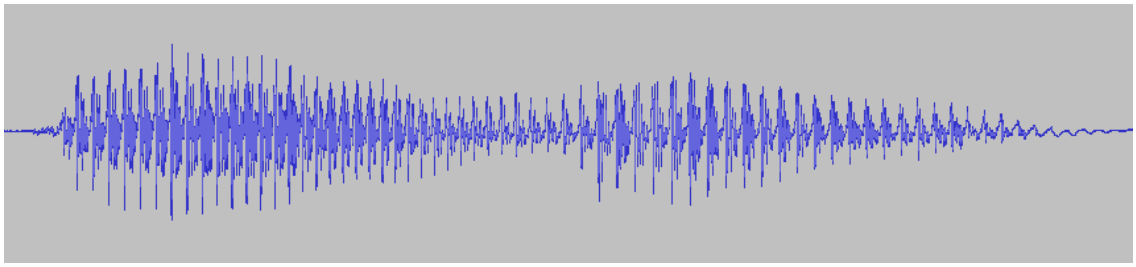
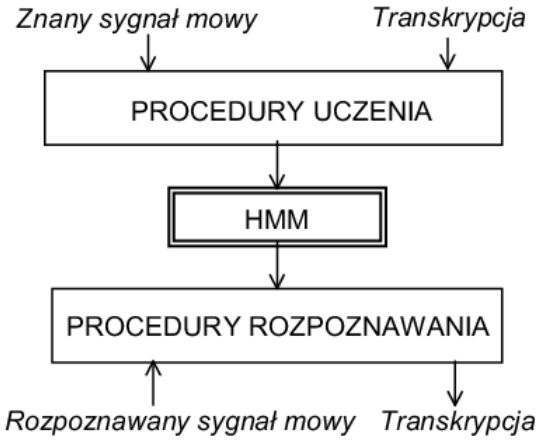
Wykład 4. Ukryte modele Markowa (HMM – Hidden Markov Model)

Współczesne systemy rozpoznawania mowy opierają się na założeniu, że mowa jest sekwencją pewnych elementarnych dyskretnych jednostek (tzw. jednostek fonetycznych). W zależności od przeznaczenia systemu jako jednostki fonetyczne można przyjąć: słowa, sylaby, fonemy (ewentualnie trifony lub diafony[1]). Proces generowania słowa „Ala” oraz jego identyfikacji przedstawia poniższy rysunek:



wektor cech:	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}	w_{13}	w_{14}	w_{15}
jednostki:				a				l					a		

Dysponując ciągiem dyskretnych obrazów akustycznych (wektorów cech) nie jesteśmy w stanie przyporządkować mu jednoznacznej sekwencji jednostek mowy. Rozwiązaniem tego problemu może być zastosowanie ukrytych modelu Markowa (HMM). Proces rozpoznawana mowy za pomocą HMM przedstawia poniższy schemat:



[1] Diafon (difon, tranzem) – para fonemów, przejście między fonemami

Realizację systemu automatycznego rozpoznawania mowy można podzielić na dwie fazy:

- uczenie (trening), które polega na estymacji parametrów zbioru modeli HMM za pomocą wypowiedzi uczących,
- rozpoznanie, które polega na wyznaczeniu transkrypcji fonetycznej (ewentualnie gramatycznej) nieznanych, rozpoznawanych wypowiedzi.

Ukryte modele Markowa są statystyczną metodą klasyfikacji sekwencji zdarzeń. Do celów rozpoznawania mowy po raz pierwszy zostały użyte przez Bakera, Jelinka i Levinsona (firma IBM) w połowie lat siedemdziesiątych. W metodzie tej sygnał mowy traktowany jest jako ciąg kolejnych wektorów parametrów określanych w krótkich odcinkach czasu (np. ramkach trwających 10 – 30 ms). Zakłada się, że sygnał mowy w czasie odpowiadającym długości ramki jest stacjonarny. Rozpoznawany fragment mowy np. słowo można wówczas przedstawić jako ciąg T obserwacji:

$$O = O_1, O_2, \dots , O_T$$

Dla każdej klasy rozpoznawanych sygnałów mowy należy utworzyć oddzielny model. Rozpoznanie polega wówczas na dopasowaniu analizowanego ciągu do zapamiętanych ciągów wzorcowych. Każdy model (ciąg wzorcowy) jest automatem skończonym posiadającym N stanów:

$$Q = \{q_1, q_2, \dots ,q_N\}$$

Automat skończony reprezentujący HMM definiuje się jako $\lambda = <\pi, A, B>$, gdzie:

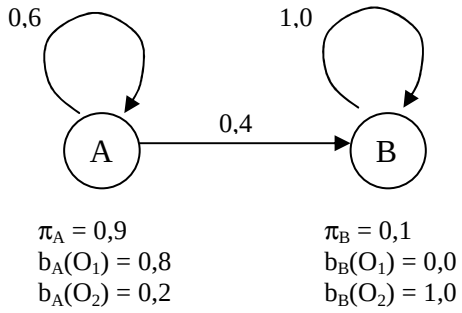
$\pi = [\pi_1, \pi_2, \dots , \pi_N]$ opisuje rozkład prawdopodobieństw znalezienia się w stanie q_i w chwili $t = 0$,

$A = [a_{ij}]$ $(i,j =1.. N)$ jest macierzą opisującą prawdopodobieństwa przejść między stanami,

$B = [b_i(o_j)]$ $(i = 1.. N, j = 1.. M$; M – liczba możliwych zdarzeń generowanych przez dany stan) jest macierzą opisującą prawdopodobieństwa pojawienia się j-tej obserwacji w stanie q_i:

$$b_i(o_j) = P(o_j| q_i)$$

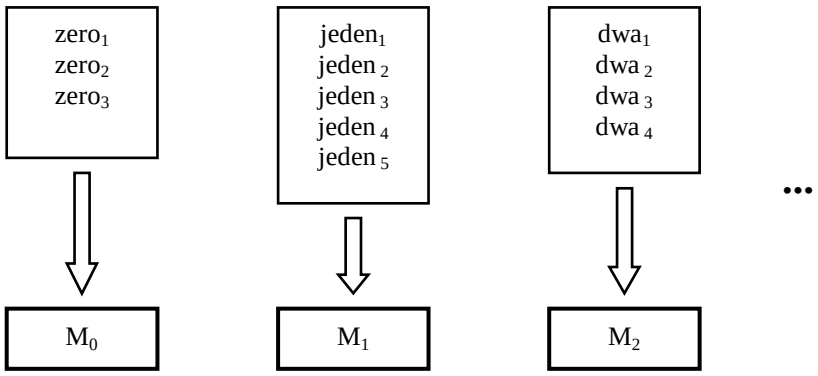
Poniższy rysunek przedstawia przykładowy automat o dwóch stanach reprezentujący model pewnego słowa:



Tworzenie systemu automatycznego rozpoznawania mowy obejmuje następujące etapy:

- przygotowanie danych (tworzenie plików transkrypcyjnych, kodowanie danych, budowa słownika, zdefiniowanie gramatyki),
- tworzenie ukrytych modeli Markowa (zdefiniowanie początkowych modeli, estymacja ostatecznych modeli),
- opracowanie procedury rozpoznawania (np. rozpoznawanie jednostek izolowanych),
- weryfikacja i testowanie systemu (rozpoznawanie danych testowych np. wykorzystanych w procesie uczenia, praca w czasie rzeczywistym).

Postać zbioru uczącego, gdy modele tworzone są na poziomie słów przedstawia rysunek:

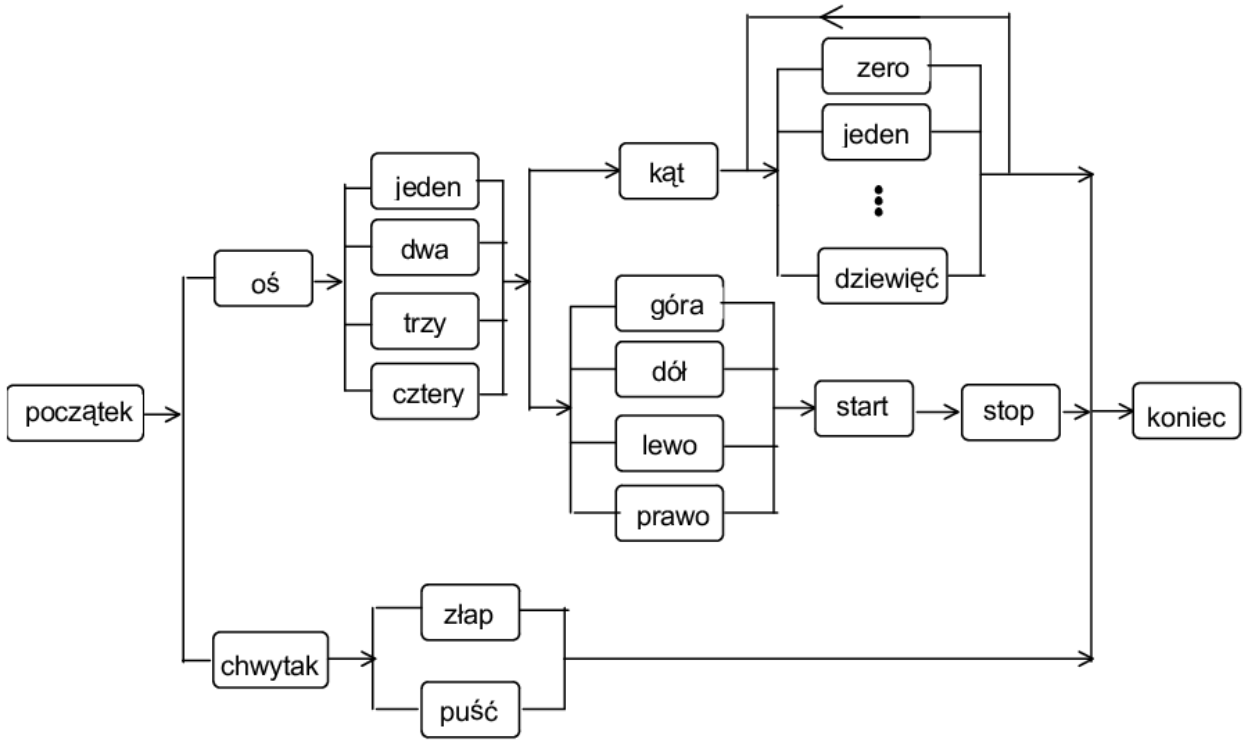


Poszczególne wypowiedzi należące do zbioru uczącego zapisywane są w oddzielnych plikach (np. rozszerzenie .wav). Dla każdego pliku dźwiękowego należy podać postać tekstową wypowiedzi (pliki .txt). Dodatkowo, dla wszystkich zarejestrowanych wypowiedzi należy podać ich transkrypcje fonetyczne (ewentualnie gramatyczne). Transkrypcję fonetyczną tworzą ciągi fonemów:

TRZY	tSy
CZTERY	Ctery
PIĘĆ	pjeńć
SZEŚĆ	Seść
SIEDEM	śedem
DZIEWIĘĆ	Źewjeńć
KĄT	kont
DÓŁ	duł
CHWYTAK	hwytak

Dla nielicznego zbioru uczącego transkrypcje fonetyczne można wyznaczyć ręcznie. Najczęściej jednak (szczególnie w przypadku dużych zbiorów) transkrypcje fonetyczne uzyskuje się automatycznie: na podstawie plików tekstowych oraz słownika wymowy danego języka.

Transkrypcje gramatyczne są wykorzystywane, gdy wypowiedź jest złożona z kilku wyrazów (są opisem występujących w wypowiedzi wyrazów). Przykład gramatyki zdefiniowanej dla systemu sterowania robotem za pomocą głosu przedstawia rysunek:



Sygnał mowy powinien zostać poddany również segmentacji (indeksacji). Proces ten polega na określeniu momentów rozpoczęcia i zakończenia się poszczególnych jednostek fonetycznych tworzących daną wypowiedź (zgodnie z transkrypcją fonetyczną). Segmentacja może być przeprowadzona ręcznie lub automatycznie.

Przed rozpoczęciem procesu tworzenia modeli HMM należy dokonać ekstrakcji charakterystyk (parametrów) z sygnałów mowy (np. opis sygnału mowy za pomocą współczynników LPC, MCCC).

Ukryte modele Markowa HMM tworzone są dla wszystkich wyróżnionych jednostek fonetycznych (słów, fonemów, trifonów itd.). Wymaga to wykonania dla każdego modelu następujących operacji:

- określenia złożoności modelu (liczby stanów modelu),
- estymacji wartości parametrów modelu w procesie uczenia.

Proces uczenia polega na wyznaczeniu elementów macierzy A, wektora prawdopodobieństw początkowych π oraz wektora rozkładów wyjściowych $B = [B_i(O_t)]$. Jest to proces złożony i przebiega w

kilku etapach (wyznaczenie początkowych wartości parametrów, reestymacja parametrów). Do tworzenia modeli wykorzystywane są następujące algorytmy:

- algorytm Viterbiego,
- algorytm Baum-Welcha,
- algorytm „embedded training”.

Problemy z utworzeniem odpowiednich modeli HMM wynikają z następujących czynników:

- segmentacja sygnału mowy (podział na stałe odcinki czasowe nie jest najlepszym rozwiązaniem, stąd stosuje się algorytmy automatycznej segmentacji, których celem jest maksymalizacja prawdopodobieństwa wygenerowania danej obserwacji dla danego modelu HMM),
- konieczność kwantyzacji wektorowej obserwacji (oszacowanie prawdopodobieństw występowania dowolnej obserwacji w dowolnym stanie jest praktycznie nie realizowalne).

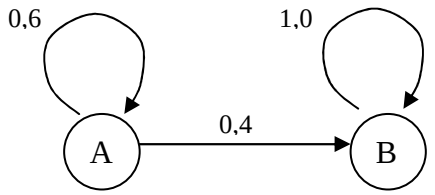
Rozpoznanie polega na stwierdzeniu, który model do danego ciągu wejściowego. Jest to ten model, który posiada największe prawdopodobieństwo wygenerowania danego ciągu zdarzeń.
Obliczenie prawdopodobieństwa jest procedurą iteracyjną:

$$\alpha_1(i) = \pi_1 \, b_i(O_1)$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(O_{t+1})$$

<u>Przykład</u>

Rozpatrzmy dwustanowy model słowa o stanach $q_1 = A$ i $q_2 = B$ o parametrach jak na rysunku:



$\pi_A = 0{,}9$ $\quad$ $\pi_B = 0{,}1$
$b_A(O_1) = 0{,}8$ $\quad$ $b_B(O_1) = 0{,}0$
$b_A(O_2) = 0{,}2$ $\quad$ $b_B(O_2) = 1{,}0$

Dany jest ciąg obserwacji $O = \{O_2, O_2, O_2\}$. Obliczyć prawdopodobieństwo wystąpienia takiej obserwacji w powyższym modelu.

Metoda 1:

Możliwe 3-elementowe sekwencje stanów to: AAA, AAB, ABB, BBB. Prawdopodobieństwa ich wystąpienia wynoszą odpowiednio:

$P(AAA) = \pi_A * \alpha_{AA} * \alpha_{AA} = 0{,}9 * 0{,}6 * 0{,}6 = 0{,}324$

$P(AAB) = \pi_A * \alpha_{AA} * \alpha_{AB} = 0{,}9 * 0{,}6 * 0{,}4 = 0{,}216$

$P(ABB) = \pi_A * \alpha_{AB} * \alpha_{BB} = 0{,}9 * 0{,}4 * 1{,}0 = 0{,}360$

$P(BBB) = \pi_B * \alpha_{BB} * \alpha_{BB} = 0{,}1 * 1{,}0 * 1{,}0 = 0{,}100$

Dla każdej z powyższych sekwencji obliczamy prawdopodobieństwo wystąpienia obserwacji O:

$P(O/AAA) = b_A(O_2) * b_A(O_2) * b_A(O_2) = 0{,}2 * 0{,}2 * 0{,}2 = 0{,}008$

$P(O/AAA) = b_A(O_2) * b_A(O_2) * b_B(O_2) = 0{,}2 * 0{,}2 * 1{,}0 = 0{,}04$

$P(O/AAA) = b_A(O_2) * b_B(O_2) * b_B(O_2) = 0{,}2 * 1{,}0 * 1{,}0 = 0{,}2$

$P(O/AAA) = b_B(O_2) * b_B(O_2) * b_B(O_2) = 1{,}0 * 1{,}0 * 1{,}0 = 1{,}0$

Następnie obliczamy całkowite prawdopodobieństwa wystąpienia poszczególnych sekwencji:

$P(O/AAA)*P(AAA) = 0{,}008 * 0{,}324 = 0{,}002592$

$P(O/AAB)*P(AAB) = 0{,}04 * 0{,}216 = 0{,}00864$

$P(O/ABB)*P(ABB) = 0{,}2 * 0{,}36 = 0{,}072$

$P(0/BBB)*P(BBB) = 1{,}0 * 0{,}1 = 0{,}1$

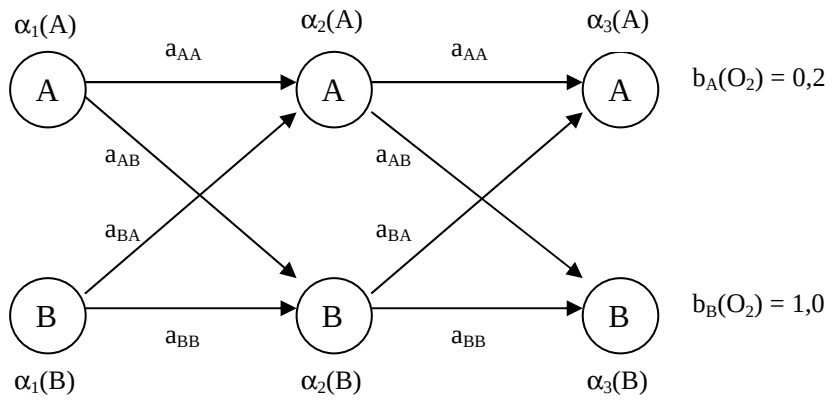
Całkowite prawdopodobieństwo wystąpienia obserwacji O w modelu M równe jest sumie obliczonych powyżej prawdopodobieństw:

$P(O/M) = 0{,}002592 + 0{,}00864 + 0{,}072 + 0{,}1 = 0{,}183232$

Wadą powyższej metody jest duża liczba obliczeń, którą należy wykonać. Liczba sekwencji, które należy rozważyć wynosi N^T (dla N = 10 i T = 30 otrzymujemy 10^{30} kombinacji).

Metoda 2:

Przedstawmy model Markowa w ten sposób, że nad każdą obserwacją z ciągu O znajduje się zbiór wszystkich stanów q_i:



Czas	t_1	t_2	t_3
Obserwacja	O_2	O_2	O_2

Oznaczmy przez $\alpha_i(n)$ prawdopodobieństwo wygenerowania ciągu obserwacji od początku do chwili t zakładając, że obserwacja O_t w chwili t pochodzi od stanu n. Wówczas:

$\alpha_1(A) = \pi_A * b_A(O_2) = 0,9 * 0,2 = 0,18$

$\alpha_1(B) = \pi_B * b_B(O_2) = 0,1 * 1,0 = 0,10$

$\alpha_2(A) = (\alpha_1(A) * a_{AA} + \alpha_1(B) * a_{BA}) * b_A(O_2) = (0,18 * 0,6 + 0,1* 0,0) * 0,2 = 0,0216$

$\alpha_2(B) = (\alpha_1(A) * a_{AB} + \alpha_1(B) * a_{BB}) * b_B(O_2) = (0,18 * 0,4 + 0,1* 1,0) * 1,0 = 0,172$

$\alpha_3(A) = (\alpha_2(A) * a_{AA} + \alpha_2(B) * a_{BA}) * b_A(O_2) = (0,0216 * 0,6 + 0,172* 0,0) * 0,2 = 0,002592$

$\alpha_3(B) = (\alpha_2(A) * a_{AB} + \alpha_2(B) * a_{BB}) * b_B(O_2) = (0,0216 * 0,4 + 0,172* 1,0) * 1,0 = 0,18064$

Ostatecznie otrzymujemy:

$P(O/M) = \alpha_3(A) + \alpha_3(B) = 0,183232$

Druga metoda wymaga zdecydowanie mniejszej liczby obliczeń. Aby obliczyć wszystkie współczynniki $\alpha_i(n)$ należy wykonać jedynie N*T iteracji (dla N=10 i T=300 daje to 300 obliczeń).

Polecana strona:

http://htk.eng.cam.ac.uk/ - the Hidden Markov Model Toolkit (wersja 3.4)