ПОРОДЖЕННЯ, ОБЧИСЛЕННЯ ПАРАМЕТРІВ ТА ВІДБІР МОДЕЛЕЙ ФОНЕМ НА ЕТАПІ РОЗВ'ЯЗАННЯ ЗАДАЧІ НАВЧАННЯ

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

АНОТАЦІЯ

Розглядається процес генерації моделей на основі наявної інформації про навчальну вибірку. Виходячи з постановки задачі навчання розпізнаванню сигналів мовлення обчислюються параметри будь-якої моделі. Внаслідок великої кількості можливих моделей пропонується критерій оцінки їхньої відповідності сигналам мовлення.

1. ВСТУП

В розпізнаванні сигналів мовлення використовують різні методи. Характерною рисою методу, що буде описаний далі, є ієрархічний принцип формування модельних сигналів та їх порівняння з пред'явленим для розпізнавання сигналом. На першому рівні використовують моделі, що відповідають фонемам, на другому – словам тощо. Але основою методу є нижній рівень – рівень фонем, який слід детальніше опрацювати.

2. ОПИС МОВНОГО СИГНАЛУ

При розпізнаванні оперують, як правило, не з початковим мовним сигналом, котрий отримують на виході мікрофона, а з так званим описом мовного сигналу. Після попередньої обробки мовний сигнал буде представляти собою послідовність елементів-скалярів $J_{0l} = (j_1, j_2, ..., j_s, ..., j_l)$, l – довжина мовного сигналу. Підпослідовності елементів (сегменти) $J_{\mu\upsilon} = (j_{\mu+1}, j_{\mu+2}, ..., j_\upsilon), 0 \le \mu < \upsilon \le l$, спостережуваного сигналу J_{0l} розглядаються як реалізації образів першого рівня ієрархії – фонем. Образами другого рівня ієрархії будуть слова, третього – речення. Образи другого й старшого рівня ієрархії задаються транскрипціями в алфавіті образів на одиницю меншого. Так, будь-яке слово задається фонетичною транскрипцією:

$$k^2 = (k_1^1, k_2^1, ..., k_s^1, ..., k_{q(k^2)}^1),$$

де $k^2 \in K^2$ - слово k^2 зі словника слів K^2, k_s^1 - образ першого рівня (фонема) з алфавіту фонем K^1, котра займає s-те місце в транскрипції слова k^2, $q(k^2)$ - довжина транскрипції слова k^2 (кількість фонем у слові).

Рішення про образи приймається за методом найбільшої правдоподібності. Так, якщо спостережуваний сигнал J_{0l} є реалізацією слова зі словника K^2, то вирішувальне правило задається виразом:

$$k^2(J_{0l}) = \arg\max_{k^2 \in K^2} \max_{\{\mu_s\}} \prod_{s=1}^{q(k^2)} P(J_{\mu_{s-1}\mu_s} / k_s^1),$$

де $\{\mu_s\}$ - можливі границі сегментів фонем в сигналі J_{0l} згідно транскрипції слова k^2: $\mu_0 = 0, \mu_{q(k^2)} = l, \mu_{s-1} < \mu_s, s = \overline{1; q(k^2)}$,

$T_{\min}(k_s^1) \le \mu_s - \mu_{s-1} \le T_{\max}(k_s^1)$.

Отже, в цій моделі необхідно задати ймовірнісні розподіли $P(J_{\mu\upsilon} / k^1)$ сегментів $J_{\mu\upsilon}$ для всіх фонем $k^1 \in K^1$, а також обмеження довжин $(T_{\min}(k^1), T_{\max}(k^1))$ сегментів фонем.

3. МОДЕЛІ СЕГМЕНТІВ ФОНЕМ

Сегменти фонем задаються стохастичними автоматними породжувальними граматиками [3]. Ці граматики (моделі) можуть мати різну складність, що визначається кількістю станів. Наприклад, модель з одним станом графічно зобразиться наступним чином:

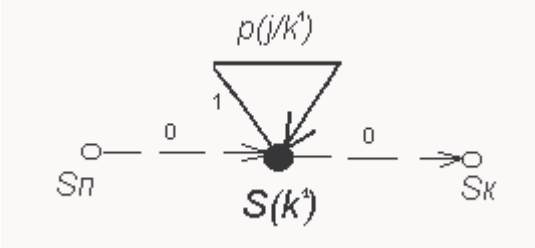


Рис.1.

де: *Sп* – початковий стан моделі;
Sк – кінцевий стан моделі;
$S(k^1)$ - основний стан моделі фонеми k^1.

Переходи виконуються за стрілками в дискретному часі: за пунктирної – за 0 тактів часу, за неперервної – за 1 такт; при переході за один такт генерується елемент j з ймовірністю $p(j/k^1)$. Параметрами цієї моделі`` є ймовірнісний розподіл $p(j/k^1), j = 1:J$, й обмеження довжин генерованого сегмента фонеми k^1 -

$(T_{\min}(k^1), T_{\max}(k^1))$. При цьому ймовірність сегмента $J_{\mu\upsilon}$ при умові фонеми k^1 й незалежності спостережень елементів j обчислюється за виразом:

$$P(J_{\mu\upsilon}/k^1)=\begin{cases}\prod_{i=\mu+1}^{\upsilon} p(j_i/k^1), якщо \quad T_{\min}(k^1)\le \upsilon-\mu\le T_{\max}(k^1);\\ 0, \quad в \quad інших \quad випадках.\end{cases}$$

Модель з двома станами має вигляд:

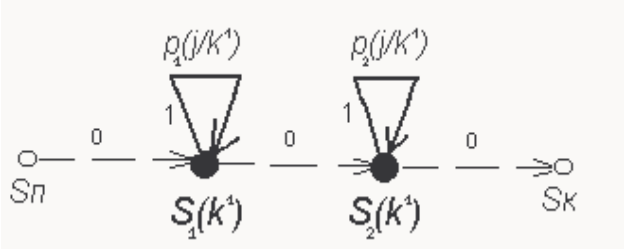


Рис.2.

Ця модель має два послідовних основних стани $S_1(k^1)$ й $S_2(k^1)$. Це значить, що стан $S_1(k^1)$ генерує першу частину (першу фазу) сегмента $J_{\mu\upsilon}$ фонеми k^1, а стан $S_2(k^1)$ - другу частину (другу фазу). Кожен з цих двох станів має свій ймовірнісний розподіл $p_1(j/k^1)$ й $p_2(j/k^1)$, $j=1:J$, а також свої обмеження довжин - $(T_{\min,1}(k^1), T_{\max,1}(k^1))$ й $(T_{\min,2}(k^1), T_{\max,2}(k^1))$. Для цієї моделі ймовірність сегмента $J_{\mu\upsilon}$ при умові фонеми k^1 запишеться наступним чином:

$$P(J_{\mu\upsilon}/k^1)=\begin{cases}\max\limits_{y\le v\le z}(\prod_{i=\mu+1}^{v} p_1(j_i/k^1)\times\prod_{i=v+1}^{\upsilon} p_2(j_i/k^1)),\\ якщо \quad T_{\min,1}(k^1)+T_{\min,2}(k^1)\le \upsilon-\mu\le\\ \le T_{\max,1}(k^1)+T_{\max,2}(k^1);\\ 0, \quad в \quad інших \quad випадках.\end{cases}$$

де:

$y=\max(T_{\min,1}(k^1);(\upsilon-\mu)-T_{\max,2}(k^1))+\mu$,

$z=\min(T_{\max,1}(k^1);(\upsilon-\mu)-T_{\min,2}(k^1))+\mu$.

Наступна за складністю модель – з трьома станами:

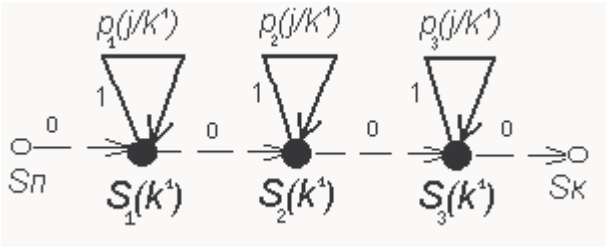


Рис.3.

Для цієї моделі маємо формулу:

$$P(J_{\mu\upsilon}/k^1)=\begin{cases}\max\limits_{\substack{y_1\le v_1\le z_1\\ y_2\le v_2\le z_2}}(\prod_{i=\mu+1}^{v_1} p_1(j_i/k^1)\times\prod_{i=v_1+1}^{v_2} p_2(j_i/k^1)\times\prod_{i=v_2+1}^{\upsilon} p_3(j_i/k^1)),\\ якщо\\ T_{\min,1}(k^1)+T_{\min,2}(k^1)+T_{\min,3}(k^1)\le \upsilon-\mu\le\\ \le T_{\min,1}(k^1)+T_{\min,2}(k^1)+T_{\min,3}(k^1)\\ 0, \quad в \quad інших \quad випадках\end{cases} \qquad (1)$$

де:

$y_1=\max(T_{\min,1}(k^1);(\upsilon-\mu)-T_{\max,2}(k^1)-T_{\max,3}(k^1))+\mu$,

$z_1=\min(T_{\max,1}(k^1);(\upsilon-\mu)-T_{\min,2}(k^1)-T_{\min,3}(k^1))+\mu$,

$y_2=\max(T_{\min,2}(k^1);(\upsilon-\mu-v_1)-T_{\max,3}(k^1))+\mu+v_1$,

$z_2=\min(T_{\max,2}(k^1);(\upsilon-\mu-v_1)-T_{\min,3}(k^1))+\mu+v_1$.

Складніші моделі – з чотирма, п'ятьма станами тощо. Кожна модель буде характеризуватися своєю кількістю станів й, тим самим, відповідною кількістю своїх параметрів (ймовірнісні розподіли та обмеження довжин підсегментів). Отже, при застосуванні будь-якої певної моделі постає питання визначення (оцінки) саме цих параметрів.

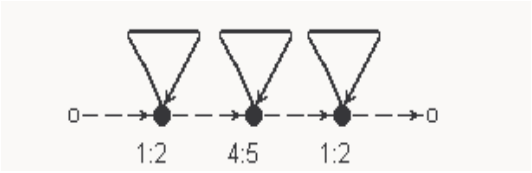
4. ПОРОДЖЕННЯ МОДЕЛЕЙ ФОНЕМ

Обмеження довжин сегментів фонем визначаються для моделі з одним станом безпосередньо з навчальної вибірки (НВ). Навчальна вибірка – наговорений текст у мікрофон, накопичений на твердих носіях. НВ експертом розмічається на сегменти, що відповідають фонемам. Кожну фонему з НВ буде представляти декілька сегментів. З аналізу довжин цих сегментів статистично визначаються обмеження довжин сегментів всіх фонем, а саме $(T_{\min}(k^1), T_{\max}(k^1))$, $k^1\in K^1$. Подальше породження складніших моделей фонем (з двома, трьома станами тощо) буде пов'язане з цими визначеними на першому кроці параметрами $(T_{\min}(k^1), T_{\max}(k^1))$ фонеми k^1. В процесі генерації моделей необхідно обов'язково дотримуватися умови відповідності довжин:

$$\sum_{i=1}^{m} T_{\min,i}(k^1)=T_{\min}(k^1), \qquad \sum_{i=1}^{m} T_{\max,i}(k^1)=T_{\max}(k^1),$$

m – кількість станів.

При цьому дві моделі з трьома (наприклад) станами й з різними розподілами довжин вважаються різними, наприклад (Рис.5.):



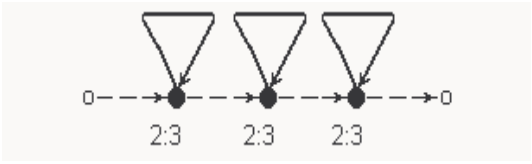


Рис.5.

хоча вони й походять з однієї моделі з одним станом:

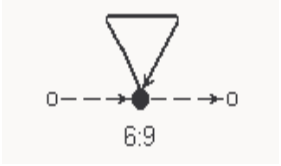


Рис.6.

Отже, з однієї моделі з одним станом, наприклад:

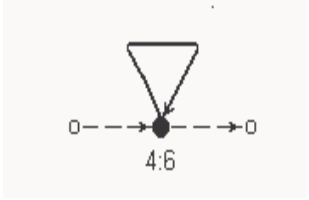


Рис.7.

можна згенерувати 18 моделей з трьома станами(Рис.8.):

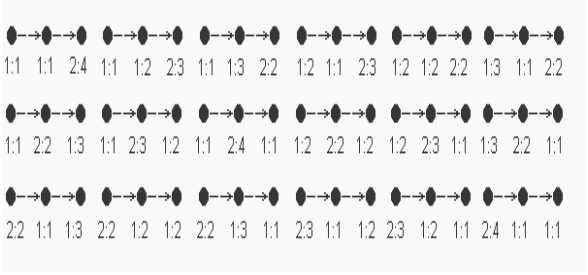


Рис.8.

Взагалі, оскільки ми маємо початкове обмеження довжин $(T_{\min}(k^1),T_{\max}(k^1))$ для фонеми k^1 (модель з одним станом), то загальна кількість моделей з трьома станами визначиться за формулою:

$Q(k^1)=S_{\min}\times S_{\max}$,

де:

$S_{\min}$ - сума арифметичної прогресії з $a_1=1,q=1,$

$S_{\min}=\frac{1+n}{2}\times n, n=T_{\min}(k^1)-2$;

$S_{\max}=\frac{1+n}{2}\times n, n=T_{\max}(k^1)-T_{\min}(k^1)+1.$

5. ОБЧИСЛЕННЯ ЙМОВІРНІСНИХ ПАРАМЕТРІВ МОДЕЛЕЙ.

Процес розв'язання цієї задачі покажемо на прикладі моделей з трьома станами.

Обчислення параметрів моделей фонем необхідно вести згідно постановки задачі навчання[4]. Для найпростішого випадку, коли модель фонеми має один стан, задача навчання формулюється наступним чином:

нехай дана НВ з сегментів $J^r_{\mu^r\upsilon^r}, r=1:U_{k^1}$, з U_{k^1} реалізацій фонеми k^1. Треба знайти такий розподіл $p(j/k^1), j=1:J$, щоб

$$L(k^1,1)=\prod_{r=1}^{U_{k^1}}P(J^r_{\mu^r\upsilon^r}/k^1)=\prod_{r=1}^{U_{k^1}}\prod_{i=\mu^r+1}^{\upsilon^r}p(j_i/k^1)\to\max$$

за умови

$$\sum_{j=1}^{J}p(j/k^1)=1,\quad 0\le p(j/k^1)\le 1, j=1:J .$$

Вище було сказано, що параметри $(T_{\min}(k^1),T_{\max}(k^1))$ визначаються безпосередньо з навчальної вибірки : $T_{\min}(k^1)=\min_{r=1:U_{k^1}}(\upsilon^r-\mu^r)$, $T_{\max}(k^1))=\max_{r=1:U_{k^1}}(\upsilon^r-\mu^r)$.

Ймовірнісний розподіл еталонних елементів $p(j/k^1), j=1:J$, за умови даної фонеми обчислюється за формулою:

$$p(j/k^1)=\frac{n(j/k^1)}{\sum_{i=1}^{J}n(i/k^1)},\quad j=1:J,$$

де: $n(j/k^1)$ - кількість j-го еталонного елемента в сегментах фонеми k^1,

$\sum_{i=1}^{J}n(i/k^1)$ - загальна кількість еталонних елементів в сегментах фонеми k^1,

тобто, максимум $L(k^1,1)$ досягається, коли $p(j/k^1), j=1:J$, - це частоти зустрічаємості елементів $j, j=1:J,$ в сегментах фонеми k^1 (це доводиться шляхом застосування функції Лагранжа). Маючи окремі сегменти фонеми k^1 з НВ, можна обчислити ці частоти, при цьому чим більше сегментів певної фонеми, тим краща статистика.

Для випадку, коли модель фонеми має більше одного стану, задача навчання сформулюється:

нехай дана НВ з сегментів $J^r_{\mu^r\upsilon^r}, r=1:U_{k^1}$, з U_{k^1} реалізацій фонеми k^1. Треба знайти такий розподіл $p_s(j/k^1), j=1:J, s=1:m, m$ – кількість станів в моделі фонеми, щоб

$$L(k^1,m)=\prod_{r=1}^{U_{k^1}}P(J^r_{\mu^r\upsilon^r}/k^1)\to\max$$

за умови

$$\sum_{j=1}^{J}p_s(j/k^1)=1,\quad 0\le p_s(j/k^1)\le 1, j=1:J, s=1:m .$$

В даному випадку ймовірність сегмента обчислюється за формулою (1). Максимізація відбувається по границях

v_1, v_2. При фіксованих границях v_1, v_2 критерій відповідності $L(k^1, m)$ досягає максимума при ймовірнісному розподілі

$$p_s(j/k^1) = \frac{n_s(j/k^1)}{\sum_{i=1}^{J} n_s(i/k^1)}, s = 1:3, \quad j = 1:J,$$

тобто знов при частотах, але вже по кожному стану окремо. Таким чином, $L(k^1, m)$ є функцією границь розбиття кожного сегмента фонеми k^1, адже

$$L(k^1, m) = \prod_{r=1}^{U_{k^1}} P(J^r_{\mu^r v^r} / k^1) =$$

$$= \prod_{r=1}^{U_{k^1}} \max_{v_1^r, v_2^r} \left(\prod_{i=\mu^r+1}^{v_1^r} p_1(j_i/k^1) \times \prod_{i=v_1^r+1}^{v_2^r} p_2(j_i/k^1) \times \prod_{i=v_2^r+1}^{v^r} p_3(j_i/k^1) \right).$$

Отже, пропонується алгоритм оптимального розбиття сегментів фонем на підсегменти для визначення ймовірнісних параметрів моделей фонем (на прикладі моделей з трьома станами):

1. Генеруємо повний набір всіх можливих моделей фонеми k^1.
2. Для кожної моделі:
   2.1. Розбиваємо всі U_{k^1} сегментів фонеми k^1 на 3 підсегменти довільно, але щоб виконувалась умова обмеження довжин відповідно моделі (початкове розбиття).
   2.2. Для кожного сегмента фонеми k^1 (по порядку, починаючи з першого):
   розглядаємо всі можливі розбиття відповідно до моделі, для кожного розбиття підраховуємо розподіли $p_s(j/k^1)$, $j = 1:J$, $s = 1:3$, й підраховуємо значення $L(k^1, 3)$. Фіксуємо те розбиття, при якому досягається максимум $L(k^1, 3)$.
   2.3. Пункт 2.2. повторюємо доти, поки не перестане змінюватися розбиття кожного сегмента.
   (Для цієї певної моделі, таким чином, зроблено оптимальне розбиття сегментів на підсегменти й тим самим обчислили ймовірнісні розподіли $p_s(j/k^1), j \in J, s = 1:3$).
3. Серед повного набору всіх можливих моделей фонеми k^1 та модель буде найкраща, на якій буде досягатися найбільше значення $L(k^1, 3)$.

6. ЕКСПЕРИМЕНТАЛЬНА ЧАСТИНА

Поданий алгоритм був запрограмований. Нижче наводиться приклад оптимального розбиття 19 сегментів фонеми $S\_$ згідно моделі **(1:2; 3:5; 1:2)** : $T_{\min}(S\_) = 5, T_{\max}(S\_) = 9$, кількість можливих моделей $Q(S\_) = \left[\frac{1+3}{2} \times 3\right] \times \left[\frac{1+5}{2} \times 5\right] = 6 \times 15 = 90$.

Довжина|	Сегменти $J_{\mu^r v^r}, r = 1:19$ (послідовності еталонних елементів, J=128)																		
8	5		83	14	7	7	7								14	46			
6	5		14	14	7										7	5			
9	2	5	15	83	83	83	83								5	17			
7	2	5	14	14	66	15									17				
5	17		15	14	15										7				
6	1		2	15	17										17	18			
6	14		14	14	15	14									17				
8	1	2	5	5	5	5									1	1			
6	5		14	5	2										1	1			
6	5		14	7	7										7	7			
6	5		15	15	7										7	7			
7	17	17	14	14	15	14									17				
6	14	17	14	15	14										2				
7	1	2	5	5	5										7	7			
6	14		14	15	15										17	17			
6	17		15	117	117	5									5				
6	15		83	46	46										7	14			
8	5		7	7	7	7	7								46	46			
7	14		83	7	83	66	15								17				

Початкове значення критерію (для моделі з одним станом) - $\ln(L(S\_,1)) = -277{,}05$, значення критерію для цієї моделі - $\ln(L(S\_,3)) = -234{,}91$. Значення критерію збільшилося (покращало).

7. ВИСНОВОК

Запропонований алгоритм дозволяє для кожної фонеми побудувати, оцінити й відібрати кращі моделі для подальшого використання в розпізнаванні сигналів мовлення.

6. ЛІТЕРАТУРА

1. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. – Киев: Наукова думка, 1987, 264с.
2. Винцюк Т.К. Сравнительный теоретический анализ ИКДП- и НММ-методов распознавания речи. – Автоматическое распознавание слуховых образов : 15-й Всесоюзный семинар. – Таллинн, 1989, С.18-24.
3. Вінцюк Т.К., Юхименко О.А. Ймовірнісні моделі фонем та критерії їх вибору. – Ймовірнісні моделі та обробка випадкових сигналів і полів: міжнародний симпозіум. – Тернопіль, 1992, ч.2, С.28-32.
4. Вінцюк Т.К., Юхименко О.А. Робастні постановки задачі навчання розпізнаванню сигналів мовлення. – Обробка сигналів і зображень та розпізнавання образів: Перша всеу українська конференція. – Київ, 1992, С.78-80.