мовленнєвого сигналу. В якості звукового образу фонемного рівня бралась фонема-трифон, тобто фонема в контексті з попередньою та наступною фонемами.

Загальна кількість фонем-трифонів у алфавіті теоретично дорівнює кількості базових фонем у степені три (125000 для алфавіту з 50 фонем). Практично ж можна обмежитися декількома тисячами фонем-трифонів. Відсутні фонем-трифони замінюються найближчою згідно з фонемно-трифонною ієрархією.

Оцінка параметрів моделей фонем-трифонів здійснюється за навчальною вибіркою, яка супроводжується фонемно-трифонною транскрипцією.

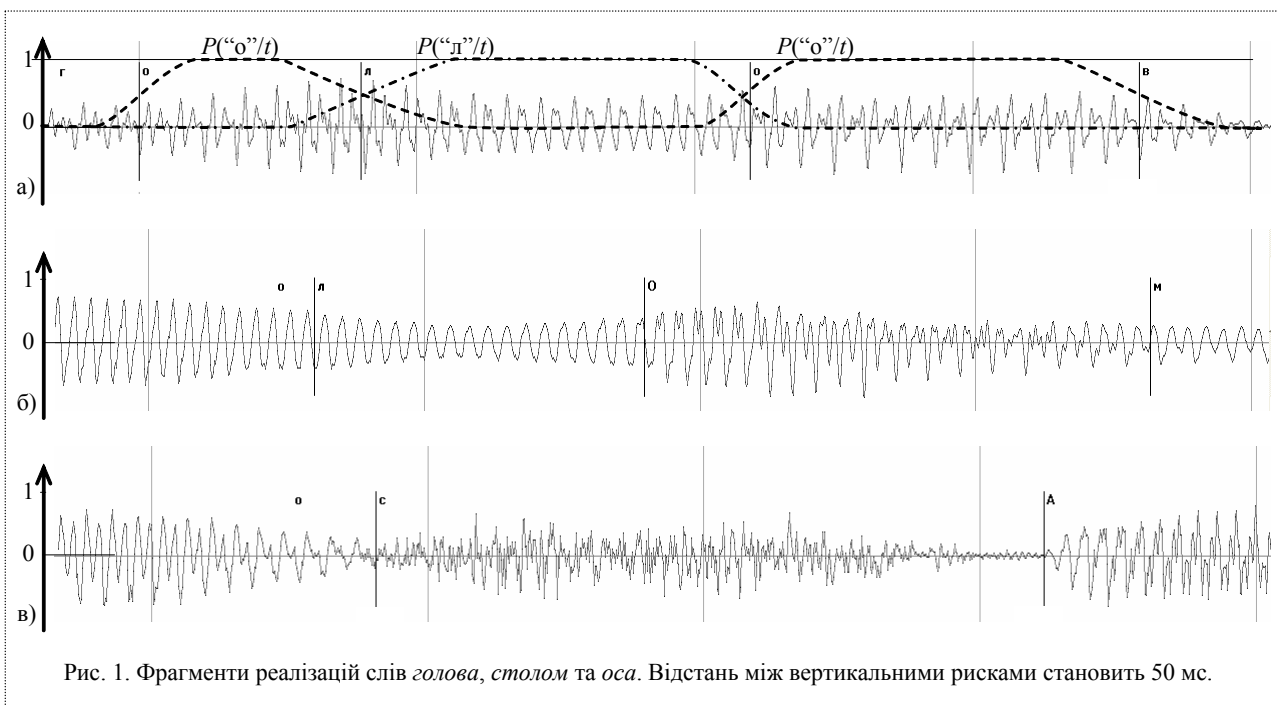
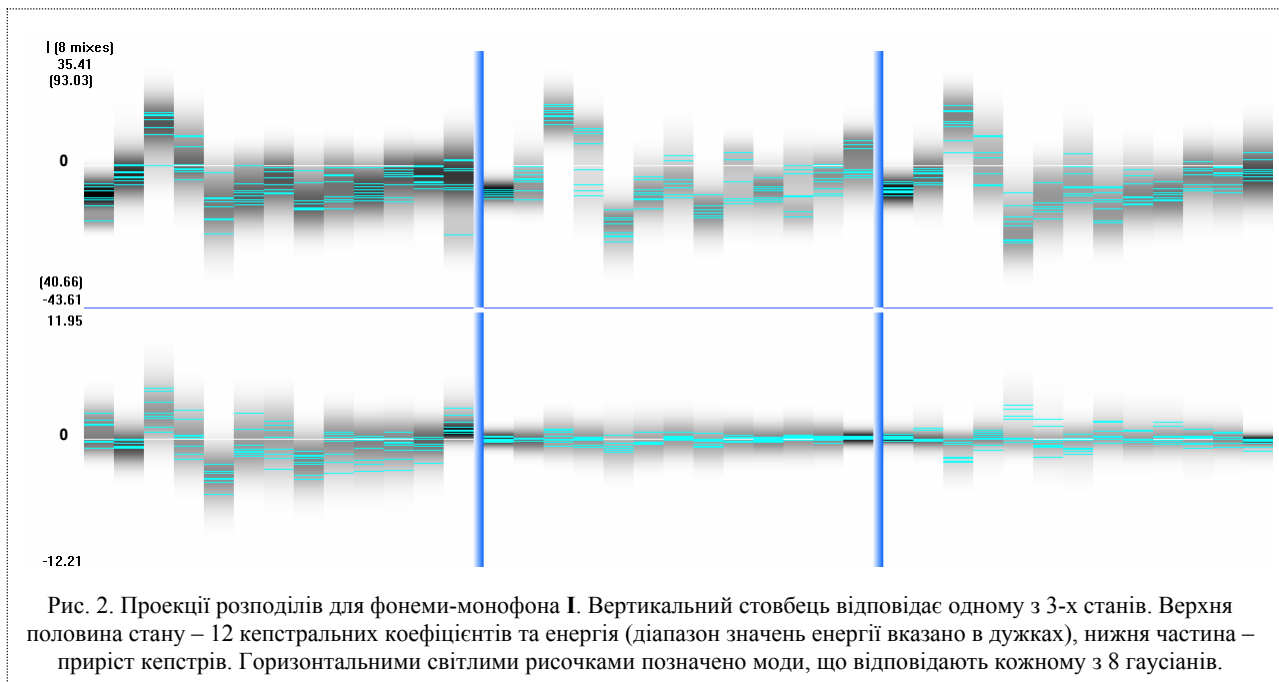


Рис. 1. Фрагменти реалізацій слів голова, столом та оса. Відстань між вертикальними рисками становить 50 мс.



Під час оцінювання параметрів для багатьох фонем-трифонів може трапитись дуже мало вимірів. Тому до нарощування гаусіанів проводимо кластеризацію фонем-трифонів за середнім і дисперсією. Кожний кластер фонем-трифонів представлено своїм центром. Матриця переходів між станами є спільною для всіх фонем-трифонів з однаковою центральною фонемою.

Використання контекстно-незалежної моделі – монофона – виправдано у випадках невеликої навчальної вибірки та в системах з обмеженими ресурсами.

4 Моделювання допустимих послідовностей фонем

У попередніх роботах при побудові граматики на першому рівні обмежень на порядок слідування фонем не накладалося. Не дивлячись на швидке виконання, робастність системи розпізнавання є далекою від бажаної, особливо для кооперативу дикторів. Тепер пропонується розглядати склади як альтернативні мовленнєві образи, які все ще слабо, залежать від словника.

Проаналізовані два шляхи вибору складів: на основі правил складоподілу та відкриті склади.

Вибір на основі правил складоподілу впливає з евристичних тверджень лінгвістичної науки щодо розміщення меж складів в залежності від сполучень фонем. Відкриті склади закінчуються голосним звуком або фонемою-паузою. Вибір складів на основі масиву даних також знаходиться у сфері інтересів і планується в подальших дослідженнях.

Словники складів були сформовані автоматично на базі частотного словника української мови. Хоча порядок слідування складів вільний, все ж на відкриті склади накладається додаткове обмеження: склади, які закінчуються фонемою паузою завжди слідують за складом, що закінчується голосним звуком.

Таблиця 1 ілюструє, що для окремо вимовлюваних слів поскладова граматики істотно покращує коректність пофонемного розпізнавання (до 1,6 разів) порівняно з розпізнаванням в умовах вільного порядку слідування фонем. Середня довжина українського слова складає 7,43 фонем, а максимальна – 20 фонем. В усіх випадках розпізнавання проведено з використанням системи Julius-Julian [6], виконується у реальному часі, який приблизно однаковий для розглянутих видів складів. Слід також зазначити, що склади, вибрані на основі правил, дають кращий результат.

Таблиця 1. Фонемна коректність при розпізнаванні різних контрольних вибірок при вільній граматиці слідування різних елементів словника.

Контрольна вибірка	Вид елемента словника	Обсяг словника	Фонемна коректність
11000 слів	монофон	55	46,0
11000 слів	склад на основі правил складоподілу	9 436	79,5
11000 слів	відкритий склад	4 966	78,3
100 речень	монофон	55	49,3
100 речень	склад на основі правил складоподілу	9 436	56,8
100 речень	відкритий склад	4 966	55,5

5 Адаптація моделей до мікропроцесорів ЦОС

В рамках програми “Образний комп’ютер” розробляється ряд портативних пристроїв широкого вжитку на основі двоядерних цифрових сигнальних процесорів *BlackFin BF-561* компанії *Analog Devices*. Реалізація алгоритмів пофонемного розпізнавання в цих пристроях є надзвичайно актуальним. Перш за все, це стосується алгоритму розпізнавання великих словників, тобто пофонемне розпізнавання окремо вимовлених слів, причому кількість слів, які система може розпізнати (словник), складає 1 000 одиниць та більше. Навчання розпізнаванню відбувається на персональному комп’ютері. Система може бути навченою як на одного диктора, так і на кооператив дикторів. В другому випадку безпосередньо на самому пристрої може відбуватись процедура адаптації (налаштування) до конкретного диктора, для того щоб покращити надійність розпізнавання.

В якості акустичних моделей фонемного рівня використовуються модифіковані моделі програмного комплексу *HTK*. Сама програма розпізнавання написана на мові програмування *C* на основі [6] для персонального комп’ютера та переписана для можливості крос-компіляції в мікропрограмний код операційного середовища *uClinux* сигнального процесора *BF-561*.

Результати розпізнавання однакових фрагментів мовлення на ПК та на портативних пристроях збігаються з точністю до 6-го знаку після коми. Уніфікація програмного коду дає змогу всі дослідження проводити на персональному комп’ютері. Надійність розпізнавання монодикторної системи зі словником в 1 000 слів з сімама коефіцієнтами та чотирма сумішами гаусіанів складає 98,5%.

Оскільки процесор *AD BlackFin BF-561* не підтримує апаратно операцій з плаваючою комою, то основною проблемою при розпізнаванні на портативних пристроях виявилась швидкість. Для розпізнавання мовленнєвого сигналу довжиною 1 секунда при словнику в 1 000 слів необхідно було більше 50 секунд програмного часу. Використання в програмі розпізнавання спеціалізованої бібліотеки *Libbfdsp* для роботи з плаваючою комою дало змогу скоротити цей час до 1.5 сек. При цьому для розпізнавання мовленнєвого сигналу довжиною 1 сек. при словнику в 10 000 слів необхідно усього лише близько 2 сек (Таблиця 2). Це пов’язано з тим, що найбільше часу іде на процедуру препроцесингу, яка не залежить від розміру словника.

Таблиця 2. Залежність тривалості обчислень від словника при розпізнаванні ізольованих слів.

Обсяг словника	Час обчислень на 1 сек. мовлення, сек			
	Препроцесинг	1-й етап	2-й етап	Загалом
100 слів	1,05	0,25	0,02	1,42
1000 слів	1,05	0,37	0,09	1,51
10000 слів	1,05	0,58	0,35	1,98

Наступним кроком є адаптація основних функцій препроцесингу та розпізнавання шляхом переведення їх на фіксовану кому, що за прогнозами дозволить добитись розпізнавання мовленнєвих сигналів у реальному часі для словників розміром 10 000 слів та більше.

6 Висновки

Модель пофонемного розпізнавання дає змогу не лише включати до словника нові слова без навчання на них, а й є актуальною для мов з великою кількістю словоформ та відносно вільним порядком слідування слів, до яких відносяться і слов’янські мови.

Реалізована фонемно-трифонна модель враховує явище коартикуляції, дає змогу зменшити кількість варіантів транскрипцій на одне слово та покращує перспективи використання словників з дрібніших ніж слово елементів, таких як склади, морфеми та фонем, що дає змогу перейти до побудови моделей словотвору.

Планується розширити дослідження з розпізнавання злитого мовлення та настроювання на голос диктора з подальшою адаптацією до мікропроцесорів ЦОС.

Література

1. Т.К. Винцюк. Анализ, распознавание и смысловая интерпретация речевых сигналов. — Киев: Наукова думка, 1987.
2. Taras K. Vintsiuk, Mykola M. Sazhok: Speaker Voice Passport for a Spoken Dialogue System. — Proceedings of the 3rd International Workshop “Speech and Computer” – SPECOM’98, St.-Petersburg, 1998.
3. Taras K. Vintsiuk, Tetiana V. Liudovyk, Mykola M. Sazhok: Phonetic Knowledge Base for Ukrainian.— Proceedings of the 3rd International Workshop “Speech and Computer” - SPECOM’98, St.-Petersburg, 1998.
4. Taras Vintsiuk, Mykola Sazhok, Taras Vintsiuk, Gerard Chollet: Acoustic-Phonetic Model Application for Syllable Speech Recognition Output Post-Processing. — Proc. of the 11th International Conference “Speech and Computer” – SPECOM’2007, Moscow, 2007, pp. 182–187.
5. Taras K. Vintsiuk, Mykola M. Sazhok: Multi-Level Multi-Decision Models in ASR”, Proc. of the 10th International Workshop “Speech and Computer”, SPECOM’2005, Patras, 2005, pp. 69–76.
6. Lee, T. Kawahara and K. Shikano: Julius – an open source real-time large vocabulary recognition engine. In Proc. European Conference on Speech Communication and Technology (EUROSPEECH), 2001, pp. 1691–1694.