Багаторівнева багатозначна модель розуміння спонтанного мовлення

Микола Сажок

Відділ розпізнавання та синтезу звукових образів

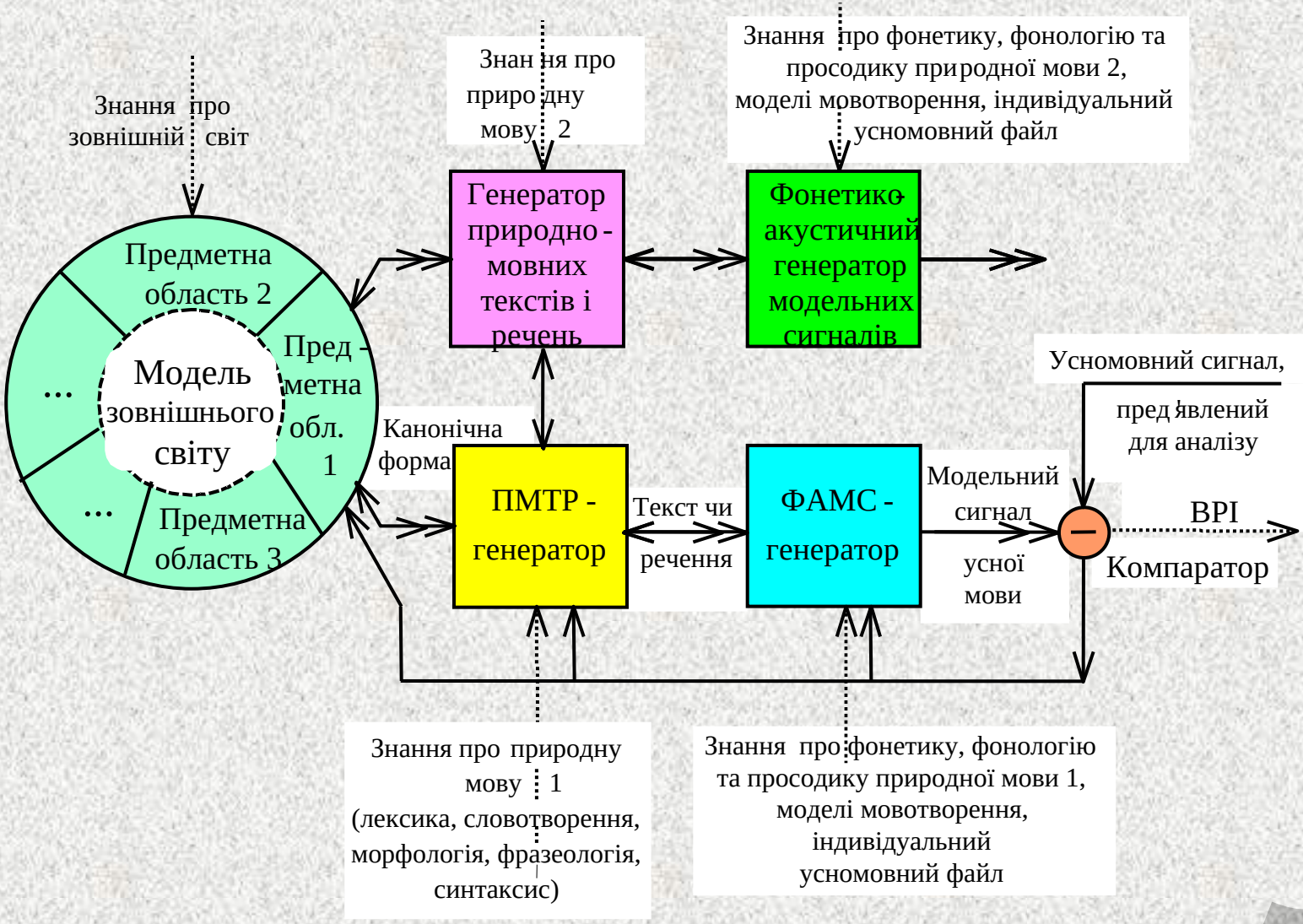
Київ - Жукин

2011

Зміст

- Генеративна модель розпізнавання, розуміння та синтезу мовного сигналу
- Математична модель мовного сигналу
  - Автоматичний фонетичний стенограф
  - Оцінка параметрів моделі на мовному корпусі
- Лінгвістична модель мовного сигналу
  - Модель вимови
  - Обмеження на послідовності слів
- Смислова інтерпретація в межах предметної галузі
- Технічні деталі – особливості реалізації
- Доповнення та відступи

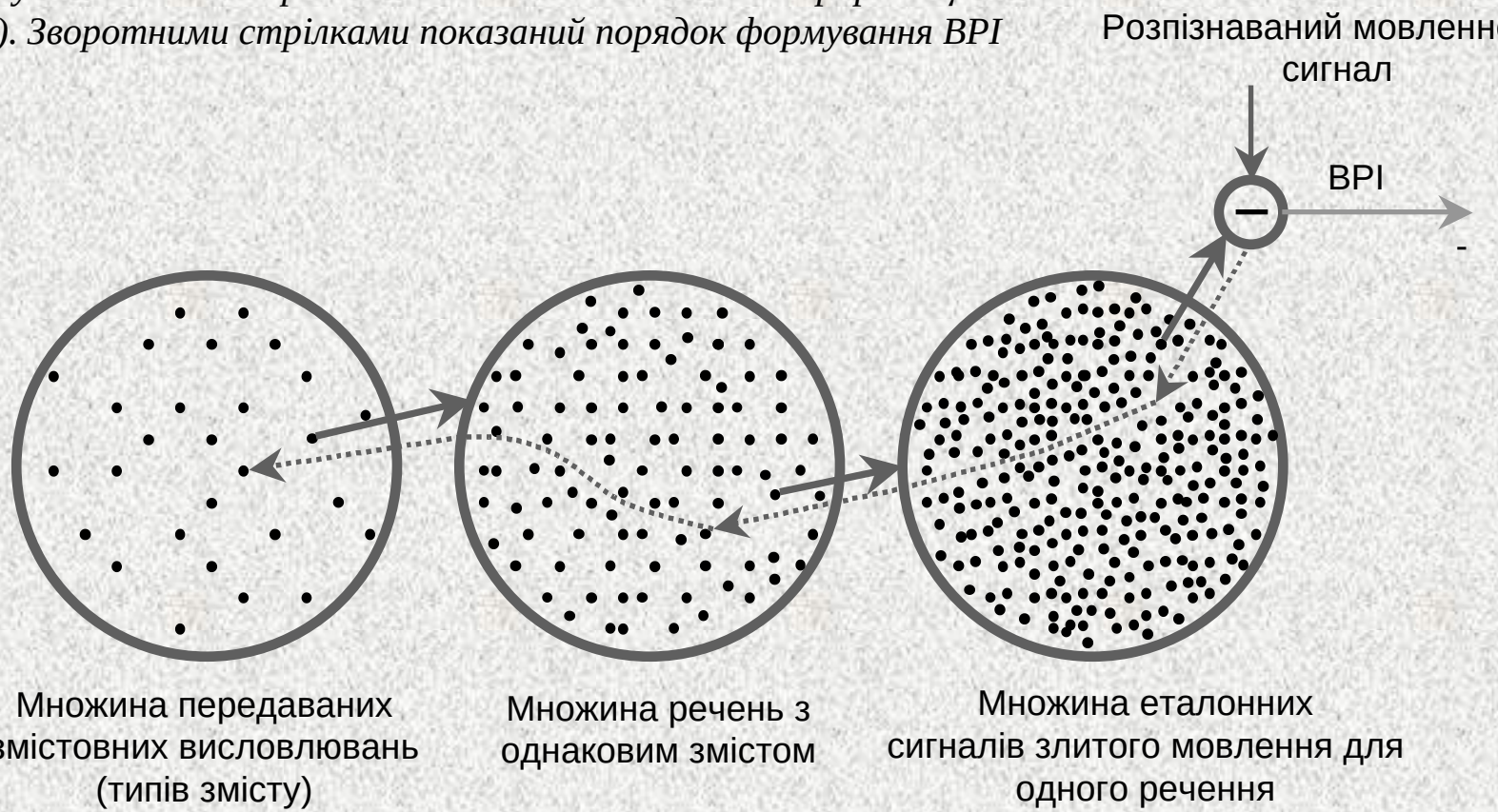
Генеративна модель автоматичного розуміння мовлення



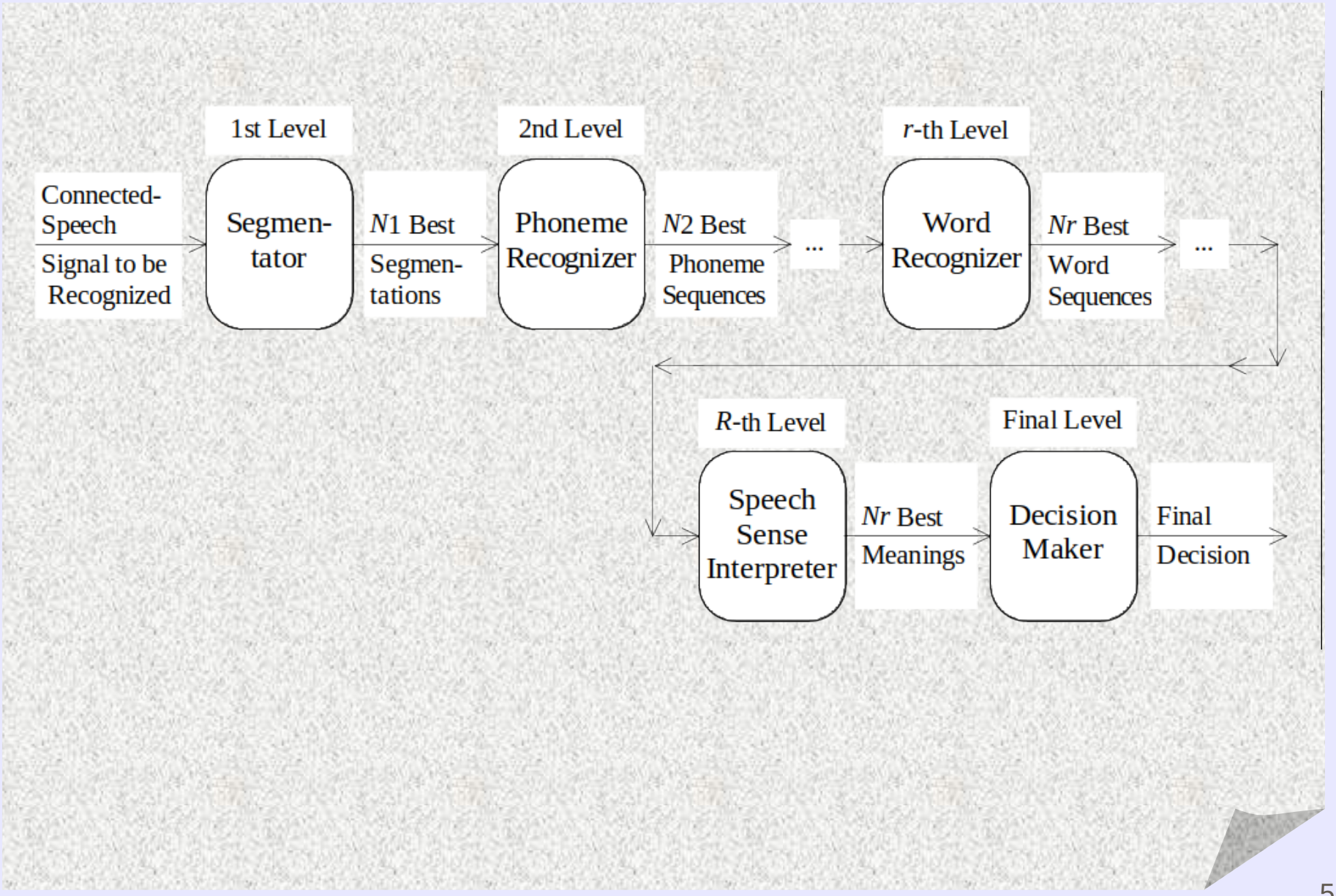
Структура диктувальної машини та машини для усного перекладу.

Генеративна модель автоматичного розуміння мовлення

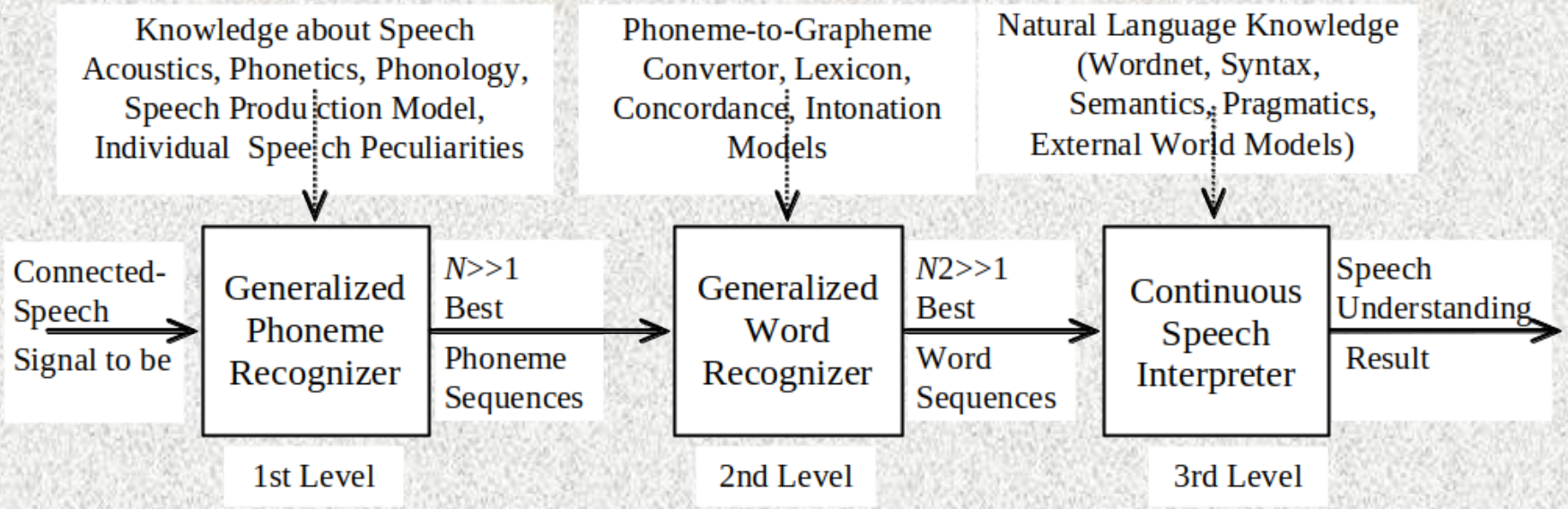
Ілюстрація функціонування генеративних моделей та формування відповіді розпізнавання та смислової інтерпретації (ВРІ). Зворотними стрілками показаний порядок формування ВРІ



The Multi-level Multi-decision ASR&U Structure



Three-Level ASR&U System Structure

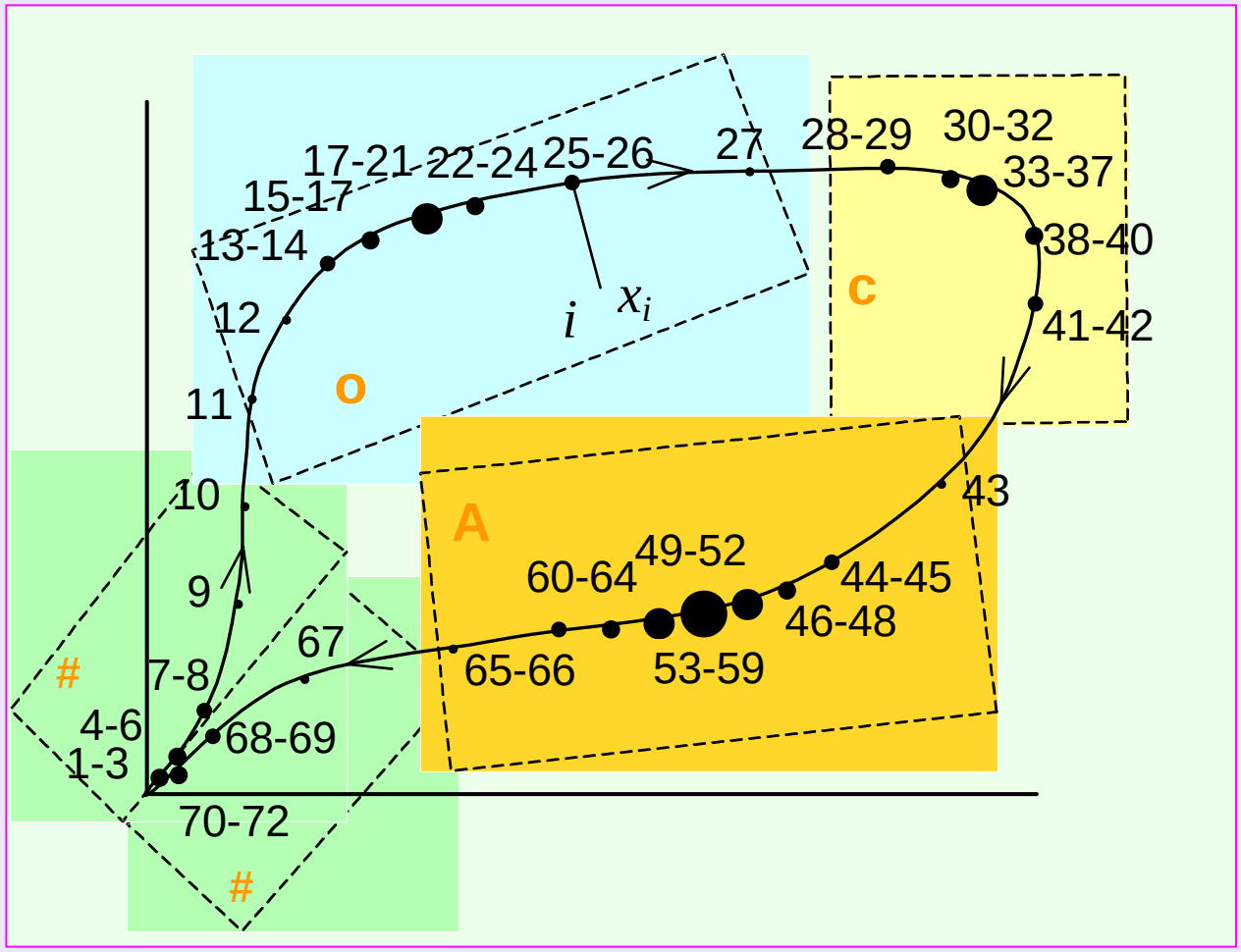


Рівні дають змогу ефективно розподіляти роботу між дослідниками.
Рівні обробляються в єдиному процесі, але форма постпроцесора полегшує реалізацію

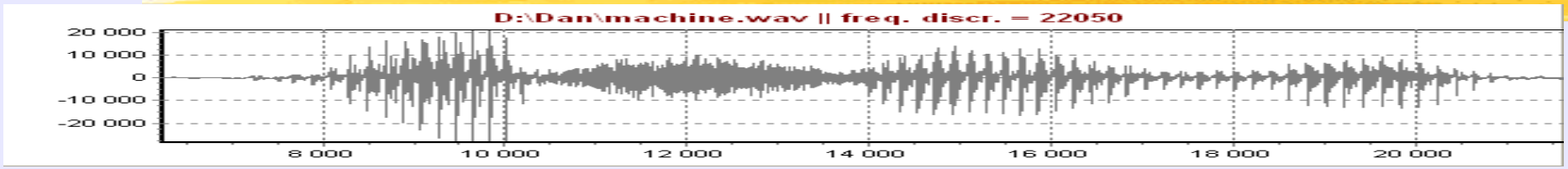
Математичні моделі сигналів фонем

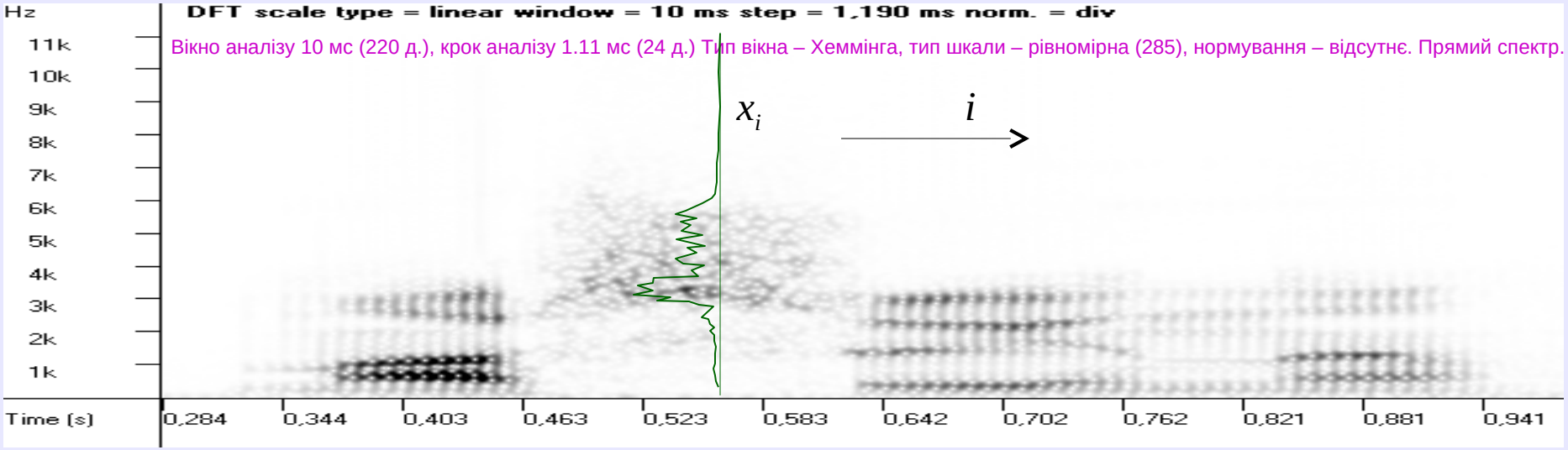
Опис мовленнєвих сигналів послідовностями елементів-векторів

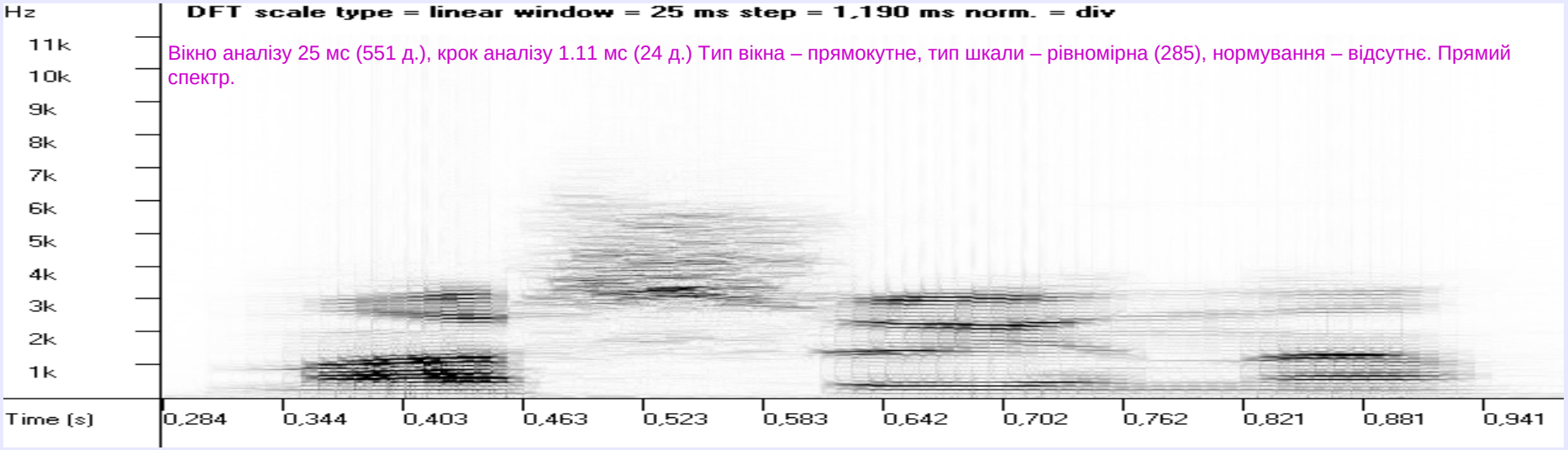
$$X_{ol} = (x_1, x_2, \ldots, x_i, \ldots, x_l)$$



Математичні моделі мовленнєвого сигналу







Математичні моделі мовленнєвого сигналу

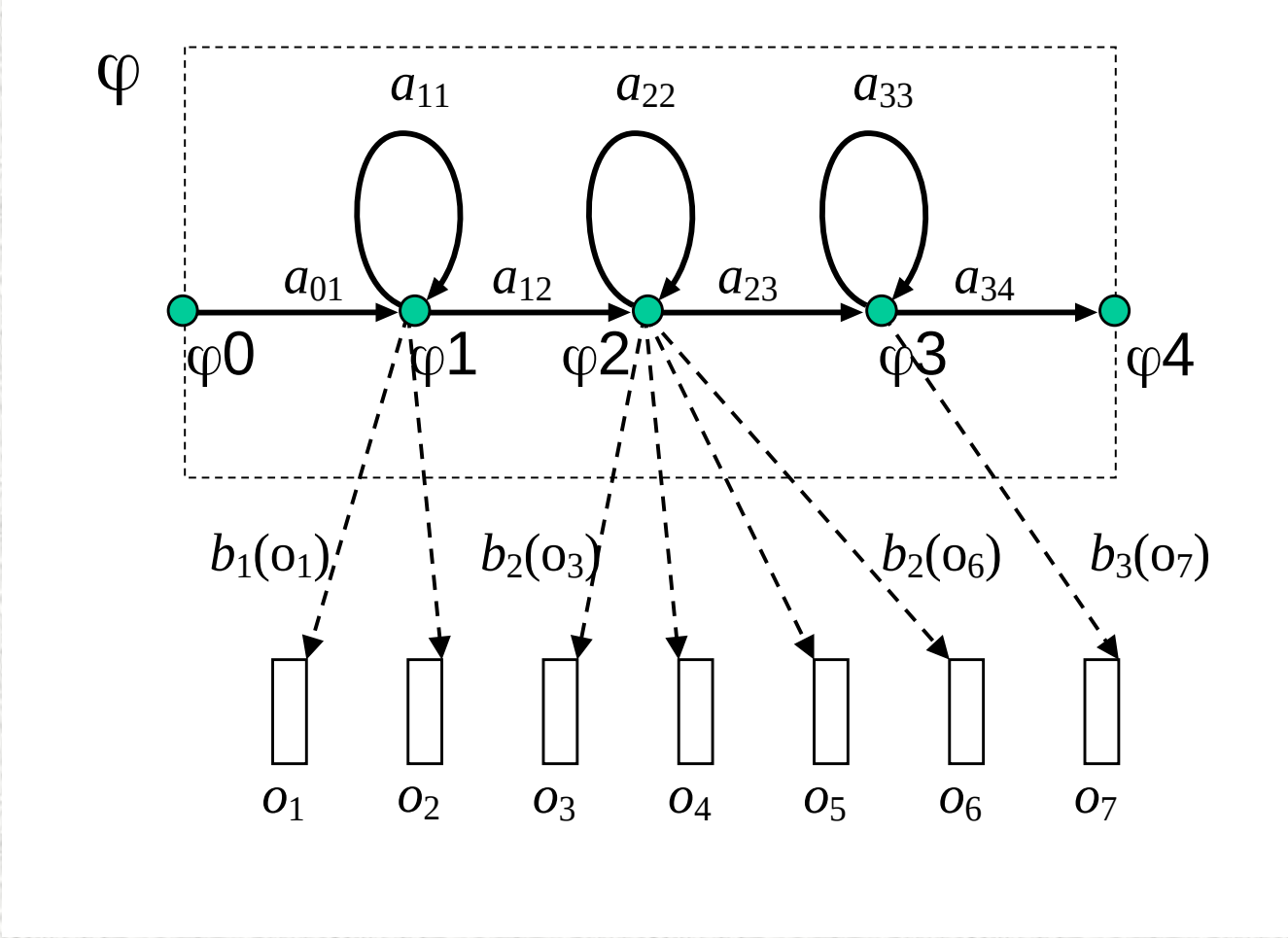
Простори первинного опису мовного сигналу:

- Кепстри (аудиторна модель) – MFCC;
- Лінійне передбачення (артикуляторна модель) - LPCC;
- PLP

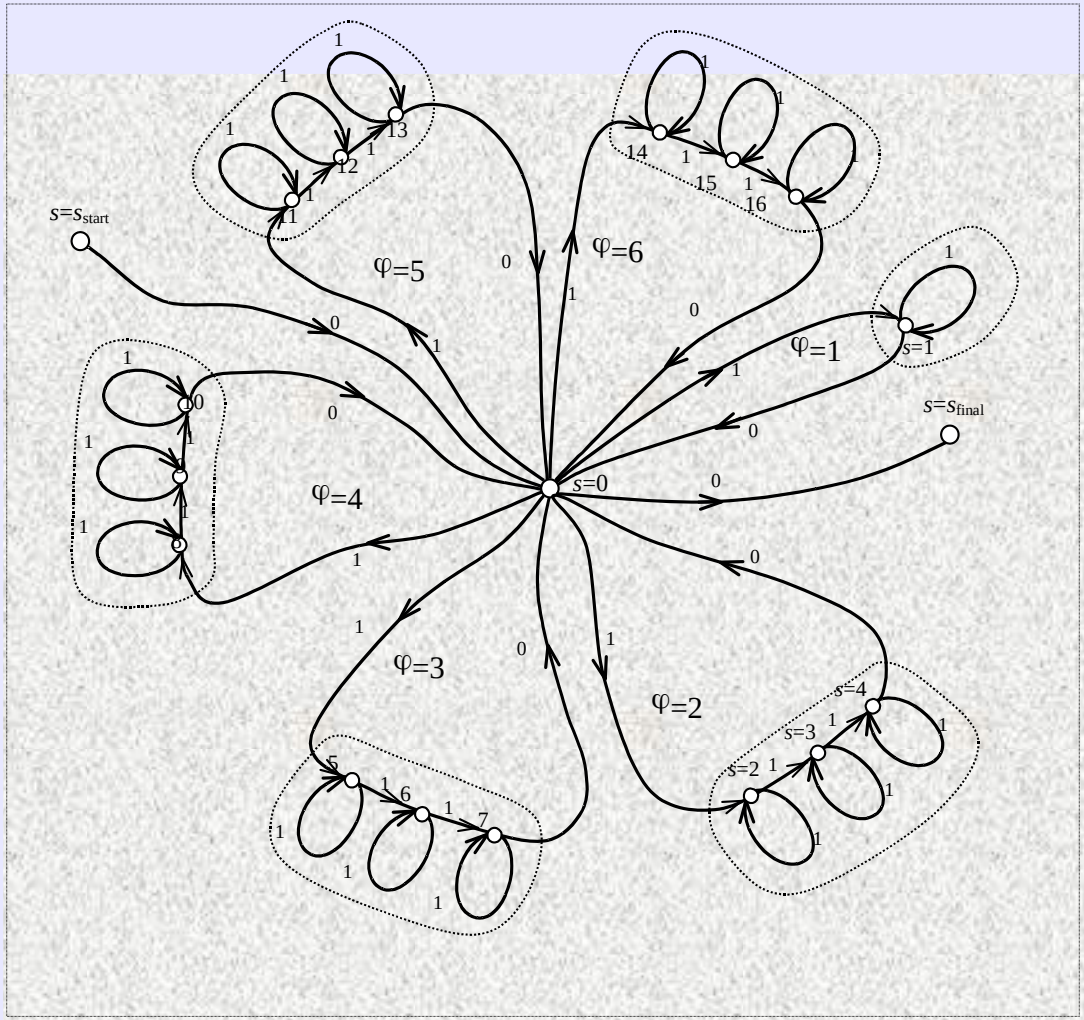
Адаптація моделі до каналу, фону та диктора:

- VTLN;
- MLLR, fMLLR;
- Віднімання середнього кепстру;
- RASTA

Генеративна модель фонемного рівня



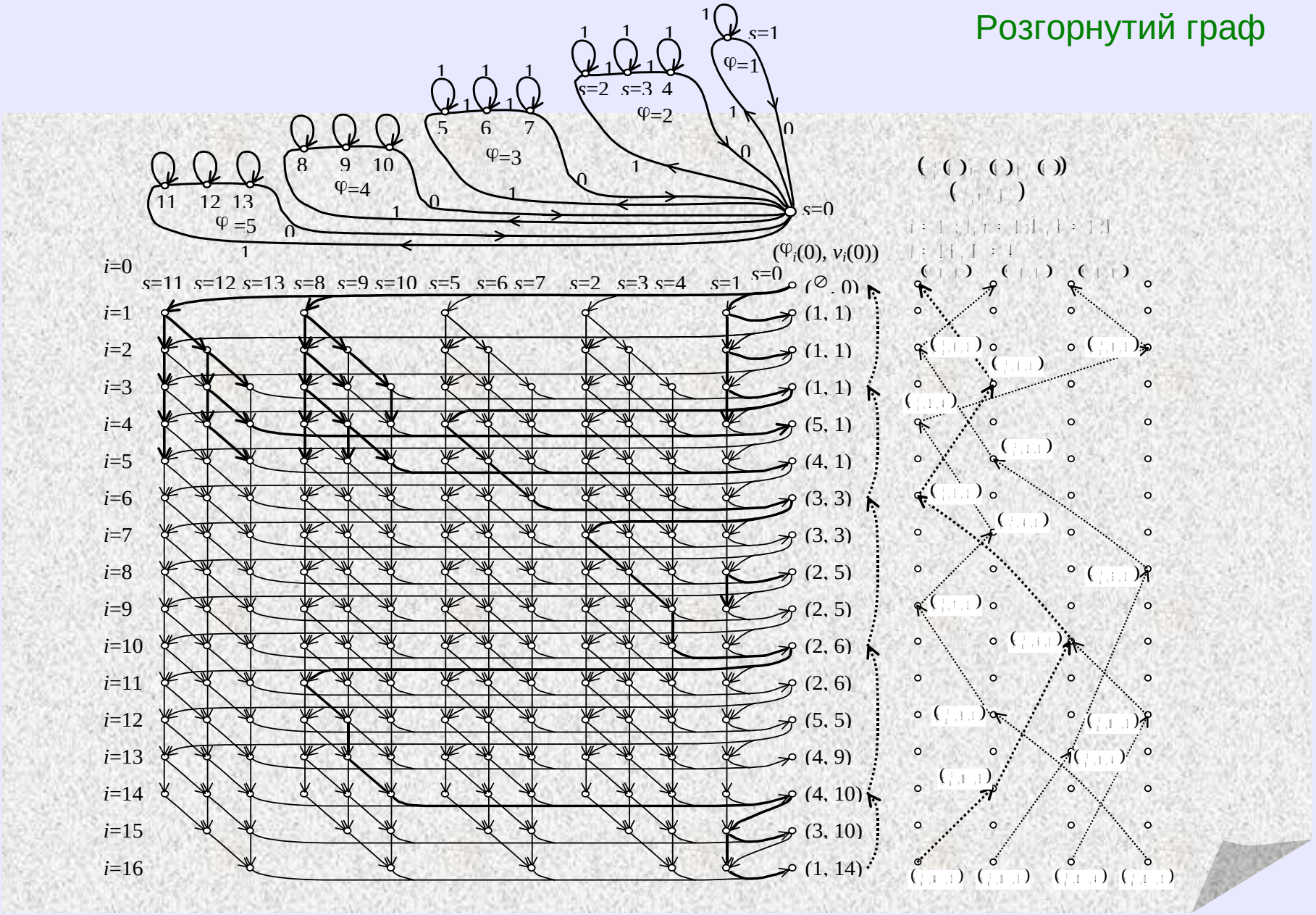
Автоматичний фонетичний стенограф



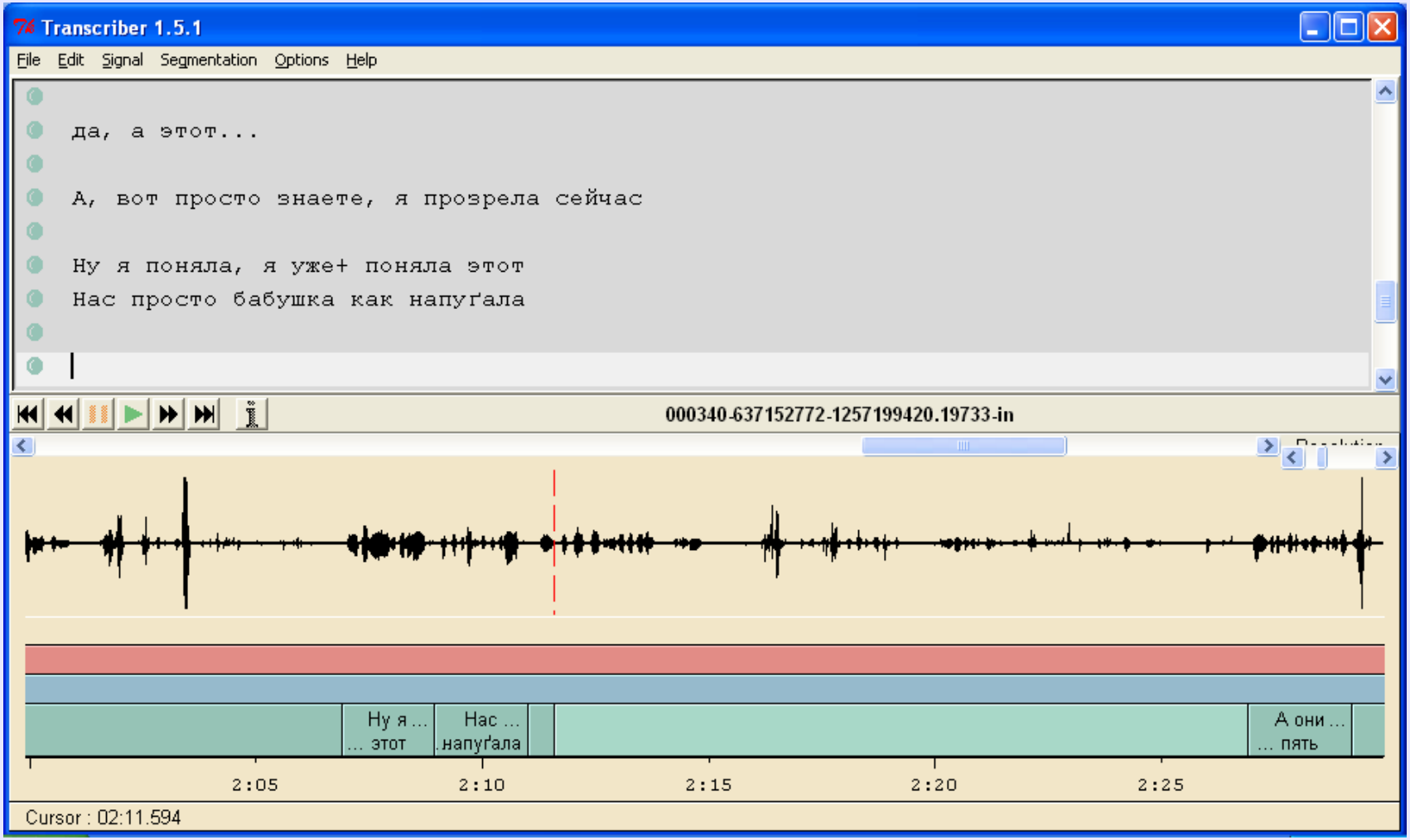
Породжувальна граматика вільного порядку слідування фонем

Автоматичний фонетичний стенограф

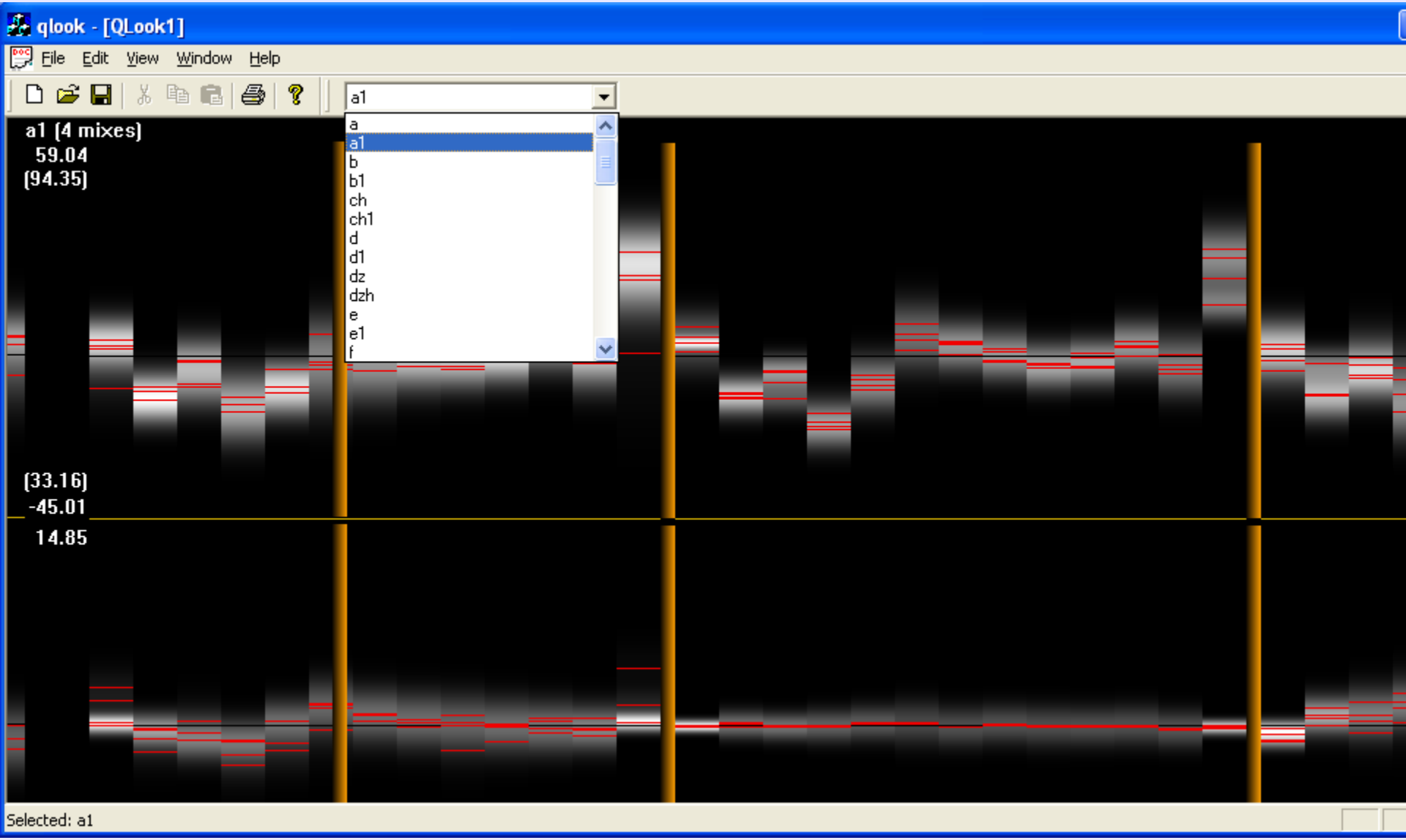
Розгорнутий граф



Оцінка параметрів АМ на основі мовленнєвого корпусу



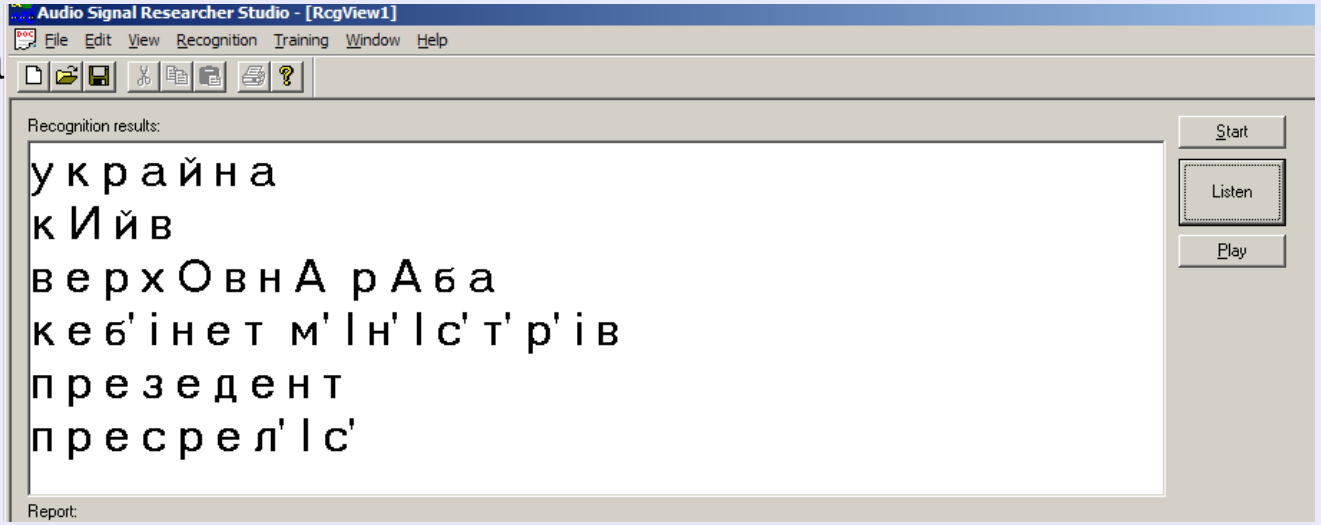
Оцінка параметрів АМ на основі мовленнєвого корпусу



Вихідні дані фонемного рівня

Українська мова

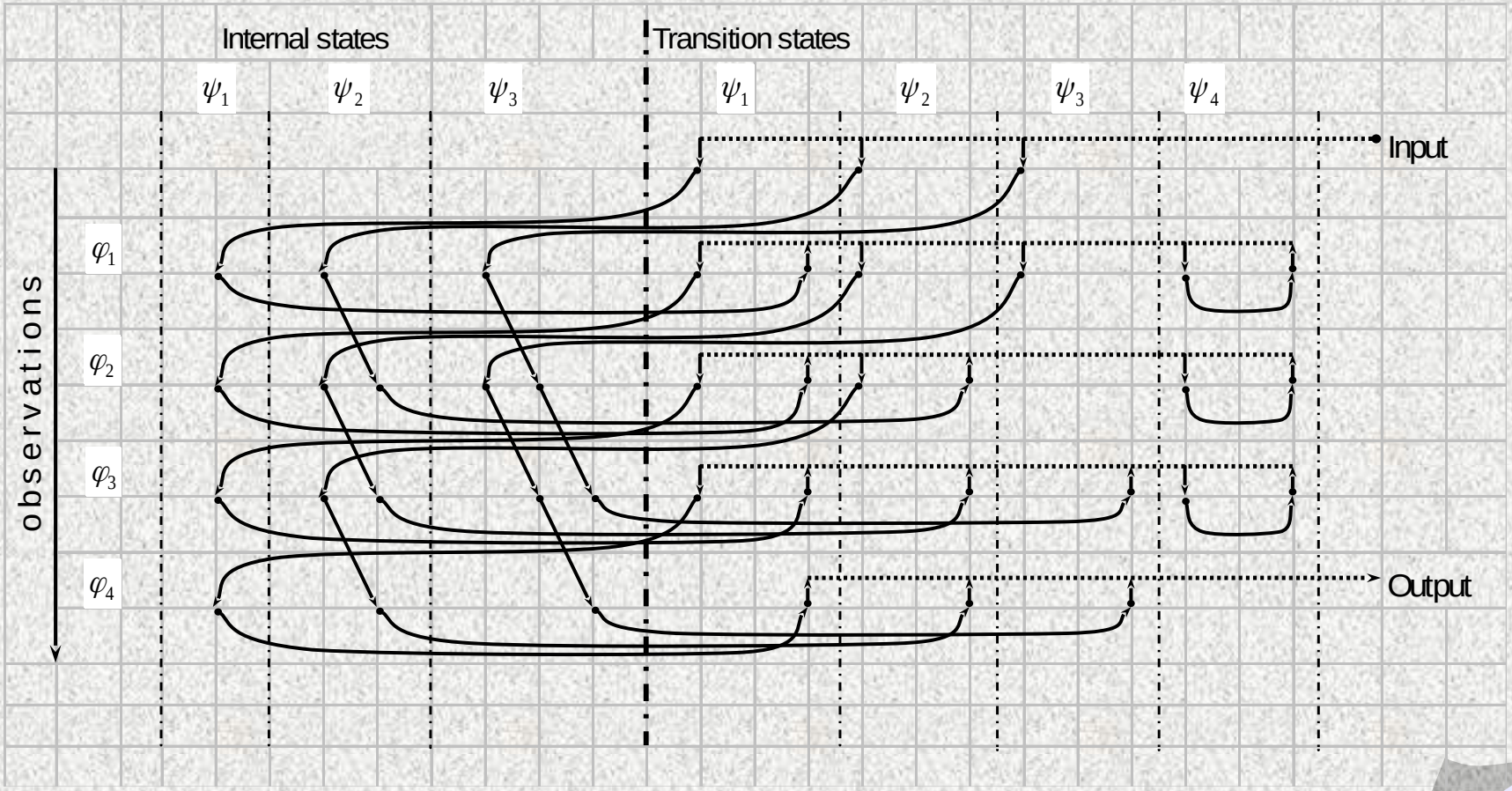
55 фонем



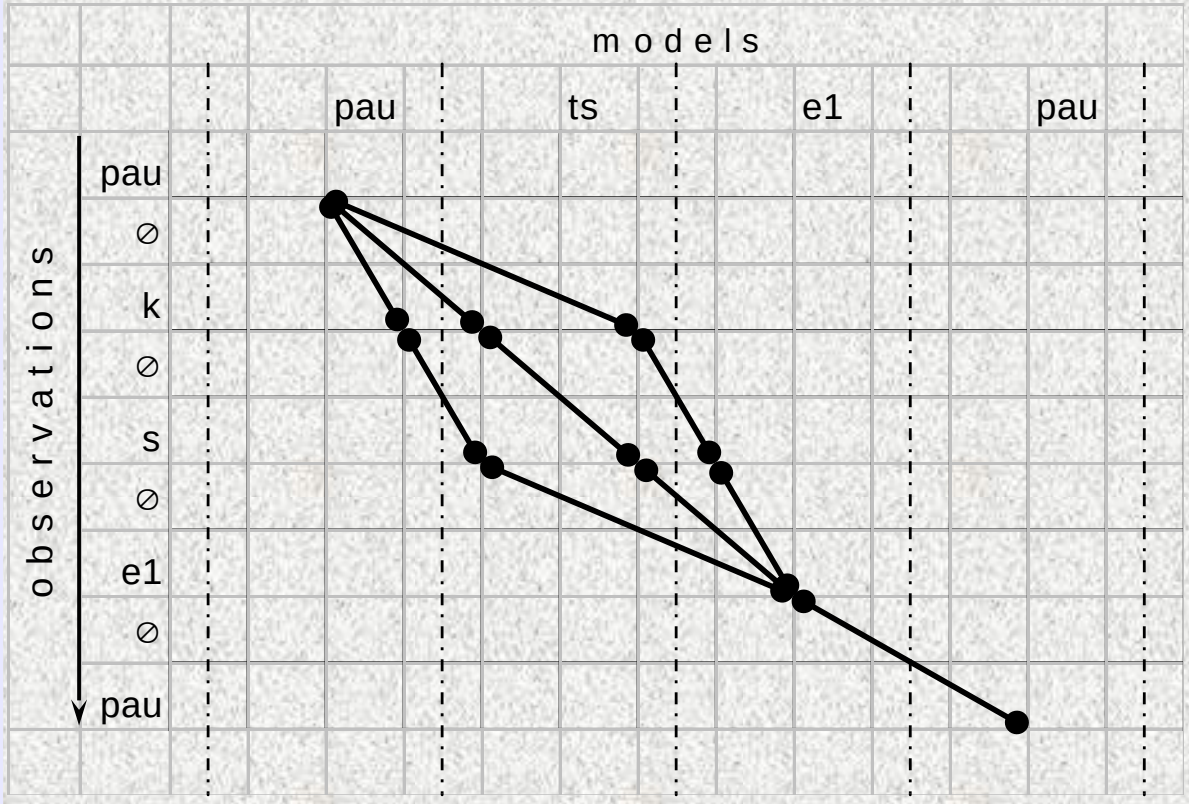
Вихідні дані фонемного рівня

from	to	n_score	item
[0	82]	-21.481287	pau
[83	97]	-27.774561	sh
[98	113]	-32.287659	ch
[114	118]	-30.129492	o1
[119	126]	-28.091370	r
[127	141]	-30.110937	o1
[142	151]	-33.818970	b
[152	157]	-32.969891	y1
[158	176]	-30.190096	t1
[177	181]	-31.983593	ts1
[182	193]	-30.517944	ch
[194	197]	-33.677856	l1
[198	224]	-27.012894	a1
[225	246]	-26.453724	s
[247	258]	-28.731241	f
[259	425]	-21.601635	pau

Each Phoneme Recognizer output sequence is processed with the phonetic-acoustic filter by means of dynamic programming. The *N* best phoneme sequences of the first level are converted to *N*2 word sequences.



Ініціалізація моделей згідно з мінімальною правкою



Phonetic-acoustic model prototypes with the following phonemic descriptions:

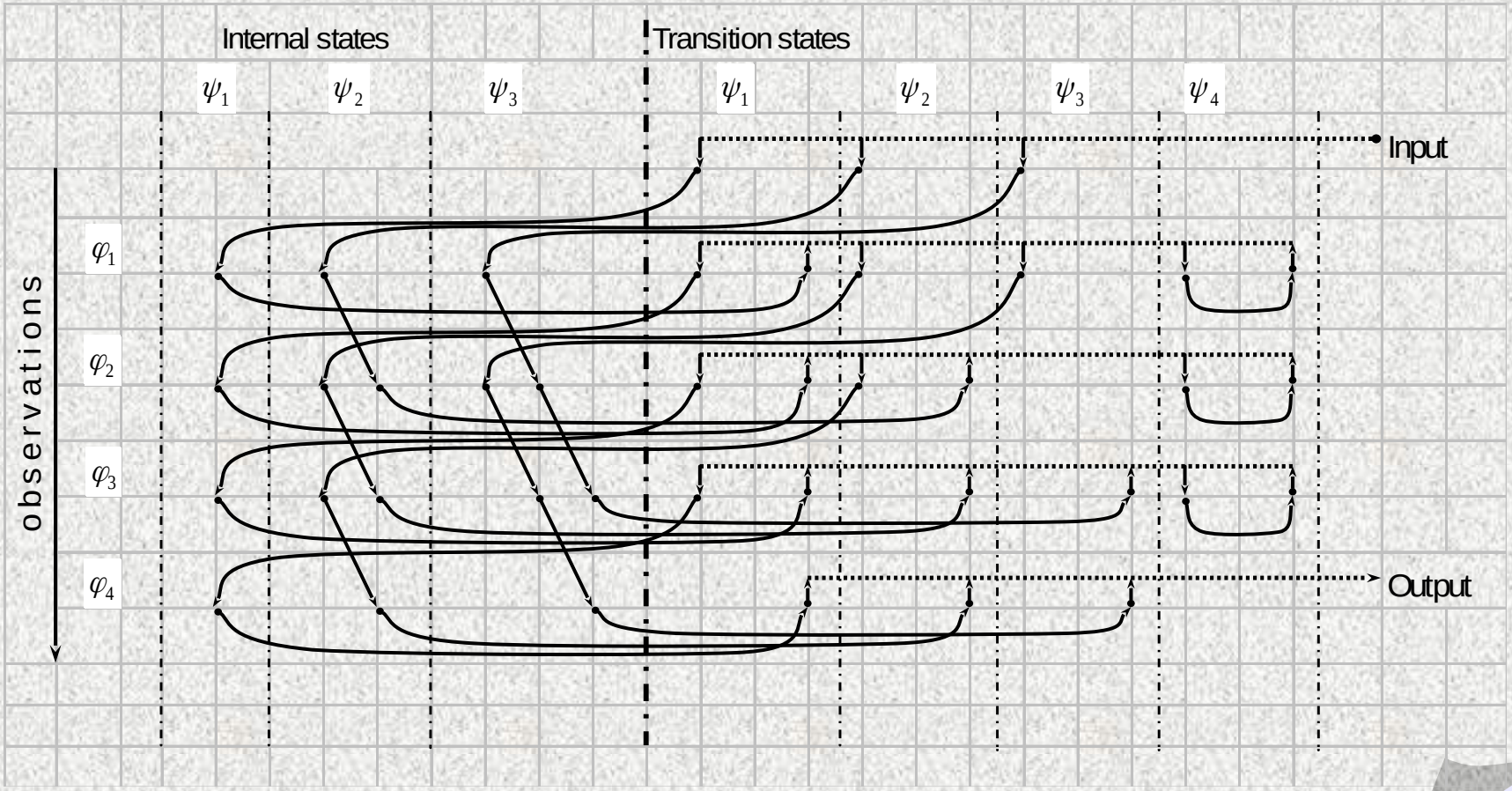
```
1:(PAU, k / pau, 4),
2:(PAU, k / pau, 5),
3:(pau, PAU / pau, 6);
4:(k / 1, ts, 7),
5:(k, s / 2, ts, 8),
6:(s / 3, ts, 9);
7:(s, E1 / 3, e1, 9),
8:(E1 / 4|5, e1, 9);
9:(PAU / 6|7|8, pau).
```

Вихідні дані фонемного рівня

Ініціалізація моделей згідно з мінімальною правкою

[0 50]	-22.154732	pau	pau	pau
[51 66]	-30.191032	m	m	m
[67 75]	-34.106201	y1	y1	y1
[76 83]	-30.016556	n	n	n (.5)
[84 86]	-33.170490	y1		n+y1 (.5)
[87 90]	-32.715576	e	e	y1-e, e
[91 98]	-28.277893	h	h	h
[99 104]	-31.310791	l	l	l, l+u
[105 111]	-35.321602	l	u	u, *
[112 132]	-27.987200	kh1	kh1	kh1, u-kh1
[133 143]	-29.788618	i1	i1	i1 (.5)
[144 161]	-26.959866	i		i1+i (.5)
[162 176]	-24.318165	t1	sp	sp, i-sp

Each Phoneme Recognizer output sequence is processed with the phonetic-acoustic filter by means of dynamic programming. The *N* best phoneme sequences of the first level are converted to *N*2 word sequences.



One of the Phoneme Recognizer level result is $N>>1$ best observed phoneme sequences $\Phi^r_{0Q^r}=\left(\varphi^r_1,\varphi^r_2,...,\varphi^r_u,...,\varphi^r_{Q^r}\right)$, $r=1{:}N$ where Q^r is a length of the r-th observed sequence. Moreover, as the result of the first level, each phoneme observation φ^r_u might be accomplished with information about its duration d^r_u, probability ΔF^r_u and may be other parameters like energy, pitch movement etc.

At the second level Word Recognizer must extract for all $\Phi^r_{0Q^r}$, $r=1{:}N$ total $N1>>1$ hidden phoneme sequences $\Psi^{r1}_{0Q^{r1}}=\left(\psi^{r1}_1,\psi^{r1}_2,...,\psi^{r1}_s,...,\psi^{r1}_{Q^{r1}}\right)$, $r1=1{:}N1$, $\psi\in\Psi\equiv\Phi$ and associate them with word sequences $J^{r2}_{0Q^{r2}}=\left(j^{r2}_1,j^{r2}_2,...,j^{r2}_k,...,j^{r2}_{Q^{r2}}\right)$, $r2=1{:}N2$, $N2>>1$ and $j^{r2}_k\in J$ where J is a word dictionary. To avoid loosing the actual word sequence $N2>>1$ recognition responses are taken.

Thus, we interpret observed phoneme subsequences $\Phi^r_{u_{s-1}u_s}=\left(\varphi^r_{u_{s-1}+1},\varphi^r_{u_{s-1}+2},...,\varphi^r_{u_s}\right)$, $u_{s-1}\le u_s$, as a transformed hidden s-th phoneme ψ^{r1}_{ks} from the k-th word regular transcription $j_{0q_k}=\left(\psi^{r1}_{k1},\psi^{r1}_{k2},...,\psi^{r1}_{ks},...,\psi^{r1}_{kq_k}\right)$. The probability of that that an observed subsequence $\Phi^r_{s_{k-1}s_k}=\left(\varphi^r_{s_{k-1}+1},\varphi^r_{s_{k-1}+2},...,\varphi^r_{s_k}\right)$, where $(s_k-s_{k-1})=l$ is length of the observation, is a realization of the hidden k-th word transcription $j_{0q_k}=\left(\psi^{r1}_{k1},\psi^{r1}_{k2},...,\psi^{r1}_{ks},...,\psi^{r1}_{kq_k}\right)$ assigns to a product of independent distortions maximized by hidden phoneme ψ^{r1}_{ks} bounds $\{u_s\}$:

$$P\left(\Phi^r_{s_{k-1}sk}\,/\,j_{0q_k}\right)=\max_{\{u_s\}}\prod_{s=1}^{q_k}P\left(\Phi^r_{u_{s-1}u_s}\,/\,\psi^{r1}_{ks}\right). \qquad (12)$$

$$P\left(\Phi^{r}_{s_{k-1}sk} \middle/ j_{0q_k}\right) = \max_{\{u_s\}} \prod_{s=1}^{q_k} P\left(\Phi^{r}_{u_{s-1}u_s} \middle/ \psi^{r1}_{ks}\right). \qquad (12)$$

Each factor $P(\Phi_{\mu\nu}/\psi)$ is equal to 0 if $\Phi_{\mu\nu} = (\varphi_{\mu+1}, \varphi_{\mu+2}, \ldots, \varphi_{\nu})$ is not associated with the hidden ψ, otherwise it is computed as a function of both a $\Phi_{\mu\nu}$ to ψ mapping occurrence frequency and acoustic parameter normal laws.

Each Phoneme Recognizer output sequence is processed with the described phonetic-acoustic filter by means of dynamic programming. Therefore, the *N*>>1 best phoneme sequences of the first level are converted to *N*2>>1 word sequences. The phonetic-acoustic filter parameters are estimated by training samples.

Породження гіпотетичних послідовностей фонем

```
[   0   94]   -22.115217  pau
[  95  101]   -31.519426  p
[ 102  109]   -34.916626  o1
[ 110  115]   -29.365845  v
[ 116  120]   -30.317236  a
[ 121  128]   -33.157440  r
[ 129  138]   -32.764305  n
[ 139  146]   -26.986633  e1
[ 147  156]   -30.091162  t
[ 157  161]   -29.844091  y1
[ 162  174]   -29.700758  s1
[ 175  191]   -29.325712  p
[ 192  210]   -30.579050  o1
[ 211  218]   -34.433228  t1
[ 219  230]   -29.935343  k1
[ 231  237]   -30.297153  i
[ 238  260]   -26.166079  m
[ 261  407]   -21.985432  pau
```

Породження гіпотетичних послідовностей фонем

```
[   0   94]  -22.115217  pau
[  95  101]  -31.519426  p
[ 102  109]  -34.916626  o1
[ 110  115]  -29.365845  v
[ 116  120]  -30.317236  a
[ 121  128]  -33.157440  r
[ 129  138]  -32.764305  n
[ 139  146]  -26.986633  e1
[ 147  156]  -30.091162  t
[ 157  161]  -29.844091  y1
[ 162  174]  -29.700758  s1
[ 175  191]  -29.325712  p
[ 192  210]  -30.579050  o1
[ 211  218]  -34.433228  t1
[ 219  230]  -29.935343  k1
[ 231  237]  -30.297153  i
[ 238  260]  -26.166079  m
[ 261  407]  -21.985432  pau
```

Порождение гипотетических последовательностей фонем

[95 101]	p							
[102 109]	o1	o	e	y	a			
[110 115]	v							
[116 120]	a	e	a1					
[121 128]	r							
[129 138]	n							
[139 146]	e1	e						
[147 156]	t							
[157 161]	y1	e						
[162 174]	s1	s						
[175 191]	p							
[192 210]	o1	o						
[211 218]	t1	*						
[219 230]	k1	*						
[231 237]	i	i1						
[238 260]	m	n	b	j				

Правила перехода от произношения к написанию

```
; jotation: "й а" = "я","'я","й а","йя";
й [аеіуАЕІУ] 2й ?   2[яєїюЯЄЇЮ]  2+ ' + [яєїюЯЄЇЮ]
й [аеуАЕУ]   2+ й + [яєюЯЄЮ]          ; майя
; palatalization "н' а" = "ня","нь а"
? ' [аеіуАЕІУ]      2? + ь | ?   2? + [яєіюЯЄІЮ]
; и might be і at word beginning after non-palatalized
[бвгґджзйклмнпрстфхцчш] [иИ]     2?/[іІ]      1?
[jz] [иИ]    2д [жз] | [іІ]      1д [жз]
; phone combinations
; вмиває*шся*, ми*тцю* (-ям, і[в])
ц ' ц ' [аАуУіІ]    5+ т + ь + с + [яЯюЮіІ]
с ' с ' [аА] 5+ ш + с + [яЯ]
с [лнм]      1с + т 1с
ш ч    2щ     1ш/
ш ч ' [аеіуАЕІУ]    3+щ+[яєіюЯЄІЮ]      1ш/
ш' ч' [іІ]   3+щ+[іІ]
ш н    2 + ч + н +  1ш
...
```

Правила перехода от произношения к написанию

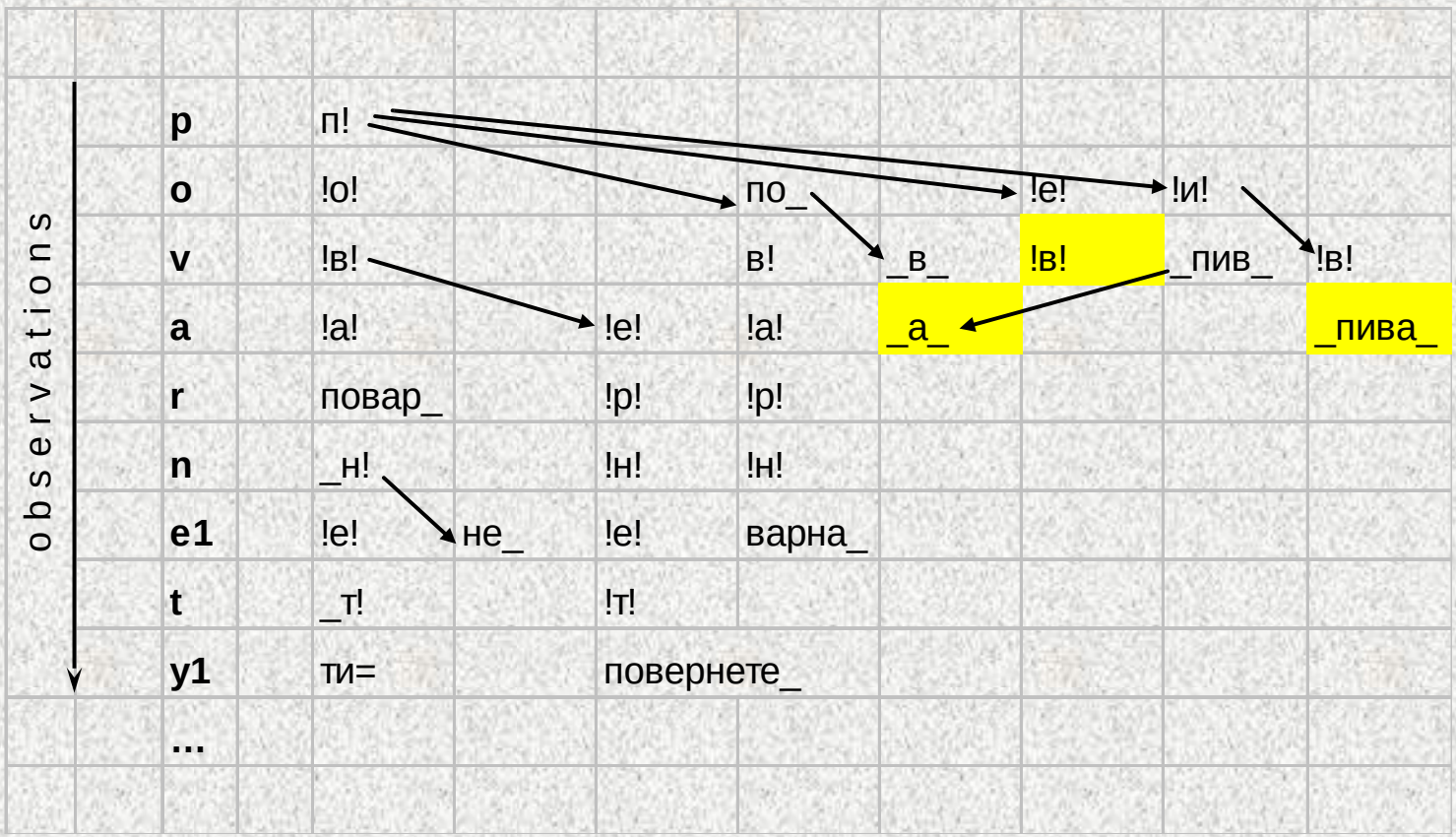
```
; vocalisation assimilation ф у д б О л
[бдгґзжjz] [бдгґзжjz]      1[птхксшчц]  1?
; softness assimilation
 [дтзснц] ' [дтзснц] '     2? + ь 2?
; sibilant assimilation
;;[жшч] [жшч]1[зсц] 1?
[шч] [шч]     1[ст]  1[жj]  1?
[жj] [жj]     1[ст]  1[шч]  1?
[сц] [сц]     1[ст]  1[жj]  1?
[зz] [зz]     1[ст]  1[шч]  1?
```

…

Пример: *повернете спокій*

		[95 101]	p						
		[102 109]	o1	o	e	y	a		
		[110 115]	v						
		[116 120]	a	e	a1				
		[121 128]	r						
		[129 138]	n						
		[139 146]	e1	e					
		[147 156]	t						
		[157 161]	y1	e					
		[162 174]	s1	s					
		[175 191]	p						
		[192 210]	o1	o					
		[211 218]	t1	*					
		[219 230]	k1	*					
		[231 237]	i	i1					
		[238 260]	m	n	b	j			

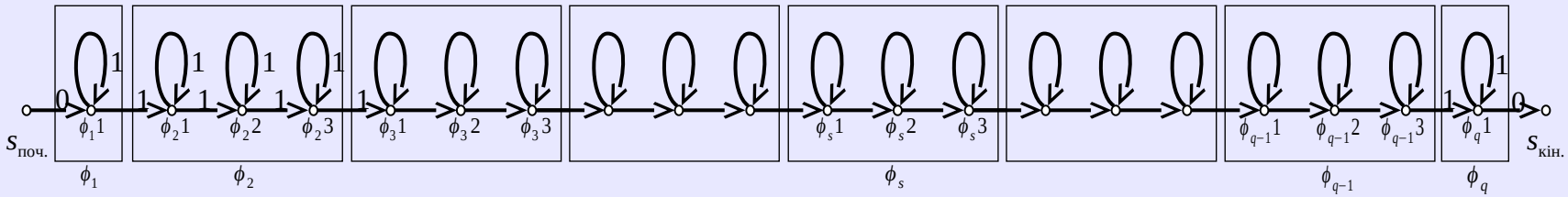
Грамматика перехода от произношения к написанию



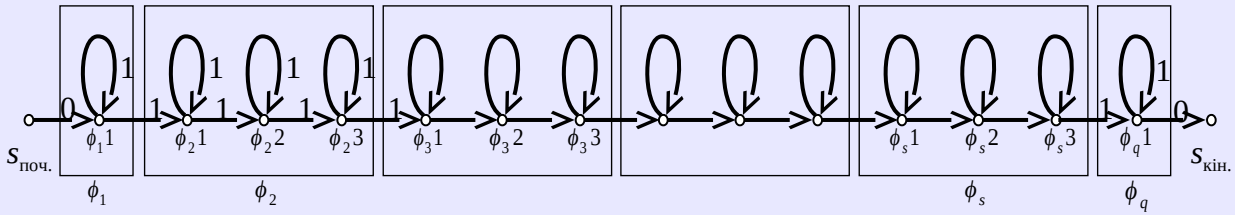
Формування еталонних сигналів слів із фонем відповідно до моделі вимови

$$k\left(X_{0l}\right)=\underset{k}{\operatorname{argmax}}\max_{s=1:q_k} ¿\left(\mu_{ks},w_{k1s},w_{k2s}\right), ¿¿\sum_{s=1}^{q_k}\left[\left(\right)+\left(\right)+\left(\right)\right]¿$$

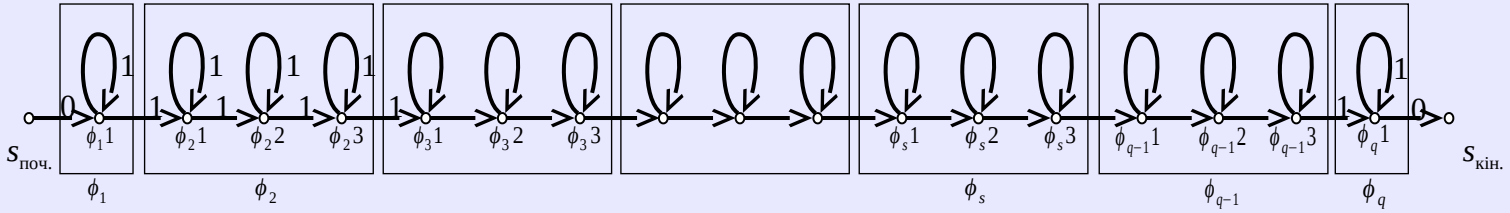
$k = 1$



$k = 2$

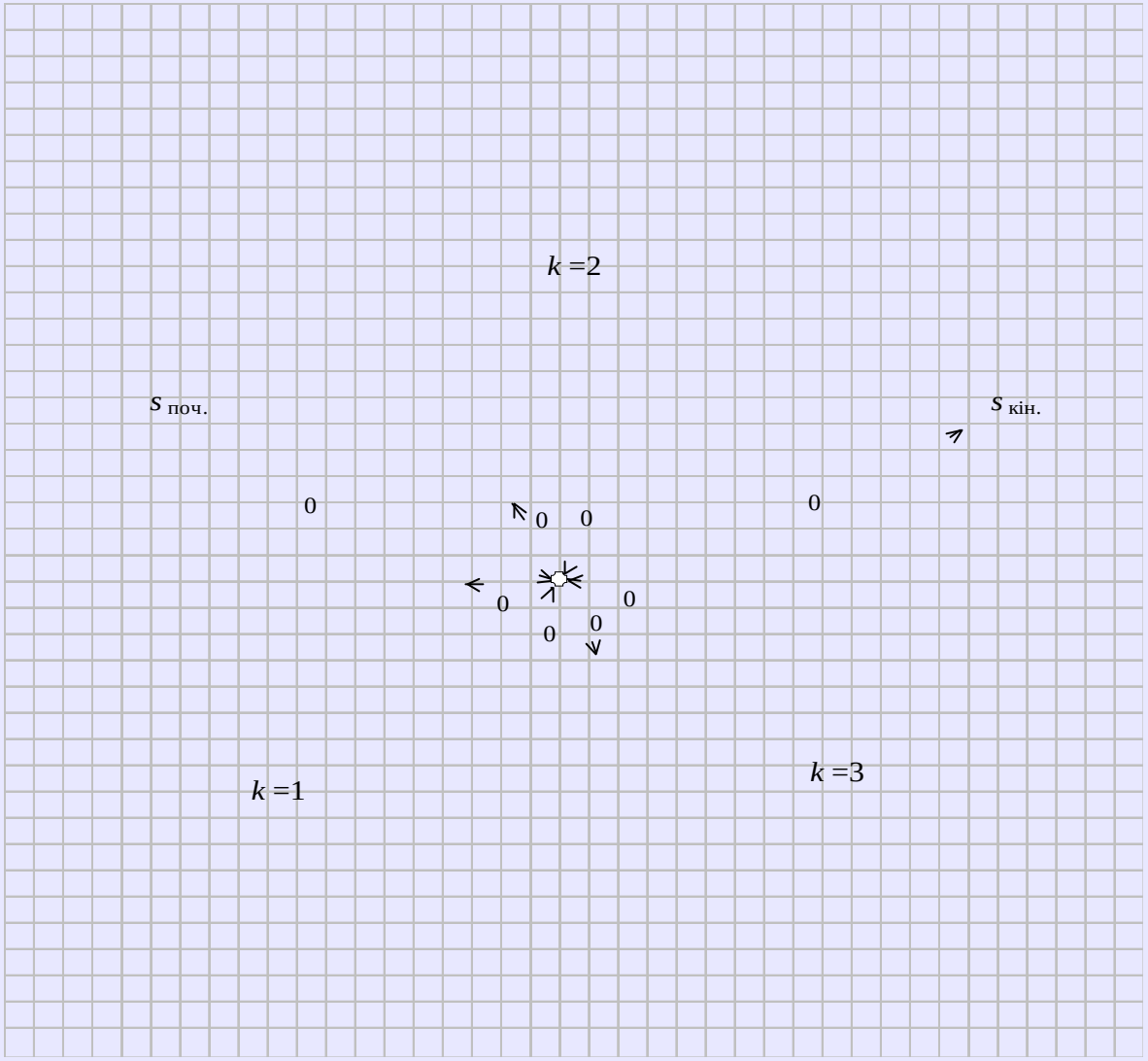


$k = 3$



N-ка, $N > 1$

Пофонемне розпізнавання злитої мови, що складається зі слів обраного словника, у випадку вільного порядку слідування слів



Лінгвістична модель

$k \in K$ – словарь

$k \in V_{\text{поч.}} = \{1, 2, 5\}$,

$k \in V_{k=1.} = \{1, 4, 5\}$,

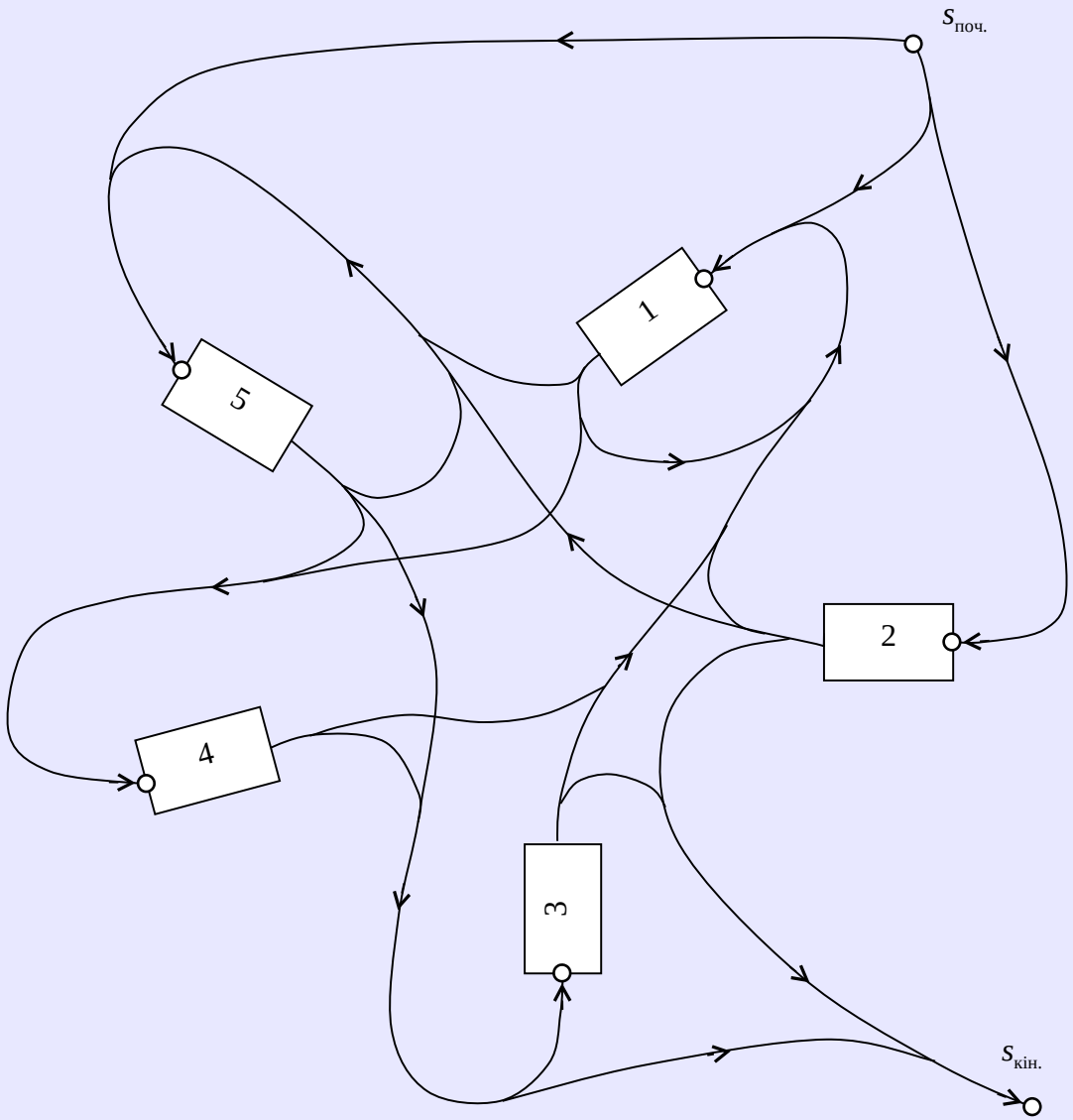
$k \in V_{k=2.} = \{1, 5\}$,

$k \in V_{k=3.} = \{1\}$,

$k \in V_{k=4.} = \{3, 1\}$,

$k \in V_{k=5.} = \{3, 4, 5\}$,

$k \in V_{\text{кін..}} = \{2, 3, 4, 5\}$,



Граматики,

n-grams

Лексико-синтаксический уровень

Основна ідея лінгвістичної моделі

Оцінювання імовірності речення із l слів $W_{0,l} = (w_1, w_2, \ldots w_l)$ за умови спостережень мовленнєвого сигналу $O = (o_1, o_2, \ldots o_m)$:

$$P(W_{0,l}/O) = \frac{P(O/W_{0,l})P(W_{0,l})}{P(O)},$$

Основна ідея лінгвістичної моделі

Імовірність речення із l слів $W_{0,l} = (w_1, \ w_2, \ \ldots \ w_l)$ розкладається на добуток умовних імовірностей:

$$P(w_1, w_2, ..., w_l) = \prod_{i=1}^{l} P(w_i / w_1, w_2, ..., w_{i-1}) = \prod_{i=1}^{l} P(w_i / W_{0,i-1}),$$

Основна ідея лінгвістичної моделі

Обмежуємо контекст:

$$P(w_1, w_2, ..., w_l) \cong \prod_{i=1}^{l} P(w_i / w_{i-n+1}, ..., w_{i-1}) = \prod_{i=1}^{l} P(w_i / W_{i-n,i-1}),$$

де n – порядок моделі.

При $n = 1$ модель вироджується до відсутності елементарних контекстів, тобто розглядаються **уні-грами**.

При $n = 2$ отримуємо **бі-грамну** модель, коли в елементарний контекст входить граматична функція одного попереднього слова.

Збільшуючи далі параметр n, ми розширюємо елементарний контекст, підвищуючи точність моделі.

Основна ідея лінгвістичної моделі

Враховуючи:

$$\hat{P}(W_{0,l}) = \prod_{i=1}^{l} P(w_i / W_{i-n,i-1}),$$

Оцінюємо ймовірність послідовності слів при надходженні наступного слова w_{i+1}

$$\hat{P}(W_{0,i+1}) = \hat{P}(W_{0,i}) P(w_{i+1} / W_{i-n,i})$$

Основна ідея лінгвістичної моделі

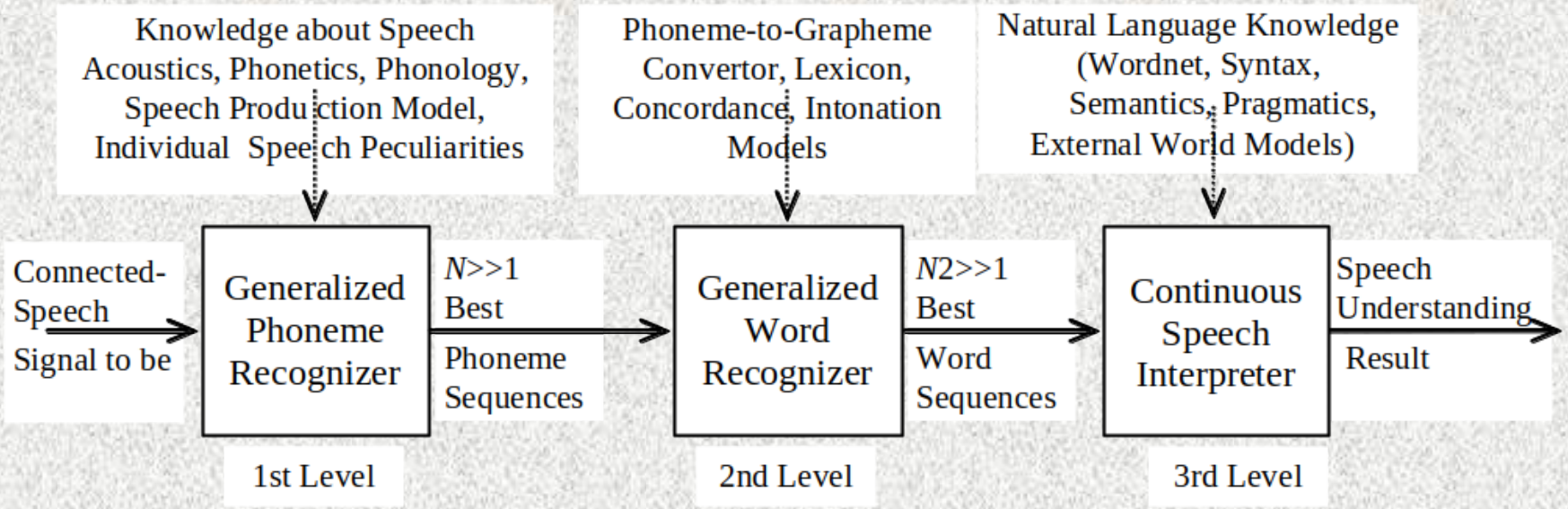
Оцінка параметрів лінгвістичної моделі

$$P(w_i / W_{i-n,i-1}) = \begin{cases} \alpha(W_{i-n,i-1}) & : \quad c(W_{i-n,i}) = 0 \\ d_{c(W_{i-n,i})} \dfrac{c(W_{i-n,i})}{c(W_{i-n,i-1})} & : \quad 1 \le c(W_{i-n,i}) \le k \\ \dfrac{c(W_{i-n,i})}{c(W_{i-n,i-1})} & : \quad c(W_{i-n,i}) > k \end{cases}$$

Основна ідея лінгвістичної моделі

Розбиття слів на класи $g = G(w)$

Three-Level ASR&U System Structure



Рівні дають змогу ефективно розподіляти роботу між дослідниками. Рівні обробляються в єдиному процесі, але форма постпроцесора полегшує реалізацію

Архитектура процессора;
Распараллеливание;
Платформа (библиотеки wxWidgets, PortAudio)

Инструментарий:
HTK;
Julius;
Sphinx, ISIP, STK