# ВИЗНАЧЕННЯ КІЛЬКОСТІ СТАНІВ В МОДЕЛЯХ ФОНЕМ

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

## АНОТАЦІЯ

Для кожної фонеми можна завести модель з одним станом, двома, трьома станами або більшою кількістю станів. Мета – підвищення надійності розпізнавання мовних сигналів. В збільшенні кількості станів в моделях є свої переваги, але також й недоліки. Пропонується певний компроміс для визначення кількості станів в моделях фонем, а також алгоритм відбору найкращих моделей.

## 1. ВСТУП

Метод, в рамках котрого буде розглядатися це питання, 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. Базовим є перший рівень – рівень фонем, котрому й буде приділена додаткова увага.

## 2. МОВНИЙ СИГНАЛ

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. 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 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 $J_{\mu\upsilon}$ за умови фонеми $k^1$ й незалежності спостережень еталонних елементів $j$ обчислюється за виразом:

1) для моделі з одним станом -

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

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 -

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

де:

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

$\vec{v} = (v_0, v_1, ..., v_m)$ - границі розбиття сегмента $J_{\mu\upsilon}$ на $m$ підсегментів, повинні знаходитись в рамках обмеження

довжин $(T_{\min,s}(k^1),T_{\max,s}(k^1))$ по станам моделі $s=1:m$, $v_0=\mu, v_m=\upsilon$ [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.

## 4. АЛГОРИТМ ВИЗНАЧЕННЯ КІЛЬКОСТІ СТАНІВ ТА ВІДБОРУ МОДЕЛЕЙ.

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

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

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

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

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

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;

$\sum\limits_{i=1}^{J}n_s(i/k^1)$ - 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 має один недолік.

В роботі [5] показано, що значення функції $L(k^1,2)$ для всіх моделей з двома станами буде більшою, ніж значення функції $L(k^1,1)$, й, можливо, тільки для однієї моделі з двома станами ці значення будуть однаковими (що в реальності є рідкісним явищем). 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.

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 $p^*(j/k^1)=\frac{n(j/k^1)}{n(k^1)}$, $j=1:J$, - 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.

Взагалі, для чого треба вводити моделі фонем, складніші за модель з одним станом? Очевидно, щоб поліпшити розпізнавання. Та уявімо, що існує всього три фонеми – Ф1,Ф2,Ф3. І нехай в навчальній вибірці кожна фонема представлена такими еталонними елементами (нехай їх буде всього 10):

Ф1 – 1,2,7;

Ф2 – 3,6,9;

Ф3 – 4,5,8,10.

Очевидно, що в даному випадку нема потреби ускладнювати модель (збільшувати кількість станів), бо еталонні елементи, що представляють фонеми, для кожної фонеми свої і моделі з одним станом й так дадуть 100% розпізнавання.

Інша справа, коли в сегментах різних фонем зустрічаються однакові еталонні елементи. Якраз ця ситуація й дає помилку розпізнавання. Введення складніших моделей дозволяє в певній мірі виправити цю помилку. При цьому, нема потреби заводити складніші моделі, якщо вони не зменшують цю помилку, оскільки збільшується кількість оцінюваних параметрів моделей (а це впливає на швидкодію), моделі стають більш детермінованими, а також погіршуються статистичні оцінки ймовірнісних параметрів (збільшується похибка оцінки) внаслідок зменшення кількості спостережень еталонних елементів по станам.

Отже, у відборі моделей фонем (а саме, їхньої складності та параметрів) пропонується орієнтуватися на таку постановку задачі навчання:

нехай дана НВ з сегментів $\{J^r_{\mu' v'}, k^1(r)\}, r=1:U, k^1(r)\in K^1$, з $U$ реалізацій сегментів всіх фонем $k^1\in K^1$. $k^1(r)$ - фонема, до якої належить сегмент з номером *r*. Треба знайти такий розподіл $p_s(j/k^1), j=1:J, s=1:m$, й кількість станів $m$ в моделі для кожної фонеми $k^1\in K^1$, щоб виконувалось якомога більше нерівностей:

$$P(J^r_{\mu' v'}/k^1(r))>P(J^r_{\mu' v'}/t), \forall t: t\in K^1, t\neq k^1(r), r=1:U \qquad (4)$$

за умови

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

Тобто, по всій НВ ймовірність кожного сегмента повинна бути за умови своєї фонеми більша, ніж за умови чужої. Це забезпечить якомога меншу помилку розпізнавання.

Ймовірність сегментів $P(J^r_{\mu' v'}/k^1), k^1\in K^1$, в формулі (4) обчислюється за формулою (2) згідно моделі фонем.

Звідси пропонується наступний алгоритм відбору моделей в процесі розв'язання задачі навчання:

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

2. Для кожної з цих моделей окремо:

а) обчислюємо ймовірнісні параметри $p_s(j/k^1), j=1:J, s=1:m,$ згідно критерія навчання (3) та визначаємо кількість нерівностей, що виконуються по всій НВ;

б) відбираємо всі ті моделі (з одним станом, двома, трьома станами тощо), для котрих виконується найбільша кількість нерівностей по всій НВ;

в) серед цих моделей відбираються моделі з найменшою кількістю станів;

г) серед цієї множини обирається одна модель, для котрої критерій навчання (3) є максимальним. Цю модель будемо вважати найкращою.

Й так робиться для кожної фонеми.

З цього алгоритму випливає , що збільшення кількості станів в моделі не тільки може не поліпшити ситуацію, а навіть й погіршити, бо дасть більшу ймовірність правдоподібності не тільки на сегментах своєї фонеми, а й на сегментах інших фонем.

# 5. ЕКСПЕРИМЕНТАЛЬНА ЧАСТИНА

Для проведення експерименту з НВ було відібрано 6 фонем (позначимо їх як Ф1-Ф6) таких, що в їхніх сегментах було чимало спільних елементів. Характеристика цих фонем:

| фонема | кількість сегментів | $T_{\min}$ | $T_{\max}$ |
|---|---|---|---|
| Ф1 | 7 | 7 | 11 |
| Ф2 | 8 | 9 | 12 |
| Ф3 | 11 | 4 | 10 |
| Ф4 | 7 | 3 | 10 |
| Ф5 | 13 | 4 | 10 |
| Ф6 | 8 | 3 | 11 |

Загальна кількість сегментів в цій (обмеженій) НВ становить 54. На початку роботи алгоритму ( кожна фонема має модель з одним станом) виконувалось 43 нерівності з, відповідно, 54 можливих. В результаті роботи алгоритму стало виконуватись 53 нерівності. Експеримент проводився для моделей з двома та трьома станами. В наступній таблиці наведені дані по моделям для кожної фонеми:

| фонема | кількість моделей з двома станами | кількість моделей з трьома станами | найбільша кількість станів в моделі |
|---|---|---|---|
| Ф1 | 30 | 225 | 7 |
| Ф2 | 32 | 280 | 9 |
| Ф3 | 21 | 84 | 4 |
| Ф4 | 16 | 36 | 3 |
| Ф5 | 21 | 84 | 4 |
| Ф6 | 18 | 45 | 3 |

В результаті навчання відібрали такі моделі:

Ф1 - модель з 2 станами **(4:8; 3:3),**

Ф2 - модель з 3 станами **(6:6; 1:3; 2:3),**

Ф3 - модель з 2 станами **(1:1; 3:9),**

Ф4 - модель з 2 станами **(1:5; 2:5),**

Ф5 - модель з 2 станами **(1:4; 3:6),**

Ф6 - модель з 3 станами **(1:2; 1:6; 1:3).**

Єдина помилка виникає в фонемі Ф5, і навіть модель з трьома станами не виправляє цієї ситуації.

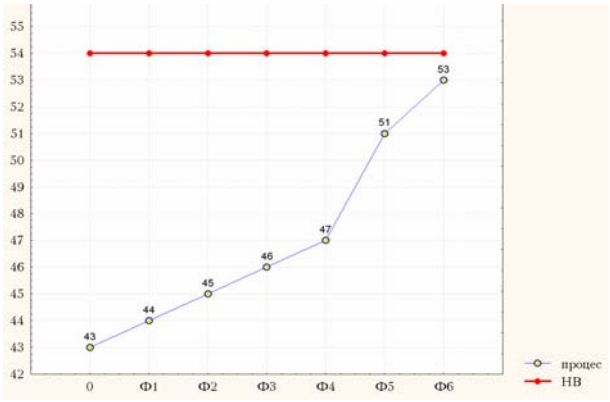


Рис.1.

На Рис.1. зображено графік, що відображає по крокам процесс збільшення кількості нерівностей, що виконуються по всій НВ, після введення найкращої моделі для кожної фонеми по черзі, починаючи з Ф1.

На Рис.2 показано інформацію для фонеми Ф4 після того, як для фонем Ф1,Ф2,Ф3 відібрали найкращі моделі. По осі абсцис йде умовна нумерація моделей (наприклад, 2.8 означає, що це модель з двома станами під номером 8 по порядку серед моделей з двома станами), по осі ординат – значення критерія навчання $L(k^1, m)$. Нижній графік – графік критеріїв навчання для кожної моделі, верхній – скільки нерівностей виконується при даній моделі по всій НВ. Видно, що згідно алгоритму слід вибрати модель з двома станами під номером 2.6 (виконується 47 нерівностей з 54 можливих). Введення моделей з трьома станами не поліпшує ситуації.

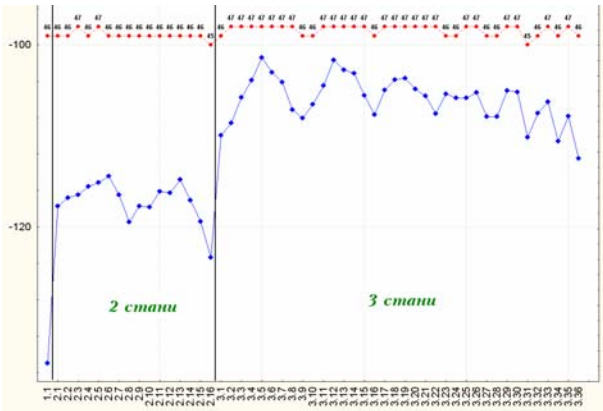


Рис.2.

# 6. ВИСНОВОК

Запропонований алгоритм дозволяє для кожної фонеми визначити кількість станів найкращої 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.
5. Юхименко О.А. Оцінка ефективності моделей фонем з врахуванням значень критерія навчання по довірчим областям. - Оброблення сигналів і зображень та розпізнавання образів: Восьма Всеукраїнська міжнародна конференція. – Київ, 2006, С.83-86.