1980.
25. *Шпак З. Я. , Рашкевич Ю. М.* Регулирование темпа подачи речевой информации//Распознавание и синтез звуковых сигналов. К: ИК АН УССР. 1987. С. 70-77.
26. *ElJaroudi . A.* Discrete all pole modeling//IEEE Trans. Acoust. Speech Signal Process. 1991. Vol. 39. - № 2. P. 411-423.
27. *Гнатив Я. Н. , Рашкевич Ю. М.* Гомоморфная модель регулирования темпа речи//Контрольно-измерительная техника. 1984. № 26. С. 84-87.
28. *Seneff S.* System to Independently Modify Excitation and/or Spectrum of Speech Waveform Without Explicit Pitch Extraction// IEEE Trans. on Acoust. Speech and Signal Proc. 1982. № 4. P. 566-578.
29. *Marques J. S. and Almeida L. B.* Frequency-varying sinusoidal modeling of speech// IEEE Trans. Acoust. Speech Signal Process. 1989. Vol. 37. № 5. P. 763-765.
30. *Lienard J. S. and d'Alessandro C.* Wavelets and granular analysis of speech. In:"Wavelets, Time-Frequency Methods and Phase Space, ed. by Combes J. M. Berlin,Springer. 1989. P. 744-754.
31. *Alessandro C.* Time-frequency speech transformation based on an elementary waveform representation// Speech Communication. 1990. № 9. P. 419-431.

УДК 621.391.19

ПРО ПІДХОДИ ДО ОРГАНІЗАЦІЇ СИСТЕМ РОЗПІЗНАВАННЯ МОВИ

© Роман Попович
НУ "Львівська політехніка", м.Львів, вул.С.Бандери, 12

Розглянуто підходи до організації сучасних систем розпізнавання суцільної мови з великим словником. Оцінено кількість трифонів для словника найуживаніших слів української мови та запропоновано питання для фонетичних вирішуючих дерев.

Approaches to organization of current large vocabulary continuous speech recognition systems have been considered. Number of triphones for Ukrainian mostly used words dictionary has been estimated and phonetic decision trees questions have been proposed.

Сучасні системи для розпізнавання суцільної мови з великим словником грунтуються на принципах статистичного розпізнавання образів [1].

На першому етапі мовний зразок перетворюється акустичним процесором на послідовність акустичних векторів $Y = y_1, y_2, ..., y_T$. Кожен вектор є стислим поданням короткочасного мовного спектра на інтервалі, як правило, близько 25 мс зі зсувом ін-

тервалів на 10 мс. Типова фраза з десяти слів по 6-7 звуків у кожному може мати тривалість близько 3 с і зображатися послідовністю з $T = 300$ акустичних векторів.

У загальному, фраза складається з послідовності слів $W = w_1, w_2, ..., w_n$. Робота системи розпізнавання полягає у визначенні найімовірнішої послідовності слів $\hat{W}$, за акустичним сигналом Y. Для цього використовується правило Байєса [1]:

$$\hat{W} = \arg\max_W P(W/Y) = \arg\max_W \frac{P(W)P(Y/W)}{P(Y)}.$$

Ця рівність показує, що для визначення найправдоподібнішої послідовності слів W повинна бути знайдена послідовність, що робить максимальним добуток $P(W)$ та $P(Y/W)$.Оскільки знаменник $P(Y)$ не залежить від W, то його при розпізнаванні ігнорують.

Перший із співмножників являє собою апріорну ймовірність спостереження W незалежно від спостереження мовного сигналу. Ця ймовірність визначається моделлю мови.

Другий співмножник являє собою ймовірність спостереження послідовності векторів Y, якщо задана послідовність слів W. Ця ймовірність визначається акустичною моделлю.

В акустичній моделі послідовності слів діляться на базові звуки - фонеми. Кожна індивідуальна фонема подається прихованою моделлю за Марковим (англійська назва - *hidden Markov model* (*HMM*)). *HMM* - модель фонеми, як правило, має три породжуючі стани та вхідний і вихідний стан. Вхідний і вихідний стани дають змогу моделям фонем об'єднуватися, щоб утворювати слова, та об'єднувати слова, щоб утворювати речення (послідовності слів).

Вважається, що кількість фонем в українській мові дорівнює 38 [2]. Здавалось би, потрібно вивчити лише 38 *HMM*-моделей. На практиці, проте, контекстні ефекти спричиняють значні зміни у способі утворення звуків (так зване явище коартикуляції). Тому, щоб досягти доброго фонетичного розрізнення, треба навчати різні *HMM* для різних контекстів.

Найзагальнішим є підхід з використанням трифонів, коли кожна фонема має окрему *HMM*-модель для кожної індивідуальної пари сусідів ліворуч та праворуч [1]. В останніх публікаціях згадується і про використання квінфонів [3, 4].

Наприклад, нехай $x - y + z$ - це фонема y, що стоїть після x і перед z. Тоді фраза "*Цей комп'ютер*" подається послідовністю фонем $\cup$ *ц е й к о м п й у т е р* $\cup$, де $\cup$ позначає паузу. Якщо використовуються *HMM*-моделі трифонів, то фраза буде моделюватися так:

И-ц+е ц-е+й е-й+к й-к+о к-о+м о-м+п м-п+й п-й+у й-у+т у-т+е т-е+р е-р+И

Розглянуті так звані трифони між словами забезпечують найкращу точність моделювання, проте роблять складним декодування. Простіші системи розпізнаваня отримуємо, використовуючи тільки трифони всередині слів.

Маючи 38 фонем, отримуємо $38^3 = 54872$ можливі трифони, проте не всі використовуються через обмеження української мови.

Загальна кількість трифонів, необхідних для практичного вживання, залежить від вибраної множини фонем, словника та граматичних обмежень. Автором розроблена програма для оцінки потрібної кількості трифонів. На першому етапі програма виконує автоматичне транскрибування, тобто перетворює орфографічний запис слів у їх фонетичну вимову(транскрипцію), а на другому - підраховує кількість трифонів. Так, близько 4000 трифонів всередині слів потрібно, коли маємо словник із 1008 найвживаніших слів української мови, взятих із частотного словника.

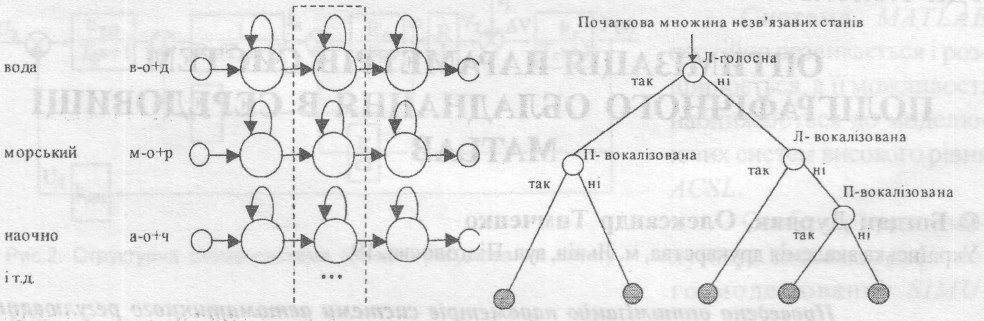
Використання лінійної комбінації багатовимірних розподілів Гаусса дає змогу промоделювати розподіл акустичного виходу для кожного породжуючого стану дуже точно [1]. Проте, використовуючи трифони, отримуємо систему з надто великою кількістю параметрів, які треба оцінити (здійснити навчання). Приблизно 10 компонент лінійної комбінації дають добрі показники для системи розпізнавання. Припускається, що всі коваріаційні матриці розподілів Гаусса є діагональними, а довжина акустичного вектора дорівнює 39 (енергія фрагмента сигналу + 12 значень кепстру + їх "дельти" та "дельти дельт"). Тоді на один стан треба 790 параметрів. Отже, 4000 трифонів з трьома породжуючими станами вимагають близько 9,5 мільйона параметрів.

Ця проблема надто великої кількості параметрів та надто малого обсягу навчальних даних є ключовою для розроблення систем статистичного розпізнавання мови. Для її вирішення використовується зв'язування станів [1,3,4]. Ідея полягає в тому, щоб зв'язати стани, які акустично не відрізняються. Це дає змогу всі дані, які відповідають кожному індивідуальному станові, об'єднати і за допомогою цього дати більш робастні оцінки параметрів зв'язаного стану. Після зв'язування ряд станів використовують один і той самий розподіл.

Вибрати, які стани зв'язувати, можна за допомогою фонетичних вирішуючих дерев. Це передбачає побудову бінарного дерева для кожного стану кожної фонеми. У кожному вузлі такого дерева ставиться запитання, на яке треба відповісти "так" або "ні".

Для трифонів питання належать до фонетичного оточення (контексту) безпосередньо ліворуч і праворуч. Одне дерево будується для кожного стану кожної фонеми, щоб розбити на підмножини всі відповідні стани всіх відповідних трифонів.

Основні питання, зв'язані з вузлами дерева, які пропонується використовувати для розпізнавання української мови, наведені нижче (П позначає правий контекст, а Л - лівий контекст):



П - пауза, Л - пауза; П - голосна, Л - голосна; П - наголошена, Л - наголошена; П -

вокалізована, Л - вокалізована; П - носова, Л - носова; П - невокалізована, Л - невокалізована; П - щілинна, Л - щілинна; П - дзвінка вибухова, Л - дзвінка вибухова.

Крім того, можна використовувати запитання, які належать до конкретних наборів контекстів. Запитання в кожному вузлі вибирають так, щоб максимізувати правдоподібність навчальних даних, які даються відповідними зв'язаними станами.

Рисунок ілюструє зв'язування центральних станів усіх трифонів фонеми [*о*] з використанням фонетичного вирішуючого дерева. Малі кола у верхній частині рисунку позначають вхідний та вихідний стани трифонів, а великі кола - породжуючі стани трифонів.

У нижній частині рисунка зображено фонетичне вирішуюче дерево. Зафарбовані кола позначають кінцеві вузли дерева. Усі стани в тому самому кінцевому вузлі дерева зв'язуються. Так, центральний стан трифону *а-о+ч* потрапить у другий зліва кінцевий вузол дерева.

Використання запропонованих оцінки кількості трифонів для словника найвживаніших слів української мови та запитань для фонетичних вирішуючих дерев дадуть змогу покращити якість навчання та роботи системи розпізнавання мови.

1. *Kapadia S.* Discriminative training of hidden Markov models. *P*hD thesis, Cambridge University, 1998.
2. *Рашкевич Ю.М.* Перетворення часового масштабу мовних сигналів. Львів, 1997.
3. *Hain T., Woodland P.C., Niesler T.R., Whittaker E.W.D.* The 1998 HTK system for transcription of conversational telephone speech// Proc. ICASSP'99, pp.57-60, Phoenix.
4. *Hain T., Woodland P.C., Evermann G., Povey D.* The CU-HTK March 2000 Hub 5E transcription system. Proc// Speech Transcription Workshop. College Park, 2000.

УДК 678.067.5

ОПТИМІЗАЦІЯ ПАРАМЕТРІВ СИСТЕМ ПОЛІГРАФІЧНОГО ОБЛАДНАННЯ В СЕРЕДОВИЩІ MATLAB

© Богдан Дурняк, Олександр Тимченко
Українська академія друкарства, м. Львів, вул. Підголоско, 19

Проведено оптимізацію параметрів системи автоматичного регулювання натягу стрічкового матеріалу прямої дії флексографської друкарської машини за допомогою пакета MATLAB.