ОЦІНКА ЕФЕКТИВНОСТІ МОДЕЛЕЙ ФОНЕМ З ВРАХУВАННЯМ ЗНАЧЕНЬ КРИТЕРІЯ НАВЧАННЯ ПО ДОВІРЧИМ ОБЛАСТЯМ

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

АНОТАЦІЯ

Пропонується для кожної фонеми підібрати свою певну модель з метою підвищення надійності розпізнавання. Виходячи з постановки задачі навчання розпізнаванню сигналів мовлення обчислюються параметри цих моделей. Внаслідок великої кількості можливих моделей пропонується правило оцінки їхньої відповідності сигналам мовлення.

1. ВСТУП

Метод, про який буде йти далі мова, використовує ієрархічний принцип формування модельних сигналів та їх порівняння з пред'явленим для розпізнавання сигналом[1,2]. На першому рівні використовують моделі, що відповідають фонемам, на другому – словам тощо. Але основою є перший рівень – рівень фонем, який слід детальніше опрацювати.

2. МОВНИЙ СИГНАЛ. ОПИС

В основі опису мовного сигналу буде лежати ієрархічна модель.

Після попередньої обробки вхідного аналогового сигналу й векторного квантування простору сигналів мовлення 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) $J_{\mu\upsilon} = (j_{\mu+1}, j_{\mu+2}, ..., j_\upsilon), 0 \le \mu < \upsilon \le l$, $((\upsilon - \mu)$ - довжина сегмента) спостережуваного сигналу J_{0l} розглядаються як реалізації образів першого рівня ієрархії – фонем. Образами другого рівня будуть слова, третього – речення. Образи другого й старшого рівня ієрархії задаються транскрипціями в алфавіті образів на одиницю меншого. 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 сегментів фонем $(T_{\min}(k^1), T_{\max}(k^1))$.

3. МОДЕЛІ СЕГМЕНТІВ ФОНЕМ

Сегменти фонем задаються стохастичними автоматними породжувальними граматиками [4]. Ці граматики (моделі) можуть мати різну складність, що визначається кількістю станів – одним, двома, трьома тощо. При цьому ймовірність сегмента $J_{\mu\upsilon}$ при умові фонеми k^1 й незалежності спостережень еталонних елементів j обчислюється за виразом:

1) для моделі з одним станом -

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

де $p(j / k^1)$ - ймовірність спостереження еталонного елемента j за умови фонеми $k^1, j = 1:J$;

2) взагалі для моделі з m станами -

$$P(J_{\mu\upsilon} / k^1) = \begin{cases} \max_{v_d, d=0:m} (\prod_{s=1}^{m} \prod_{i=v_{s-1}+1}^{v_s} p_s(j_i / k^1)), \\ якщо \quad \sum_{s=1}^{m} T_{\min,s}(k^1) \le \upsilon - \mu \le \sum_{s=1}^{m} T_{\max,s}(k^1) \\ 0, \quad в \ інших \ випадках. \end{cases} \quad (2)$$

де:

$p_s(j/k^1)$ - ймовірність спостереження еталонного елемента j по s -му стану, $j=1:J$;

$v_d, d=0:m,$ - границі розбиття сегмента $J_{\mu\upsilon}$ на підсегменти, повинні знаходитись в рамках обмеження довжин по станам моделі - $(T_{\min,s}(k^1),T_{\max,s}(k^1)), s=1:m$, $v_0=\mu, v_m=\upsilon$ [4].

Кожна модель буде характеризуватися своєю кількістю станів й, тим самим, відповідною кількістю своїх параметрів (ймовірнісні розподіли та обмеження довжин підсегментів), своєю формулою для обчислення ймовірності сегментів $P(J_{\mu\upsilon}/k^1)$. Отже, при застосуванні будь-якої певної моделі постає питання визначення (оцінки) її параметрів.

Обмеження довжин сегментів фонем $(T_{\min}(k^1),T_{\max}(k^1)), k^1\in K^1$, 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. Подальше породження складніших моделей фонем (з двома, трьома станами тощо) буде пов'язане з цими визначеними на першому кроці параметрами $(T_{\min}(k^1),T_{\max}(k^1))$. В процесі генерації моделей необхідно обов'язково дотримуватися умови відповідності довжин:

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

m – кількість станів.

Оскільки ми маємо початкове обмеження довжин $(T_{\min}(k^1),T_{\max}(k^1))$, то моделей з двома станами буде :

$Q_2(k^1)=(T_{\min}(k^1)-1)\times(T_{\max}(k^1)-T_{\min}(k^1)+1)$;

з трьома:

$Q_3(k^1)=S_{\min}\times S_{\max}$,

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

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

тощо.

4. ОБЧИСЛЕННЯ ПАРАМЕТРІВ. ОЦІНКА ТА ВІДБІР МОДЕЛЕЙ

Слід зауважити, що в процесі розв'язання задачі навчання необхідно для кожної фонеми підібрати певну модель згідно з якимсь правилом цього відбору. Для всіх моделей, які можна згенерувати для фонеми k^1, потрібно обчислити їхні ймовірнісні параметри $p_s(j/k^1), j=1:J, s=1:m$, згідно постановки задачі навчання[3], котра 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 $p_s(j/k^1), j=1:J, s=1:m,$ щоб досягався максимум критерія навчання:

$$\prod_{r=1}^{U_{k^1}}P(J^r_{\mu^r\upsilon^r}/k^1)\to\max \qquad (3)$$

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 формулі (3) ймовірність $P(J^r_{\mu^r\upsilon^r}/k^1)$ обчислюється за формулою (2).

Параметри $(T_{\min}(k^1),T_{\max}(k^1))$ визначаються: $T_{\min}(k^1)=\min_{r=1:U_{k^1}}(\upsilon^r-\mu^r)$, $T_{\max}(k^1)=\max_{r=1:U_{k^1}}(\upsilon^r-\mu^r)$.

Функція

$$L(k^1,m)=\max_{\substack{v_i^r,i=0:m\\ \vec{p}_s(k^1),s=1:m}}\prod_{r=1}^{U_{k^1}}P(J^r_{\mu^r\upsilon^r}/k^1)$$

є функцією границь розбиття сегментів $J^r_{\mu^r\upsilon^r}, r=1:U_{k^1}$, фонеми k^1 - $v_i^r, i=0:m$, й розподілу $\vec{p}_s(k^1)=(p_s(1/k^1),p_s(2/k^1),...,p_s(J/k^1)), s=1:m$ [4]. При фіксованих границях $v_i^r, i=0:m$ функція $L(k^1,m)$ досягає максимума при ймовірнісному розподілі

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

де: $n_s(j/k^1)$ - кількість зустрічей елемента j в тих підсегментах сегментів $J^r_{\mu^r\upsilon^r}, r=1:U_{k^1}$, котрі відносяться до s -го стану моделі ;

$\sum_{i=1}^{J}n_s(i/k^1)$ - загальна кількість елементів в цих підсегментах.

Тобто, це будуть частоти зустрічаємості елементів j по станам моделі. Маючи сегменти фонеми k^1 з НВ, можна обчислити ці частоти, при цьому чим більше сегментів даної фонеми, тим краща статистика. Границі розбиття на підсегменти $v_i^r, i=0:m$, повинні знаходитись в рамках обмеження довжин по станам моделі - $(T_{\min,s}(k^1),T_{\max,s}(k^1)), s=1:m$ [4]. Серед повного набору всіх можливих моделей фонеми k^1 та модель буде найкраща, на якій буде досягатися найбільше значення $L(k^1,m)$.

Але цей підхід, гарний на перший погляд, має один тонкий недолік.

Функція $L(k^1,m)$ для всіх моделей з двома станами буде не меншою, ніж $L(k^1,1)$, й обов'язково знайдуться такі моделі, для котрих буде виконуватись $L(k^1,2)>L(k^1,1)$. Серед всіх моделей з трьома станами обов'язково знайдуться такі моделі, для котрих буде виконуватись $L(k^1,3)>L(k^1,2)$. Серед всіх моделей з чотирма станами обов'язково знайдуться такі моделі, для котрих буде виконуватись $L(k^1,4)>L(k^1,3)$ тощо. Покажемо це на прикладі моделей з двома станами. Функція $L(k^1,m)$ для моделі з одним станом запишеться:

$$L(k^1,1)=\max_{\bar{p}(k^1)}\prod_{r=1}^{U_{k^1}}P(J^r_{\mu^r v^r}/k^1)=\prod_{r=1}^{U_{k^1}}(\prod_{i=\mu^r+1}^{v^r}p^*(j_i/k^1)). \quad (4)$$

Оскільки $p^*(j/k^1)=\dfrac{n(j/k^1)}{\sum_{i=1}^{J}n\ (i/k^1)}$, $j=1:J$, то (4) перепишеться :

$$L(k^1,1)=\prod_{j=1}^{J}p^*(j/k^1)^{n(j/k^1)}=\prod_{j=1}^{J}\left(\frac{n(j/k^1)}{n(k^1)}\right)^{n(j/k^1)},$$

де $n(k^1)=\sum_{i=1}^{J}n\ (i/k^1)$.

Для моделі з двома станами функція $L(k^1,m)$ запишеться:

$$L(k^1,2)=\max_{\substack{v^r,r=1:U_{k^1}\\ \bar{p}_1(k^1),\bar{p}_2(k^1)}}\prod_{r=1}^{U_{k^1}}P(J^r_{\mu^r v^r}/k^1)=\prod_{r=1}^{U_{k^1}}(\prod_{i=\mu^r+1}^{v^r}(p_1^*(j_i/k^1)\times$$

$$\times\prod_{i=v^r+1}^{v^r}p_2^*(j_i/k^1)))=\prod_{r=1}^{U_{k^1}}\prod_{i=\mu^r+1}^{v^r}p_1^*(j_i/k^1)\times\prod_{r=1}^{U_{k^1}}\prod_{i=v^r+1}^{v^r}p_2^*(j_i/k^1)=$$

$$=\prod_{j=1}^{J}p_1^*(j/k^1)^{n_1(j/k^1)}\times\prod_{j=1}^{J}p_2^*(j/k^1)^{n_2(j/k^1)}=\prod_{j=1}^{J}[p_1^*(j/k^1)^{n_1(j/k^1)}\times$$

$$\times p_2^*(j/k^1)^{n_2(j/k^1)}]=\prod_{j=1}^{J}\left[\left(\frac{n_1(j/k^1)}{n_1(k^1)}\right)^{n_1(j/k^1)}\times\left(\frac{n_2(j/k^1)}{n_2(k^1)}\right)^{n_2(j/k^1)}\right]$$

де: параметр $v^r, r=1:U_{k^1}$, обчислений так, що досягається максимум в цій формулі;

$n_1(j/k^1)+n_2(j/k^1)=n(j/k^1), n_1(k^1)+n_2(k^1)=n(k^1)$.

Оскільки справедливе твердження:

“Для $\forall m,n,l,k,r,p\in N$:

$l\ge 0, r\ge 0, m<n, l\le k, r\le p, \quad l+r=m, \quad k+p=n,$

виконується нерівність: $\left(\dfrac{m}{n}\right)^m\le\left(\dfrac{l}{k}\right)^l\times\left(\dfrac{r}{p}\right)^r$.

При чому, рівність досягається тоді, коли $\dfrac{m\cdot k}{n}$ - ціле число, й $l=\dfrac{m\cdot k}{n}$”,

(доведення цього факту ми опускаємо за нестачею тут місця)

то

$$\left(\frac{n(j/k^1)}{n(k^1)}\right)^{n(j/k^1)}\le\left(\frac{n_1(j/k^1)}{n_1(k^1)}\right)^{n_1(j/k^1)}\times\left(\frac{n_2(j/k^1)}{n_2(k^1)}\right)^{n_2(j/k^1)},\forall j=1:J,\Rightarrow$$

$$\prod_{j=1}^{J}\left(\frac{n(j/k^1)}{n(k^1)}\right)^{n(j/k^1)}\le\prod_{j=1}^{J}\left(\frac{n_1(j/k^1)}{n_1(k^1)}\right)^{n_1(j/k^1)}\times\prod_{j=1}^{J}\left(\frac{n_2(j/k^1)}{n_2(k^1)}\right)^{n_2(j/k^1)},$$

тобто $L(k^1,1)\le L(k^1,2)$ для будь-якої моделі з двома станами. З цього випливає, що будь-яка модель з двома станами апріорі краща за модель з одним станом.

Далі, якщо звернутися до моделей з двома станами, то серед них буде така, у котрої критерій навчання $L(k^1,2)$ буде максимальним. Якщо для даної фонеми можна генерувати моделі з трьома станами (нагадаємо, що максимальна кількість станів визначається найменшою довжиною сегментів фонеми з НВ - $m=T_{\min}(k^1)$), то якийсь з двох станів обов’язково можна розбити на два (причому на декілька), й тим самим отримати декілька моделей з трьома станами. Застосувавши попередні викладки до цього випадку отримаємо, що критерії навчання всіх цих моделей будуть більшими за максимальний критерій навчання всіх моделей з двома станами (і, можливо, тільки один буде дорівнювати). З усього цього випливає той висновок, що стає заздалегідь ясно – найкраща модель для фонеми буде серед моделей з максимально можливою кількістю станів. Наприклад, якщо початкове обмеження довжин фонеми k^1 є $T_{\min}(k^1)=5, T_{\max}(k^1)=10$, то найкраща модель буде з п’яттю станами.

Але! Спостерігаючи $n(j/k^1)$ разів елементи $j=1:J$ в сегментах фонеми k^1 ($n(k^1)$ - загальна кількість спостережуваних елементів для фонеми k^1 в НВ), маємо справу з поліноміальним (розмірності J) розподілом. Значення $p^*(j/k^1)=\dfrac{n(j/k^1)}{n(k^1)}$, $j=1:J$, - це точкові оцінки ймовірнісних параметрів цього розподілу. Й добре відомо, що чим більша кількість спостережень $n(k^1)$, тим точніші ці самі точкові оцінки. Коли ми заводимо моделі з двома станами, то кількість спостережень автоматично зменшується – ми маємо два поліноміальних розподіла з $n_1(k^1)$ й $n_2(k^1)$ кількістю спостережень, а $n_1(k^1)+n_2(k^1)=n(k^1)$. Чим більше заводимо станів, тим меншими стають кількості спостережень й , тим самим, зростають похибки оцінок ймовірнісних параметрів $p_s(j/k^1), j=1:J, s=1:m$, котрі призведуть до похибок в розпізнаванні. А звідси пропонується дещо уточнений алгоритм відбору моделей:

1. Для фонеми k^1 генеруємо повну кількість всіх можливих моделей.

2. Для всіх цих моделей:

- обчислюємо значення функції $L(k^1,m), m=1:T_{\min}(k^1)$;

- знаходимо довірчі області ймовірнісних параметрів $\Omega_s(k^1), s=1:m$ (наприклад, 67%-ні - σ вліво-вправо);

- знаходимо значення функції $F(k^1,m)=\min_{\bar{p}_s(k^1)\in\Omega_s(k^1),s=1:m}\prod_{r=1}^{U_{k^1}}P(J^r_{\mu^r v^r}/k^1)$, тобто мінімальне (найгірше) значення критерія навчання по довірчим областям.

3. При переході від моделей з $(m-1)$ станами до моделей з m станами до розгляду приймаємо тільки ті моделі, у котрих не тільки $L(k^1,m)$ більші за максимальне значення $L(k^1,m-1)$, а й $F(k^1,m)$ також більші за це максимальне значення (принцип “найгірше краще за найкраще”). Якщо таких моделей нема – то далі не йдемо, кінець алгоритму, якщо є – то серед них обираємо з максимальним значенням $L(k^1,m)$.

5. ЕКСПЕРИМЕНТАЛЬНА ЧАСТИНА

Для ілюстрації вищесказаного наведемо приклад з експерименту: обмеження довжин фонеми - $T_{\min}(k^1)=3, T_{\max}(k^1)=7,$ фонему представляють 40 сегментів, кількість еталонних елементів $n(k^1)=151$, присутні елементи – 1, 2, 7, 17. Найбільша кількість станів, які можна завести, - три. Кількість моделей з двома станами – 10, з трьома – 15. На Рис.1. зображено графік значень функції $L(k^1,m)$ цих моделей:

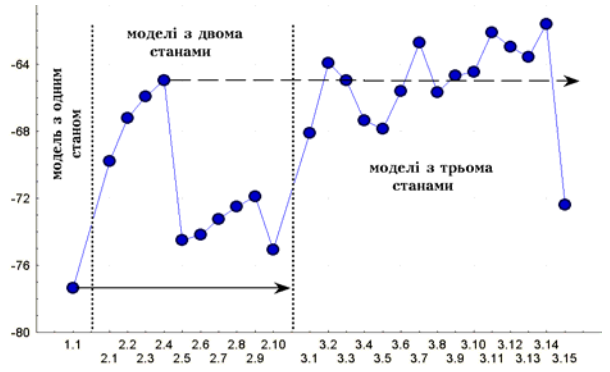

Рис.1.

Моделі з двома станами умовно занумеровані 2.1-2.10, з трьома – 3.1-3.15. Неперервна стрілка показує рівень значення $L(k^1,1)$, пунктирна – рівень максимального значення $L(k^1,2)$ серед моделей з двома станами. З графіка видно, що найкраща модель з двома станами – під номером 2.4, з трьома – 3.14. Всі моделі з номерами 3.1,3.3-3.6,3.8,3.15 безумовно гірші за модель 2.4.

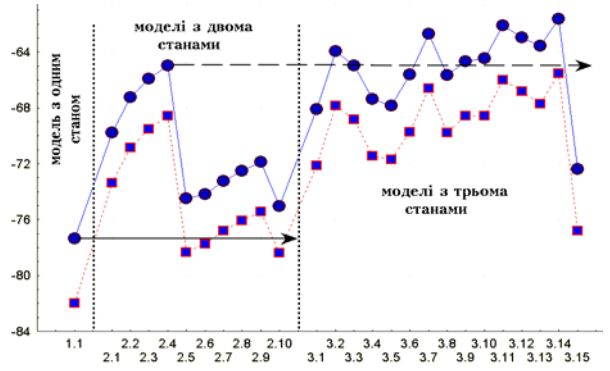

Рис.2.

На Рис.2. зображено той самий графік, але ще й зі значеннями функції $F(k^1,m)$ всіх моделей. Добре видно, що значення функції $F(k^1,3)$ всіх моделей з трьома станами менші за $L(k^1,2)$ найкращої моделі з двома станами – 2.4, для двох станів – тільки значення $F(k^1,2)$ моделей 2.5,2.6,2.10 менші за $L(k^1,1)$. Отже, висновок: для даної фонеми доцільно вибрати модель з двома станами (під номером 2.4 модель з обмеженнями довжин по станам $T_{\min,1}(k^1)=1, T_{\max,1}(k^1)=4, T_{\min,2}(k^1)=2, T_{\max,2}(k^1)=3$).

6. ВИСНОВОК

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

7. 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. Робастні постановки задачі навчання розпізнаванню сигналів мовлення. – Обробка сигналів і зображень та розпізнавання образів: Перша Всеукраїнська конференція. – Київ, 1992, С.78-80.
4. Юхименко О.А. Породження,обчислення параметрів та відбір моделей фонем на етапі розв'язання задачі навчання. - Оброблення сигналів і зображень та розпізнавання образів: Сьома Всеукраїнська міжнародна конференція. – Київ, 2004, С.103-106.