Ідентифікація мови диктора з використанням акустичної та фонетичної інформації

Дмитро Федорин

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

Dmytro Fedoryn. **Spoken Language Identification Utilizing Acoustic and Phonetic Speech Information.** The aspects of Spoken Language Identification (*LID*) are described. The benefits of using statistical models for *LID* systems are spotted. The description of knowledge and speech data base for Spoken Language Identification system evaluation is given. Experimental results are discussed.

1 Вступ

Люди здавна звикли використовувати мову як основний засіб комунікації з собі подібними. І перші ж контакти з особами що належали до іншого племені чи народу окреслили основну проблему для міжплемінного чи міжнародного спілкування – різноманітність людських мов, яка унеможливлювала спілкування між представниками різних народів без знання хоча б одним з них мови співбесідника.

Бурхливий розвиток комп'ютерних технологій в 21 сторіччі ознаменувався підвищеним інтересом до мовленнєвих технологій. І це не дивно - світова глобалізація вимагає від людства можливості швидко і доступно порозумітись незалежно від мови, якою володіють співрозмовники. При цьому в наш час співрозмовником не обов'язково повинна виступати людина. Такі мобільні інтелектуальні системи як *Siri* (*iOS*) чи *Google Voice Search* (*Android*) цілком здатні вести з людиною зв'язний діалог, та виконувати певні команди. Все це ставить перед розробниками багато нових та цікавих наукових задач. Однією з таких задач є задача автоматичного розпізнавання (або ідентифікації) мови диктора в потоці мовлення.

Автоматичне розпізнавання мови являє собою процес визначення мови, якою говорить диктор, в межах певного висловлювання (фрази, речення або певної кількості речень). Основна відмінність від більшості інших задач розпізнавання мовлення полягає в тому, що ніякої безпосередньої інформації із зазначенням змісту висловлювання чи особистості диктора не надається. Отже система розпізнавання мови повинна використовувати різні аспекти мовленнєвої інформації, що характеризують відмінності між різними мовами, а також бути достатньо гнучкою щоб нівелювати можливий вплив відмінностей мовленнєвих стилів різних дикторів.

У розділі 2 наводиться загальний опис алгоритму ідентифікації мови. У розділі 3 розглядаються акустичні характеристики сигналів, що використовуються при розпізнаванні. Розділ 4 присвячено фонетичним характеристикам. Просодичні, морфологічні та синтаксичні характеристики представлено в Розділі 5. У розділі 6 описано наявні бази даних та корпуси для експериментів з ідентифікації мови. У розділі 7 мова йде про інструментарій розробника систем ідентифікації мови. У 8-му розділі підбиваються підсумки статті.

2 Опис алгоритму розпізнавання

Загальна схема алгоритму представлена на Рисунку 1. Етап препроцессингу, на якому відбувається обчислення вектору ознак акустичної інформації для вхідного сигналу, спільний для процедур навчання та розпізнавання.

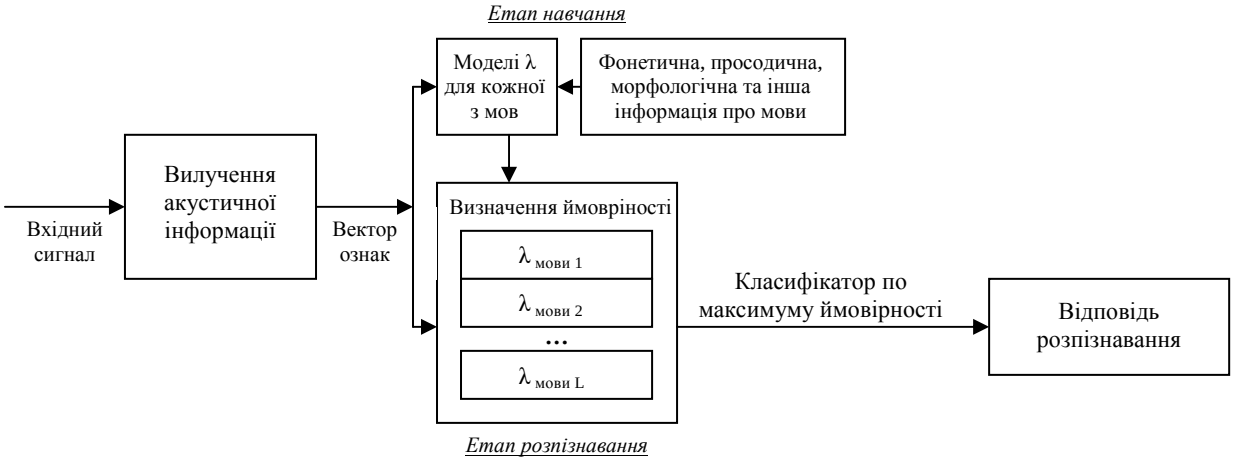

Рис.1 Загальна схема алгоритму

Процедура навчання полягає в створенні моделей для кожної з можливих мов з використанням акустичної, фонетичної, просодичної, морфологічної та іншої наявної інформації.

Процедура розпізнавання полягає в визначенні найбільш ймовірної, з числа *L* можливих, мови, яка відноситься до вхідного мовленнєвого сигналу.

Як і для більшості задач розпізнавання, для задачі ідентифікації мови може бути використаний класифікатор по максимуму ймовірності. Для того щоб класифікатор був функціональним необхідно провести процедуру навчання. Ціллю цієї процедури є обчислення мовленнєвих характеристик *X*, які представляють собою особливий тип мовної інформації. Моделі λ створюються за допомогою методів моделювання (як правило статистичних) щоб передати характеристики кожної з мов що входять до навчальної вибірки. На етапі розпізнавання, ті ж самі характеристики *X* виділяються для висловлювання, мову якого ми визначаємо. Потім набір обчислених характеристик порівнюється з набором моделей λl, l=1…*L*, де *L* це кількість можливих мов в системі. На останньому кроці необхідно визначити найбільш ймовірного претендента згідно формули:

$$\hat{L} = \arg\max_{1 \le l \le L} P(\lambda_l \mid X) \qquad (1)$$

Мова $\hat{L}$, яка представляє визначену за формулою (1) модель, і буде ідентифікована системою як мова, що використовується у висловлюванні, що розпізнається.

Акустична, фонетична та просодична інформація широко використовується в системах ідентифікації мови, оскільки вона є невід'ємним аспектом мовлення. Для моделювання цих характеристик можуть бути використані такі статистичні методи, як модель гаусових сумішей (*GMM*) [1] та приховані моделі Маркова (*HMM*) [2]. Найбільш поширеним методом для визначення фонетичних особливостей є метод пофонемного розпізнавання. Просодичні особливості, такі як період основного тону, тривалість та інтенсивність, також можуть бути використані при розробці систем ідентифікації мов. Для використання даних про синтактичні правила (наприклад такі як граматика) та різноманітність словоформ необхідні знання щодо письмового аспекту мови.

3 Акустичні характеристики

Акустичні дані надають певне уявлення щодо мовленнєвого сигналу в певний момент часу, та являють собою набір параметрів що компактно передають найважливішу акустичну інформацію вилучену за допомогою процесу параметризації. Ця інформація є найбільш примітивною з усієї мовленнєвої інформації та може бути отримана безпосередньо з запису висловлювання. В той же час вона є будівельним матеріалом для визначення фонетичних та інших особливостей більш високого рівня. Найбільш ефективними методами акустичної параметризації є методи мел-частотних кепстральних коефіцієнтів (*MFCC*) [3] та метод кепстральних коефіцієнтів лінійного передбачення (*LPCC*) [4].

Мел-частотні кепстральні коефіцієнти (*MFCC*) обчислюються з логарифму частотних амплітуд $\{m_j\}$, із застосуванням дискретного косинусного перетворення

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

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:

$$\mathrm{E} = \log \sum_{n=1}^{N} s_n^2 . \qquad (3)$$

Дельта коефіцієнти обчислюються за такою формулою:

$$d_t = \frac{\sum_{\Theta=1}^{\Theta} \Theta \left(c_{t+\Theta} - c_{t-\Theta} \right)}{2 \sum_{\Theta=1}^{\Theta} \Theta^2}, \qquad (4)$$

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

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 визначаються на етапі навчання.

4 Фонетичні характеристики

Існує скінченна кількість значимих звуків, які з'являються в людських мовах та можуть бути фізично вимовлені людиною. Не всі ці звуки з'являються в тій чи іншій мові, отже, кожна мова має свою власну скінчену підмножину значущих звуків. Фонема — найменша (неподільна) структурно-семантична звукова одиниця, що репрезентує певний звук або групу звуків. Фонема як певний знак, модель матеріалізується в мовленні у вигляді звуків, серед яких вирізняють головний вияв фонеми (інваріант) та її варіанти (алофони). Фонетичні характеристики являють собою дані щодо послідовності звукових одиниць, які можна отримати з мовленнєвого сигналу. Різні мови можуть мати різні набори звуків та фонем що їх представляють. Також фонема/звук може бути спільною для двох мов, але частота використання їх в цих мовах може різнитись. Правила, що регулюють допустимі послідовності з фонем/звуків також можуть бути різними. Таким чином, фонетична інформація є вкрай важливою інформацією для використання в якості даних при ідентифікації мови. Для задач

ідентифікації мови зручно використовувати єдиний набір фонетичних символів з усіх можливих мов. Міжнародний фонетичний алфавіт (МФА) [5] і система *Worldbet* [6] (*ASCII* кодування для МФА) призначені саме для цієї мети. Станом на 2008 рік у МФА визначено 107 окремих символів фонем, 52 діакритичних знаки і 4 знаки просодії.

В системі ідентифікації мови алфавіт фонем задається у вигляді скінченної множини K, куди входять фонеми $k \in K$, які спостерігаються в природній мові [7]. До алфавіту включено також фонему-паузу #.

5 Інші характеристики

До просодичних характеристик відноситься інформація щодо тривалості звуків, інтонації (періоду основного тону) та моделей наголосу. Звуки, що позначаються однаковими символами для багатьох мов, часто мають різні просодичні характеристики в залежності від фонетичної системи конкретної мови.

Морфологія та синтаксис також можуть бути корисними в процесі ідентифікації мови [8]. Кожна мова має свій словник та свій спосіб словотворення. Та навіть коли одне й те саме слово зустрічається в двох різних мовах, його контекст, тобто слово в оточенні слів що передують йому чи слідують за ним, буде відмінним.

Описаний вище набір мовленнєвих характеристик є достатнім для побудови системи ідентифікації мови за допомогою статистичних моделей та показує перспективні результати, що підтверджується експериментальними дослідженнями [1, 8].

6 Бази даних та корпуси

Хоча деякі з відомих мовленнєвих корпусів містять кілька мов (наприклад *SpeechDat*), проте вони з певних причин можуть не найкращим чином підходити для експериментів з ідентифікації мови. Першим же спеціалізованим корпусом для ідентифікації мови став Oregon Graduate Institute *Telephone Speech Corpus* (*OGI-TS*C) [9] представлений в 1992 році. Він містив 10 мов: англійська, фарсі, французька, німецька, японська, корейська, китайська, іспанська, тамільська та в'єтнамська, кожна з яких була представлена записами телефонних розмов 90 носіїв мови. Загальна довжина записів варіювалась від 3,5 годин для англійської мови до 1,5 годин для в'єтнамської.

Наступним корпусом був *Oregon Graduate Institute* 22 *Language Telephone Speech Corpus* (*OHSU*) [10] представлений в 1995 році. Він містить 22 мови: східно-арабська, кантонський діалект китайської, чеська, фарсі, французька, німецька, хінді, угорська, японська, корейська, малайська, мандаринський діалект китайської, італійська, польська, португальська, російська, іспанська, шведська, суахілі, тамільська, в'єтнамська та англійська. Корпус містить як висловлювання з фіксованого словника (такі наприклад як числівники, дні тижня) так і записи спонтанного злитого мовлення.

База *CALLFRIEND* представлена Консорціумом лінгвістичних даних (*LDC*) включає 12 мов, 3 з яких представлені двома різними діалектами.

База 2011 *NIST Language Recognition Evaluation* (*LRE*-2011) [11] містить 24 мови та діалекти: іракська арабська, ліванська арабська, сучасна стандартна арабська (*MSA*), арабська-магріб, індійська англійська, американська англійська, російська, фарсі, словацька, хінді, іспанська, лаоська, тамільська, бенгальська, мандаринський діалект китайської, тайська, чеська, пенджабі, турецька, дарі пушту, українська, польська та урду.

7 Інструментарій

Задача ідентифікації мови має багато спільного з задачами автоматичного розпізнавання мовленнєвих сигналів та задачею верифікації диктора. Тому для розв'язання цих задач використовується схожий інструментарій. Це можуть бути як відомі open-source пакети *HTK*, *MASV*, *Becars* [12] так і інші програми.

Пакет *HTK* може використовуватись для вилучення акустичних характеристик та побудови моделей фонем.

Пакети *MASV* та *Becars* можуть використовуватись для визначення найбільш статистично вірогідних претендентів в задачі ідентифікації, використовуючи в якості вхідних параметрів параметри моделей фонем з *HTK*.

8 Висновки

Описано алгоритм ідентифікації мови диктора з використанням акустичної та фонетичної інформації. Наведено мовленнєві характеристики, що використовуються при роботі алгоритму. Розглянуто основні мовленнєві бази даних що використовуються для створення діючих моделей ідентифікації мови та необхідний для цього інструментарій.

.

Література

1. *J. McLaughlin, D. A. Reynolds, and T. Gleason, "A Study of Computation Speed-Ups of the GMM-UBM Speaker Recognition System" / EuroSpeech, vol. 3, 1999, pp. 1215–1218.*
2. *Steve Young et al, "The HTK Book for Hidden Markov Model Toolkit (HTK) Version 3.4" / http://htk.eng.cam.ac.uk, 2006.*
3. *S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences" / IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-28, no. 4, 1980, pp. 357–366.*
4. *Дж. Д. Маркел, А. Х. Грэй, "Линейное предсказание речи" / Москва: Связь, 1980.*

5. *IPA, "Handbook of the International Phonetic Association : a guide to the use of the International Phonetic Alphabet" / Cambridge University Press, 1999.*
6. *J. Hieronymous, "ASCII Phonetic Symbols for the World's Languages: Worldbet" / Journal of the International Phonetic Association, 1993.*
7. *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.*
8. *Kim-Yung Eddie Wong, "Automatic Spoken Language Identification Utilizing Acoustic and Phonetic Speech Information" / Ph.d. thesis, Queensland University of Technology, 2004.*
9. *Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI Multi-Language Telephone Speech Corpus" / in International Conference on Spoken Language Processing, vol. 2, pp. 895–898, 1992*
10. *T. Lander, R. Cole, B. Oshika, and M. Noel, "The OGI 22 Language Telephone Speech Corpus" / in The European Conference on Speech Communication and Technology, 1995*
11. *NIST, "Spoken Natural Language Processing Group." / http://www.nist.gov/speech/, 2004-2012*
12. *N. Dehak and G. Chollet, "Support vector gmms for speaker verification" / In the proceedings of the IEEE Workshop on Speaker and Language Recognition (Odyssey), June 2006*