Пофонемне розпізнавання мовленнєвих сигналів на мікропроцесорах ЦОС сімейства *BlackFin*

Дмитро Федорин

Міжнародний науково-навчальний центр інформаційних технологій та систем
40 просп. Академіка Глушкова, Київ 03680
fedoryn@uasoiro.org.ua

Dmytro Fedoryn. **Phoneme-based speech recognition for BlackFin DSP.** The aspects of phoneme-based recognition for DSP are described. The description of knowledge and data base speech for speech recognition system evaluation is given. Experimental results are discussed.

1 Вступ

Реалізація алгоритмів пофонемного розпізнавання в портативних пристроях є надзвичайно актуальною проблемою. Перш за все, це стосується алгоритму розпізнавання великих словників, тобто пофонемного розпізнавання окремо вимовлених слів, причому кількість слів, які система може розпізнати (словник), складає 1 000 одиниць та більше. Ще більш важливою є проблема розпізнавання злитого мовлення – тобто розпізнавання довільної послідовності вимовлених слів. Ці алгоритми можуть використовуватись для керування портативними пристроями, перекладу сказанного іншими мовами, голосового пошуку та при вирішенні величезної кількості інших задач.

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]. Це особливо важливо при роботі з портативними пристроями, процедури навчання та донавчання на яких можуть бути пов'язані з певними ускладненнями для користувача.

Цю статтю присвячено опису алгоритму розпізнавання окремовимовлюваних слів та злитого мовлення та його адаптації для використання на портативних пристроях з процесорами сімейства *BlackFin* в операційному середовищі *uCLinux*.

У розділі 2 приводиться загальний опис запропонованого алгоритму розпізнавання мовленнєвих сигналів. У розділі 3 розглядаються акустичні моделі фонем. Розділ 4 присвячено принципам наповнення словника. Лінгвістичні моделі та граматики описані в розділі 5. У розділі 6 представлено можливості операційного середовища *uCLinux* для мікропроцесорів *BlackFin*. У розділі 7 акцентується увага на особливостях реалізації пофонемного розпізнавання на мікропроцесорах ЦОС.

2 Опис алгоритму розпізнавання

В якості алгоритму розпізнавання використовується двопрохідний алгоритм «прямого-зворотнього» ходу (*forward-backward*).

На першому етапі на основі лінгвістичного словника будується дерево, внутрішнім вузлам якого присвоюються заздалегідь обчисленні значення з акустичної моделі, а до вузлів, що є кінцями слів застосовується біграмні ймовірності з лінгвістичної моделі. На цьому дереві обчислюється найкращий шлях, вузли кінців слів разом зі своїми значеннями та передбачені слова-кандидати і передаються на вхід другого етапу.

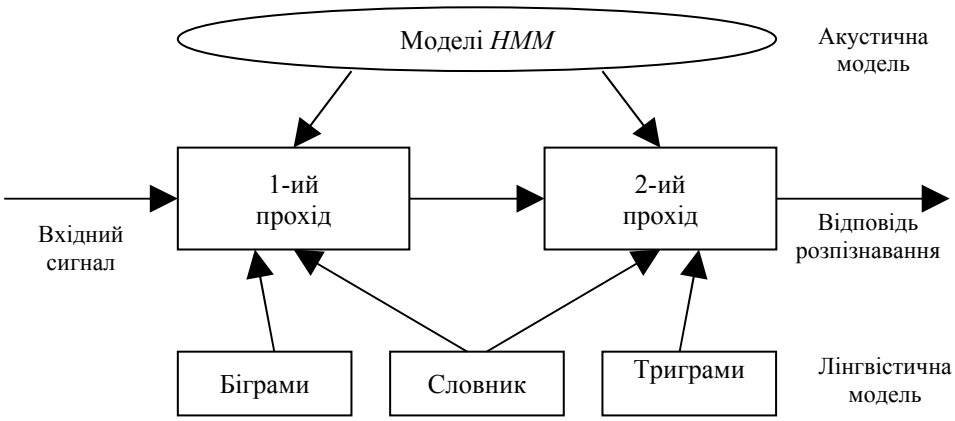

Рис.1 Загальна схема алгоритму

На другому етапі пошук відбувається в зворотньому напрямку і застосовується лінгвістичні моделі N-грам слів.

3 Акустична модель

В системі використовується опис мовленнєвих сигналів за допомогою мел-частотних кепстральних коефіцієнтів (*MFCC*). Вони обчислюються з логарифму частотних амплітуд $\{m_j\}$, із застосуванням дискретного косинусного перетворення

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right),$$

де N – кількість частотних каналів.

Для покращення якості розпізнавання крім коефіцієнтів *MFCC* також можна враховувати енергію (*_E*), коефіцієнти регресії першого порядку (дельта коефіцієнти) (*_D*) та коефіцієнти регресії другого порядку (коефіцієнти приросту) (*_A*). Енергія обчислювалася як логарифм сигналу енергії:

$$\mathrm{E} = \log \sum_{n=1}^{N} s_n^{\,2} .$$

Дельта коефіцієнти обчислюються за такою формулою:

$$d_t = \frac{\sum_{\Theta=1}^{\Theta} \Theta(c_{t+\Theta} - c_{t-\Theta})}{2\sum_{\Theta=1}^{\Theta} \Theta^2},$$

де Θ береться рівним 2. Ця ж сама формула застосовується до дельта коефіцієнтів, щоб отримати коефіцієнти приросту.

Алфавіт фонем задано у вигляді скінченної множини K, куди входять фонеми $k \in K$, 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: $k \in$ {а, о, у, е, и, і, А, О, У, Е, И, І, б, б', в, в', г, г', ґ, ґ', 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.

Акустичні моделі фонем являють собою генеративну модель, де кожна форма-елемент (стан) описується сумішшю нормальних законів. Всі фонеми мають три стани без пропусків і повернень за виключенням фонеми-паузи. Параметри нормальних законів беруться з усномовного паспорта диктора або кооперативу дикторів.

4 Словник

Словник являє собою набір транскрипцій слів, що розпізнаються, представлених у вигляді послідовностей фонем, що входять до акустичної моделі. Допускається мультиваріантність транскрипцій. Для цього різні варианти транскрипції одного і того ж самого слова просто записуються в одному рядку розділені спеціальним сепаратором.

5 Лінгвістична модель

Представлений алгоритм може бути використаний як для розпізнавання окремих слів, так і для розпізнавання злитого мовлення.

При розпізнавання окремих слів на вхід подається лише словник та акустичні моделі. Відповідь розпізнавання видається алгоритмом одразу після першого етапу. Другий етап несуттєвий.

Розпізнавання злитого мовлення може відбуватись на основі *N*-грамної моделі чи БНФ-граматик. В випадку *N*-грамної моделі для пошуку відповіді розпізнавання використовуються біграми на першому етапі та *N*-грами на другому. В випадку БНФ-граматик на вхід подається додатково спеціальний файл, в якому лінгвістична модель представлена за допомогою спеціальних форм Бекуса-Наура.

6 Особливості операційного середовища **uCLinux** для ЦОС **BlackFin**

Для сигнальних процесорів *Analog Devices* сімейства *BlackFin* існує можливість запуску на цих процесорах операційного середовища *uCLinux*, що належить до сімейства *UNIX*-подібних операційних систем, та має за основу вихідні коди ядра ОС *Linux*. Як і інші *Linux*-дистрибутиви, *uCLinux* розповсюджується згідно ліцензії *GNU GPL*. Згідно з основним принципом цієї ліцензії – «свобода запускати, копіювати, розповсюджувати, вивчати, змінювати і вдосконалювати програмне забезпечення» усі вихідні коди необхідні для компіляції операційного середовища *uCLinux* відкриті і доступні для скачування за адресою - http://blackfin.uclinux.org/gf/. Розробка ведеться групою ентузіастів до якої може долучитись кожен, при безпосередній підтримці спеціалістів *Analog Devices*. На сайті доступні три основних модуля - *GNU Toolchain* (крос-компілятор), *Das U-boot* (вихідні файли загружчика), *Linux Kernel* (вихідні файли ядра ОС *uCLinux*).

GNU Toolchain (крос-компілятор) – спеціальний компілятор що запускається в операційному середовищі *Linux* на персональному комп'ютері та продукує виконуваний код для операційного середовища *uCLinux* платформи на основі сигнального процесора *AD BlackFin*. Використовується як для крос-компіляції вихідних кодів ядра ОС *uCLinux* та загружчика *Das U-boot*, так і для крос-компіляції прикладних програм написаних на мові програмування *C\C++* для можливості їх виконання в середовищі uCLinux. До складу компілятора входять такі основні модулі: компілятори *gcc* та *gcc-elf* (версій 3.4 та 4.1, що дає широкі можливості сумісності програм) та

спеціалізовану бібліотеку для вбудованих систем *uclibc*. Для зручності крос-компілятор *GNU Toolchain* надається у двох видах – у вигляді пакету rpm та у вигляді архівів *tar.gz*. Якщо ваша система допускає роботу з пакетами rpm то зручніше використовувати їх. Також крос-компілятор існує у 2х версіях – для 32-бітних та для 64-бітних систем відповідно. Після встановлення пакетів достатньо виконати команду export PATH=$PATH :/opt/uClinux/bfin-uclinux/bin:/opt/uClinux/bfin-linux-uclibc/bin після чого увесь функціонал буде доступний для використання стандартних методів процедури make.

Das U-boot (загружчик) – комп'ютерний загружчик операційних систем орієнтований на вбудовані пристрої архітектур *MIPS*, *ARM* та інших. Після крос-компіляції може бути записаний в *Flash-ROM* платформи. Після чого загружчик виконуватиметься при старті системи, що дозволяє завантажити в пам'ять та запустити ядро ОС *uCLinux*.

Linux Kernel (ядро ОС *uCLinux*) – центральна частина операційного середовища uCLinux, що забезпечує прикладним програмам координований доступ до ресурсів комп'ютера, таким як процесорний час, пам'ять та зовнішнє апаратне забезпечення і надає сервіси файлової системи.

Для того щоб запустити на платформі заснованій на процесорах *BlackFin* операційне середовище *uCLinux* спочатку необхідно скомпілювати таку конфігурацію загружчика *U-boot* та ядра *Linux* Kernel що відповідає функціональним та технічним характеристикам саме цієї платформи. Для цього необхідно вносити правки в конфігураційні файли та (якщо це необхідно) в файли вихідного коду загружчика та ядра.

7 Адаптація моделей до мікропроцесорів ЦОС

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.

Сама програма розпізнавання написана на мові програмування *C* на основі [3,,5] 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 сімома 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 трохи більше 2,5 сек (Таблиця 1). Найбільше часу іде на процедуру препроцесингу, яка не залежить від розміру словника.

Таблиця 1. Залежність тривалості обчислень від словника при розпізнаванні ізольованих слів.

Обсяг словника	Час обчислень на 1 сек. мовлення, сек (частота дискретизації 16KHz; акустична модель - 7 коефіцієнтів, 4 мікси)			
	Препро-цесинг	1-й етап	2-й етап	Загалом
100 слів	1,00	0,32	0,03	1,35
1000 слів	1,00	0,53	0,09	1,62
10000 слів	1,00	0,90	0,70	2,60

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.

8 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.

Планується розширити дослідження з розпізнавання злитого мовлення та настроювання на голос диктора з подальшою адаптацію 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: Multi-Level Multi-Decision Models in ASR", Proc. of the 10th International Workshop "Speech and Computer", SPECOM'2005, Patras, 2005, pp. 69–76.
3. A. Lee and T. Kawahara: Recent Development of Open-Source Speech Recognition Engine Julius. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2009.

4. A. 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.
5. T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro and K. Shikano: Free software toolkit for Japanese large vocabulary continuous speech recognition. In Proc. Int'l Conf. on Spoken Language Processing (ICSLP) , Vol. 4, pp. 476–479, 2000.