AKADEMIA GÓRNICZO - HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki

PRACA DOKTORSKA

Temat pracy:

Rozpoznawanie mowy metodami niejawnych modeli Markowa HMM

PROMOTOR:
prof. zw. dr hab. inż. Bogusław Filipowicz

OPRACOWAŁ:
mgr inż. Karol Szostek

KRAKÓW 2006

Składam serdeczne podziękowania promotorowi
Panu Profesorowi Bogusławowi Filipowiczowi
za pomoc i cenne uwagi, które przyczyniły
się do powstania tej pracy

Spis treści

1. WPROWADZENIE 7

2. CELE I TEZY PRACY 14

3. WPROWADZENIE DO TEORII NIEJAWNYCH MODELI MARKOWA HMM 15

3.1. MODEL PROCESU RZUCANIA MONETĄ 16
3.2. MODEL PROCESU MARKOWA 18
3.3. NIEJAWNE MODELE MARKOWA HMM 21
3.3.1. Elementy modeli HMM 21
3.3.2. Trzy podstawowe problemy dotyczące HMM (Wyniki analizy niejawnych modeli Markowa) 22
3.3.3. Rozwiązanie problemu 1 (Wyznaczenie prawdopodobieństwa obserwacji) 23
3.3.4. Rozwiązanie problemu 2 (Wyznaczenie optymalnej sekwencji stanów) 25
3.3.5. Rozwiązanie problemu 3 (Estymacja parametrów modelu) 27
3.3.5.1. Algorytm Bauma-Welcha estymacji parametrów dla modelu podstawowego DHMM 28
3.3.5.2. Algorytm Bayesa estymacji parametrów modelu podstawowego DHMM 30
3.3.5.3. Algorytm segmental *k*-means estymacji parametrów modelu podstawowego DHMM 32
3.3.5.4. Algorytmu segmental MAP estymacji parametrów modelu podstawowego DHMM 32
3.4. ZASTOSOWANIE JAWNEJ GĘSTOŚCI ROZKŁADU PRAWDOPODOBIEŃSTWA POZOSTAWANIA W TYM SAMYM STANIE 33
3.5. MODEL HMM Z CIĄGŁA GĘSTOŚCIĄ ROZKŁADU PRAWDOPODOBIEŃSTWA OBSERWACJI W STANACH CDHMM 36
3.5.1. Algorytm Bauma-Welcha estymacji parametrów modelu HMM z rozkładami ciągłymi 38
3.5.2. Wyznaczenie rozkładu „prior" dla rozkładu normalnego z diagonalną macierzą kowariancji 40
3.5.3. Algorytm Bayesa estymacji parametrów modelu HMM z rozkładami ciągłymi 44
3.5.4. Algorytm segmental k-means estymacji parametrów modelu HMM z rozkładami ciągłymi 47
3.5.5. Algorytm segmental MAP estymacji parametrów modelu HMM z rozkładami ciągłymi 48
3.6. MODEL NIESTACJONARNY HMM „*TRENDED HMM*" 50
3.6.1. Algorytm Bauma-Welcha estymacji parametrów modelu niestacjonarnego 50
3.6.2. Algorytm segmental k-means estymacji parametrów modelu niestacjonarnego 52
3.6.3. Wyznaczenie rozkładu „prior" dla niestacjonarnego rozkładu normalnego z diagonalną macierzą kowariancji 54
3.7. SKALOWANIE 58
3.8. WIELOKROTNE SEKWENCJE OBSERWACJI 60
3.9. METODA ADAPTACJI MLLR (*MAXIMUM LIKELIHOOD LINEAR REGRESSION*) 61
3.9.1. Modyfikacja wartości średniej 61
3.9.2. Modyfikacja macierzy kowariancji 64
3.10. REDUKCJA WEKTORA CECH METODAMI LDA, HLDA 66
3.10.1. Metoda LDA 66
3.10.2. Metoda HLDA 67

4. MODEL SIECIOWY 70

4.1. OBLICZENIE PRAWDOPODOBIEŃSTWA OBSERWACJI 71
4.2. ESTYMACJA PARAMETRÓW MODELU SIECIOWEGO 72
4.3. SPROWADZENIE ESTYMATORA MODELU SIECIOWEGO DO ESTYMATORA MODEL Z JAWNYM ROZKŁADEM PRAWDOPODOBIEŃSTWA POZOSTAWANIA W TYM SAMYM STANIE 75
4.4. SPROWADZENIE ESTYMATORA MODELU SIECIOWEGO DO ESTYMATORA MODELU PODSTAWOWEGO 77
4.5. SKALOWANIE W MODELU SIECIOWYM 78
4.6. UCZENIE MODELU SIECIOWEGO DO WIELOKROTNYCH SEKWENCJI OBSERWACJI 80

5. SZACOWANIE ZŁOŻONOŚCI OBLICZENIOWEJ WYZNACZANIA PRAWDOPODOBIEŃSTWA OBSERWACJI ORAZ ESTYMACJI PARAMETRÓW MODELU HMM 82

5.1. SZACOWANIE ZŁOŻONOŚCI DLA MODELU PODSTAWOWEGO HMM 82
5.2. SZACOWANIE ZŁOŻONOŚCI DLA MODELU Z JAWNYM ROZKŁADEM PRAWDOPODOBIEŃSTWA POZOSTAWANIA W TYM SAMYM STANIE 83
5.3. SZACOWANIE ZŁOŻONOŚCI DLA MODELU SIECIOWEGO 83

6. ALGORYTMY SZTUCZNEJ INTELIGENCJI 86

6.1. ALGORYTM GENETYCZNY 86
6.1.1. Tworzenie nowej populacji 86
6.1.2. Algorytm konstrukcyjny 86
6.1.3. Kodowanie chromosomów 87
6.1.4. Funkcja oceny 87
6.1.5. Reprodukcja z krzyżowaniem 88
6.1.6. Mutacja 88
6.1.7. Selekcja 89
6.2. ALGORYTM ITERACYJNY 90
6.2.1. Otoczenie rozwiązania 90

7. ZALEŻNOŚCI WYKORZYSTYWANE DO WYZNACZENIA ALGORYTMU BAUMA-WELCHA 92

8. TWORZENIE SEKWENCJI UCZĄCYCH 93

8.1. KRÓTKO-CZASOWA TRANSFORMACJA FOURIERA (STFT) 94
8.2. METODA PREDYKCJI LINIOWEJ (LPC) 96
8.3. METODA (MFCC) 100
8.4. POMIAR ODLEGŁOŚCI POMIĘDZY WIDMAMI SYGNAŁU 102
8.5. POCHODNE CEPSTRUM 106
8.6. METODA MODYFIKACJI ALGORYTMU PRZETWARZANIA SYGNAŁU 107
8.7. KWANTYZACJA WEKTOROWA (VQ) 108

9. OPIS SYSTEMU ROZPOZNAWANIA 110

10. WYNIKI OBLICZEŃ KOMPUTEROWYCH 113

10.1. OPTYMALIZACJA MODELI DO SEKWENCJI, DLA KTÓRYCH ISTNIEJE OPTYMALNY MODEL HMM 113
10.1.1. Optymalizacja modelu podstawowego HMM 113
10.1.2. Optymalizacja modelu z jawnym rozkładem prawdopodobieństwa pozostawania w tym samym stanie 115
10.1.3. Optymalizacja modelu sieciowego 116
10.2. EFEKTY PRZEUCZENIA MODELI HMM DLA NIELICZNYCH DANYCH TRENINGOWYCH 118
10.3. PORÓWNANIE WYNIKÓW ROZPOZNAWANIA SYLAB W POSTACI PRZEBIEGÓW CZASOWYCH ORAZ PO PRZETWARZANIU METODĄ LPC ORAZ MFCC. 120
10.4. PORÓWNANIE WYNIKÓW ROZPOZNAWANIA CYFR DLA MODELI HMM 123
10.4.1. Model podstawowy HMM 123
10.4.2. Model HMM z jawnym rozkładem prawdopodobieństwa pozostawania w tym samym stanie 124
10.4.3. Model sieciowy HMM 125
10.5. ZASTOSOWANIE ALGORYTMÓW SZTUCZNEJ INTELIGENCJI DO OPTYMALIZACJI MODELI HMM 126

11. PODSUMOWANIE I WNIOSKI 128

11.1. OTRZYMANE WYNIKI 128
11.2. PROPOZYCJE DALSZYCH BADAŃ 129

12. LITERATURA 131

Wykaz oznaczeń stosowanych w pracy

$\mathbf{O}=(o_1,o_2,\ldots,o_T)$	sekwencja obserwacji dla modelu DHMM
$\mathbf{O}=(\mathbf{o}_1,\mathbf{o}_2,\ldots,\mathbf{o}_T)$	sekwencja obserwacji dla modelu CDHMM
o	składowa sekwencji dla modelu DHMM wartość całkowita od 1 do n
$\boldsymbol{o}$	składowa sekwencji dla modelu CDHMM wektor $\boldsymbol{o} \in \Re^D$
$\{\boldsymbol{x}_i\}$, $\{\boldsymbol{y}_i\}$	zbiory wektorów obserwacji $\boldsymbol{x}_i, \boldsymbol{y}_i \in \Re^D$
$\Re$	zbiór liczb rzeczywistych
$c_1,\ldots,c_L$	składowe wektora obserwacji współczynniki (*cepstrum*)
$\Delta c_1,\ldots,\Delta c_L$	składowe wektora obserwacji pochodne współczynników (*cepstrum*)
$\Delta^2 c_1,\ldots,\Delta^2 c_L$	składowe wektora obserwacji drugie pochodne współczynników (*cepstrum*)
h_1, h_2, h_3	współczynniki wielomianu do wyznaczenia pochodnych (*cepstrum*)
L	liczba wyznaczanych współczynników (*cepstrum*)
D	wymiar składowej sekwencji obserwacji dla modeli CDHMM (rozdział 3.5); maksymalny czas pozostawania w stanie, węźle dla modeli DHMM (rozdział 3.4 oraz 4); ilość składowych widma w transformacji Fouriera (rozdział 8)
T	długość sekwencji obserwacji
t	indeks składowej sekwencji obserwacji, dyskretna chwila pracy modelu HMM
$S=(S_1,S_2,..,S_T)$	sekwencja stanów (*ścieżka*)
$S^*=(S_1^*,S_2^*,..,S_T^*)$	optymalna sekwencja stanów
f_{ij}	liczba przejść ze stanu i do stanu j na optymalnej ścieżce
g_{ki}	liczba symboli k w stanie i na optymalnej ścieżce
M	model probabilistyczny HMM; liczba składowych (*cepstrum*) przy wyznaczaniu pochodnych (rozdział 8.5)
M^z	model probabilistyczny HMM w węzłach modelu sieciowego
$P(\mathbf{O}\|M)=P(\mathbf{O})$	wartość prawdopodobieństwa wygenerowania obserwacji **O** przez model M
$P(\mathbf{O},S\|M)$	wartość prawdopodobieństwo wygenerowania obserwacji **O** przez model M na ścieżce S
$p_i(d)$	wartość funkcji rozkładu prawdopodobieństwa pozostawania w stanie, węźle modelu HMM dla stanu (wezła) i przez okres d
d_k	czas pozostawania w stanie, węźle o indeksie k, dla modelu z jawnym rozkładem prawdopodobieństwa pozostawania w tym samym stanie oraz dla modelu sieciowego
τ_i	liczba chwil czasowych liczona od czasu przejścia do stanu i dla modelu niestacjonarnego
d_t	parametr czasu wystąpienia obserwacji w stanie i dla modelu niestacjonarnego $d_t=t-\tau_i$
H	stopień wielomianu opisującego składową wektora wartości średniej dla rozkładu niestacjonarnego
$g(M)$	funkcja „*prior*" rozkładu prawdopodobieństwa parametrów modelu HMM
$\eta_i, \eta_{ij}, \nu_{ki}$	parametry rozkładu „*prior*" dla dyskretnych rozkładów prawdopodobieństw
$\boldsymbol{m}$, τ, $\boldsymbol{\beta}$, (α)	parametry rozkładu „*prior*" dla rozkładu normalnego
$\boldsymbol{M}$, τ, $\boldsymbol{\phi}$, $\boldsymbol{\beta}$,	parametry rozkładu „*prior*" dla rozkładu normalnego niestacjonarnego
m	liczba stanów w modelu HMM
Z	liczba węzłów w modelu sieciowym HMM
δ_i^z, γ_{ij}^z, π_{ki}^z,	parametry pod modeli DHMM w modelu sieciowym
n	liczba generowanych symboli przez model DHMM
N	liczba komponentów mikstury w stanie modelu CDHMM (rozdział 3.5), liczba próbek sygnału w oknie (rozdział 8)
$\delta=\{\delta_i\}$	wektor rozkładu prawdopodobieństwa początkowego
$\Gamma=\{\gamma_{ij}\}$	macierz rozkładów prawdopodobieństw przejść
$\pi=\{\pi_{ki}\}$	macierz rozkładów symboli występujących w modelu DHMM
$\pi_{\mathbf{o}j}$	wartość prawdopodobieństwa wygenerowania wektora obserwacji **o** w stanie j modelu CDHMM

$\pi_{\mathbf{o}jk}=\mathcal{N}(\mathbf{o},\boldsymbol{\mu}_{jk},\boldsymbol{U}_{jk})$	wartość prawdopodobieństwa wygenerowania wektora obserwacji **o** przez rozkład normalny $\mathcal{N}$o indeksie k w stanie j modelu CDHMM
$\pi_{\mathbf{o}_t ik}=\mathcal{N}(\mathbf{o},\boldsymbol{B}_{jk}\boldsymbol{t},\boldsymbol{U}_{jk})$	wartość prawdopodobieństwa wygenerowania obserwacji **o** przez rozkład niestacjonarny $\mathcal{N}$o indeksie k w stanie j, czas wystąpienia obserwacji wynosi t
$\pi_{\mathbf{o}_t jkd_t}=\mathcal{N}(\mathbf{o}_t,\boldsymbol{B}_{jk}\boldsymbol{T},\boldsymbol{U}_{jk})$	wartość prawdopodobieństwa wygenerowania obserwacji **o** przez rozkład niestacjonarny $\mathcal{N}$o indeksie k w stanie j, czas wystąpienia obserwacji wynosi d_t
$\boldsymbol{\mu}_{jk},\boldsymbol{U}_{jk}$	parametry funkcji rozkładu normalnego $\mathcal{N}$o indeksie k w stanie j
$\boldsymbol{B}_{ik},\boldsymbol{U}_{ik}$	parametry funkcji rozkładu normalnego niestacjonarnego $\mathcal{N}$o indeksie k w stanie j
$C_{ik}(m)$, $\boldsymbol{V}_i(d)$	oznaczenia wprowadzone dla modelu niestacjonarnego
c_{jk}	składowa o indeksie k dyskretnego rozkładu prawdopodobieństwa dla komponentów mikstury w stanie j modelu CDHMM
c_t	współczynnik skalowania tablic $\alpha_t(i)$, $\beta_t(i)$ dla czasu t
$l=(l_1,l_2,\ldots,l_T)$	sekwencja podstanów w modelu CDHMM
$l^*=(l_1^*,l_2^*,\ldots,l_T^*)$	optymalna sekwencja podstanów w modelu CDHMM
$\boldsymbol{F}_t(i,k)$	wektor złożony z wielomianów rzędu D
$\boldsymbol{R}(i,k)$	wektor losowany według D-wymiarowego rozkładu normalnego o wartości średniej równej zeru oraz macierzy kowariancji $\boldsymbol{U}$
$\boldsymbol{W}$, $\boldsymbol{H}$	macierz modyfikacji wartości średniej oraz macierzy kowariancji w metodzie adaptacji MLLR
$\boldsymbol{W}$	macierz skalowania w metodzie przetwarzania sygnału LPC
J	liczba klas w metodach LDA oraz HLDA redukcji wektorów cech
N, N_j	liczba wszystkich wektorów oraz wektorów w klasie j w metodach redukcji wektorów cech LDA, HLDA (rozdział 3.10)
$\boldsymbol{\theta}$	macierz przekształcenia liniowego w metodzie redukcji wektora cech LDA, HLDA
θ	funkcja skalująca w metodzie przetwarzania sygnału LPC
$\overline{\boldsymbol{x}}_j$, $\overline{\boldsymbol{W}}_j$	wartości średnie oraz macierze kowariancji dla poszczególnych klas w metodzie LDA, HLDA
$\overline{\boldsymbol{x}}$, $\overline{\boldsymbol{T}}$	wartości średnie oraz macierze kowariancji dla wszystkich klas w metodzie LDA, HLDA
$Q(M\|\widehat{M})$	funkcja pomocnicza wykorzystywana do wyznaczania estymatorów Bauma-Welcha
$R(M\|\widehat{M})$	zmodyfikowana funkcja pomocnicza wykorzystywana do wyznaczania estymatorów Bayesa
$\alpha_t(i)$, $\beta_t(i)$	współczynniki w algorytmie *forward-backward*
$\alpha_t^*(i)$, $\beta_t^*(i)$	współczynniki w algorytmie *forward-backward* dla modelu z jawnym rozkładem prawdopodobieństwa pozostawania w tym samym stanie oraz modelu sieciowym
$\xi_t(i,j)$, $\chi_t(i)$	współczynniki w estymatorze Bauma-Welcha
$\eta_t(i)$, $\varphi_t(i)$	współczynniki w algorytmie Viterbiego
$\eta_t(i,k,d)$, $\varphi_t(i,k,d)$	współczynniki w algorytmie Viterbiego dla modelu niestacjonarnego
$\nu_t(i)$	współczynnik w algorytmach estymacji parametrów modelu CDHMM, różni się od iloczynu $\alpha_t(i)\beta_t(i)$ tym, że nie uwzględnia wygenerowania obserwacji $\pi_{\mathbf{o}i}$
$\gamma_t(i,k)$	współczynnik stosowany w algorytmie MLLR uwzględnia czynnik $c_{ik}\pi_{\mathbf{o}_t ik}$
$x(n)$	dyskretny sygnał dźwięku
f_p	częstotliwość próbkowania
$X(k)$	dyskretna transformacja Fouriera sygnału (*widmo sygnału*)
$\|X(k)\|$	moduł widma sygnału
$S(k)=\|X(k)\|^2$	kwadrat modułu widma sygnału
a_k	współczynniki szeregu Fouriera, współczynniki modelu predykcyjnego w metodzie LPC
p	liczba współczynników modelu predykcyjnego
$H(z)$	transmitancja modelu predykcyjnego

$e(n)$	błąd predykcji
$u(n)$	znormalizowany błąd predykcji
$G^2=\sigma^2$	współczynnik poziomu sygnału
E_n	sumaryczny błąd predykcji
E_{min}	minimalny błąd predykcji
$r(k)$	funkcja autokorelacji
$\phi_n(i,k)$	funkcja wprowadzona w metodzie przetwarzania sygnału LPC
$\mathbf{s}_p^T(m)$, $\mathbf{R}_p$, $\mathbf{a}_p$, $\mathbf{r}_p$	parametry wykorzystywane w metodzie LPC
$H_i(k)$	oznaczenie filtrów w metodzie MFCC
Y_i	współczynnik reprezentujące widmo sygnału w skali „*Mel*”
d_2^2	odległość (*cepstral distance*) dwóch widm sygnału wyznaczanych metodą LPC
d_c^2	odległość (*cepstral distance*) dwóch widm sygnału wyznaczanych metodą MFCC
$K(f)$	funkcja skalująca widmo w dziedzinie częstotliwości
L_{max}	wartość progowa wykorzystywana w algorytmach sztucznej inteligencji oraz w procesie generowania losowych modeli HMM
K	parametr normalizujący w kryterium oceny modeli

1. Wprowadzenie

W rozpoznawaniu mowy stosowane są metody bazujące na porównywaniu przybliżonego widma rozpoznawanego sygnału dźwięku z wzorcem.

Sygnał mowy charakteryzuje się pewną przypadkowością. Zarejestrowany przebieg wypowiedzi jednego mówcy przenoszący tą samą informację ma zawsze niepowtarzalny przebieg czasowy, dlatego podczas rozpoznawania mowy należy i uwzględnia się czynniki przypadkowe wpływające na sygnał mowy.

Można wyróżnić dwie kategorie metod wykorzystywanych do rozpoznawania tj. metody deterministyczne oraz niedeterministyczne. W obu metodach uwzględnia się czynniki przypadkowe wpływające na sygnał dźwięku. W metodach deterministycznych wyznaczane są błędy występujące podczas porównywana dźwięków z wzorcem, natomiast w metodach niedeterministycznych wyznacza się wartości prawdopodobieństw reprezentujące dopasowanie dźwięku do stosowanych modeli probabilistycznych.

Do metod deterministycznych można zaliczyć metodę dwupoziomowego programowania dynamicznego (*Double Level Dynamic Time Wraping*), możliwe jest również zastosowanie deterministycznych sieci neuronowych. Do metod niedeterministycznych należy algorytm Viterbiego związany z zastosowaniem niejawnych modeli Markowa (*Hidden Markov Model*), stosowane są również hybrydy wykorzystujące zarówno sieci neuronowe jak i niejawne modele Markowa.

Modele niedeterministyczne (probabilistyczne) są generalnie bardziej przydatne niż modele deterministyczne, do modelowania procesów, na które mają wpływ nieprzewidywalne i nieregularne czynniki. W praktyce dzięki zastosowaniu metod niedeterministycznych wykorzystujących niejawne modele Markowa uzyskiwane są najlepsze wyniki rozpoznawania.

Pierwsze publikacje [44]÷[59] dotyczące niejawnych modeli Markowa były opracowane przez Bauma oraz jego współpracowników w latach 60-tych oraz wczesnych latach 70-tych. Prace te dotyczyły metody estymacji parametrów modeli HMM nazywanej algorytmem Bauma-Welcha. Algorytm ten oparty jest na kryterium (ML) (*Maximum Likelihood*) maksymalizacji prawdopodobieństwa wygenerowania sekwencji obserwacji przez model HMM.

Poza algorytmem Bauma-Welcha został opracowany algorytm segmental *k*-means estymacji parametrów modeli HMM opisany w pracach [60]÷[63]. Algorytm ten jest oparty na zmodyfikowanym kryterium maksymalizacji (*segmental* ML). Modyfikacja kryterium polega na wprowadzeniu optymalnej ścieżki oraz maksymalizacji prawdopodobieństwa na optymalnej ścieżce. Stosowana w algorytmie ścieżka wyznaczana jest z wykorzystaniem algorytmu Viterbiego.

Równolegle z algorytmami Bauma-Welcha oraz algorytmem segmental *k*-means zostały opracowywane metody estymacji parametrów modeli HMM nazywane algorytmami adaptacji Bayesa opisane w pracach [64]÷[77]. Algorytm adaptacji Bayesa pozwala na modyfikację parametrów modelu HMM z uwzględnieniem wcześniej zgromadzonej wiedzy na jego temat. Algorytm adaptacji Bayesa będący odpowiednikiem algorytmu Bauma-Welcha oparty jest na kryterium maksymalizacji (MAP) (*Maksimum A Priori*), natomiast algorytm adaptacji Bayesa będący odpowiednikiem algorytmu segmental *k*-means oparty jest na kryterium (*segmental* MAP). Adaptacja Bayesa była początkowo stosowana jedynie do estymacji rozkładów prawdopodobieństw w stanach modelu HMM, dopiero w późniejszym czasie została zastosowana do wszystkich parametrów modelu np.: w pracy [74] z 1995 roku.

W systemach rozpoznawania mowy mogą zostać zastosowane zarówno modele stacjonarne jak niestacjonarne. W aplikacjach najczęściej są stosowane modele stacjonarne HMM, ze względu na to, że nie wymagają one tak dużej ilości danych treningowych oraz algorytmy estymacji tych modeli posiadają mniejszą złożoność obliczeniową niż algorytmy dla modeli niestacjonarnych.

Dla niestacjonarnych modeli HMM na początku został opracowany algorytm adaptacji oparty na kryterium (ML), który można znaleźć w pracy [59] z 1992, algorytm segmental *k*-means można znaleźć w pracy [63] z 1994 roku natomiast algorytmy adaptacji Bayesa oparte na kryterium (MAP) oraz (*segmental* MAP) zostały opublikowane w pracy [77] z 2001.

W wymienionych pracach zostały opisane metody optymalizacji modeli HMM nazywane metodami poprawiającymi (*corective training*). Poza metodami poprawiającymi można znaleźć równie liczne prace [90]÷[112] poświęcone algorytmom gradientowym.

W algorytmach poprawiających, do których zalicza się algorytmy: Bauma-Welcha, algorytm segmental *k*-means oraz algorytm Bayesa modyfikowane są jedynie modele zgodne z sekwencją treningową, dlatego naturalne było zaprojektowanie algorytmów pozwalających modyfikować zarówno modele zgodne z daną sekwencją treningową w celu poprawy prawdopodobieństwa obserwacji tej sekwencji oraz modele nie zgodne z sekwencją w celu zmniejszenia prawdopodobieństwa obserwacji sekwencji. Algorytmy, w których modyfikuje się jednocześnie kilka modeli nazywane są algorytmami dyskryminacyjnymi (*discriminative training*). W algorytmach dyskryminacyjnych, stosowane są metody gradientowe optymalizacji modeli HMM. W związku z zastosowaniem metod gradientowych oraz modyfikacją kilku modeli tj. modelu zgodnego z sekwencją treningową oraz modeli nie niezgodnych z sekwencją treningową algorytmy te działają dużo wolniej niż algorytmy poprawiające. W związku z zastosowaniem algorytmów dyskryminacyjnych zostały wprowadzone kryteria optymalizacji takie jak (MMI), (MDI), (MCE). W algorytmie opartym na kryterium (MMI) (*Maximum Mutual Information*) wykorzystuje się funkcje entropii parametrów modelu, informacje na temat algorytmów opartych na tym kryterium można znaleźć w pracach [92]÷[102]. Algorytmy oparte na kryterium (MCE) (*Minimum Classification Error*) inaczej nazywanym (MDI) (*Minimum Discrimination Information*) są opisane w pracach [103]÷[111].

W metodach rozpoznawania mowy bazujących na modelach HMM można wyróżnić dwa etapy:

- modelowania i optymalizacji
- rozpoznawania

Podczas modelowania i optymalizacji tworzone są modele HMM odpowiadające elementom języka tj.: *wyrazom, fonemom, sylabom, zdaniom*. Do tworzenia wzorców wykorzystuje się algorytmy estymacji parametrów modeli HMM. Metody estymacji parametrów, co już wcześniej zostało zaznaczone można podzielić na dwie grupy *tj*.: metody poprawiające (*corective training*) oraz metody dyskryminacyjne (*discriminative training*).

W praktycznych zastosowaniach stosuje się algorytmy poprawiające oparte na czterech kryteriach tj.: (ML), (*segmental* ML), (MAP) oraz (*segmental* MAP). Kryterium (ML) jest (*Maximum Likelihood*) wykorzystane w algorytmie Bauma-Welcha. Kryterium to polega na maksymalizacji prawdopodobieństwa obserwacji $\max_{M} P(\mathbf{O} \mid M)$. Kryterium (*segmental* ML) jest wykorzystywane w algorytmie segmental *k*-means. Kryterium to polega na maksymalizacji prawdopodobieństwa obserwacji na optymalnej ścieżce $\max_{M}(\max_{S} P(\mathbf{O},S \mid M))$.

Najczęściej stosowane są algorytmy estymacji parametrów modeli oparte na kryterium (MAP) oraz (*segmental* MAP). W algorytmach tych wykorzystywana jest pamięć o poprzednich

danych treningowych w postaci rozkładu prawdopodobieństwa modelu $g(M)$. Rozkład prawdopodobieństwa modelu $g(M)$ nazywany jest rozkładem „*prior*". Kryterium (MAP) polega na maksymalizacji prawdopodobieństwa modelu pod warunkiem sekwencji obserwacji $\max_M P(M \mid \mathbf{O}) = \max_M P(\mathbf{O} \mid M) g(M)$. W algorytmie segmental MAP stosuje się kryterium polegające na maksymalizacji prawdopodobieństwa modelu na optymalnej ścieżce $\max_M(\max_S P(M, S \mid \mathbf{O})) = \max_M(\max_S P(\mathbf{O}, S \mid M) g(M))$.

W algorytmie estymacji Bayesa najczęściej stosowane są następujące rozkłady „*prior*": rozkład Dirichleta dla jako „*prior*" dla dyskretnych rozkładów prawdopodobieństw, rozkład normalny-gamma jako „*prior*" dla rozkładów normalnych z diagonalną macierzą kowariancji oraz rozkład normalny-Wisharta jako rozkład „*prior*" dla rozkładów normalnych z pełną macierzą kowariancji.

Poza algorytmami optymalizacji modeli został opracowany algorytm, stosowany w celu adaptacji modeli HMM do mówcy tj. algorytm (MLLR) (*Maximum Likelihood Linear Regression*) opisany w pracach [113]÷[121]. W algorytmie szukane jest przekształcenie liniowe pozwalające modyfikować wektory wartości średniej oraz macierz kowariancji rozkładów prawdopodobieństw w stanach modelu HMM.

Drugim etapem związanym z zastosowaniem modeli HMM, jest etap rozpoznawania. W zależności od tego czy rozpoznawane są izolowane słowa czy fragmenty mowy ciągłej stosuje się odmienne metody.

Podczas rozpoznawania izolowanych dźwięków np. słów rozpoznawane słowo przyporządkowuje się do wzorca, który jest najmniej oddalony. Przyporządkowania dokonuje się na podstawie przyjętej miary odległości pomiędzy modelem HMM oraz rozpoznawanym dźwiękiem. Najczęstszą miarą odległości pomiędzy elementami ze zbioru modeli M_i i=1,2,…,v, v-ilość słów w słowniku a rozpoznawaną obserwacją $\mathbf{O}$ jest prawdopodobieństwo wystąpienia modelu M_i pod warunkiem obserwacji $P(M_i \mid \mathbf{O}) = P(\mathbf{O} \mid M_i) P(M_i) / P(\mathbf{O})$. Jeżeli założy się, że każdy model jest równie prawdopodobny wówczas miara ta jest równoważna $P(\mathbf{O} \mid M_i)$.

W przypadku rozpoznawania fragmentów mowy ciągłej wykorzystuje się model lingwistyczny oraz modele HMM reprezentujące najczęściej fonemy, w tym przypadku rozpoznawanie polega na znalezieniu optymalnej ścieżki w modelu lingwistycznym. Podczas szukania optymalnej ścieżki stosuje się algorytm Viterbiego. Parametry modelu lingwistycznego dobiera się na podstawie analizy tekstu, metody tworzenia modeli lingwistycznych zostały omówione między innymi w pracach [8], [9].

Stosowane modele HMM mogą posiadać różnorodną strukturę ze względu na rozkład prawdopodobieństwa w stanach:

- modele z dyskretnymi rozkładami prawdopodobieństwa w stanach DHMM
- modele z ciągłymi rozkładami prawdopodobieństwa w stanach CDHMM
- niestacjonarne modele HMM (*Trended* HMM)

W modelach DHMM (*Discrete* HMM) zastosowane są dyskretne rozkłady prawdopodobieństwa o 64 do 1024 wartościach, w celu wykorzystania tego typu modeli w systemie rozpoznawania mowy konieczne jest poddanie parametrów dźwięku kwantyzacji wektorowej VQ (*Vector Quantization*). W wyniku kwantyzacji występują straty przenoszonej informacji co wpływa na pogorszenie wyników rozpoznawania. W modelach CDHMM (*Continuous Density* HMM) zastosowane są wielowymiarowe ciągłe rozkłady Gaussa z diagonalną macierzą kowariancji, natomiast w modelach niestacjonarnych HMM zastosowane

są rozkłady normalne z wartością średnią zależną od czasu oraz diagonalną macierzą kowariancji. Dla każdego z wymienionych modeli HMM algorytmy estymacji parametrów oraz metoda wyznaczania prawdopodobieństwa obserwacji mają swoją specyfikę i różnią się pewnymi szczegółami.

Poza wymienionymi powyżej modelami HMM można znaleźć prace, w których są opisane modele HMM o specyficznej strukturze tj. modele z jawną gęstością rozkładu prawdopodobieństwa pozostawania w tym samym stanie opisane w pracy [3], modele te posiadają dodatkowe dyskretne rozkłady prawdopodobieństwa określające prawdopodobieństwo przebywania w stanie przez określony czas.

Opracowane zostały również modele auto-regresywne o specyficznych funkcjach rozkładu prawdopodobieństwa w stanach opisane w pracach [3], [88],[89]. Modele te są bezpośrednio związany z metodą LPC przetwarzania sygnału dźwięku. Podczas wyznaczania prawdopodobieństwa wygenerowania przez model wektora obserwacji w stanie wykorzystywany jest niezmienny rozkład normalny z wartością średnią równą zero oraz błąd linearyzacji, w przeciwieństwie do modeli z ciągłym rozkładem prawdopodobieństwa w stanach, w których prawdopodobieństwo wygenerowania wektora obserwacji wyznacza się na podstawie rozkładu normalnego o modyfikowanych podczas optymalizacji parametrach oraz wektora cech dźwięku.

W systemach rozpoznawania mowy zarejestrowany dźwięk poddawany jest obróbce cyfrowej. Celem obróbki jest wyodrębnienia istotnych cech dźwięku, wyeliminowanie poziomu zarejestrowanego sygnału oraz filtracja szumów. W metodach obróbki stosowane są między innymi metody wykorzystywane do kompresji dźwięku. Do stosowanych metod można zaliczyć: metodę (LPC) (*Linear Predictive Coding*) opisaną w pracach [12], [122], metodę filtracji przez zespół filtrów (*Filter Bank Analysis*) opisaną w pracy [123], metodę (MFCC) (*Mel Frequency Cepstral Coefficients*) opisaną w pracach [124],[125], metodę (PLP) (*Perceptual Linear Prediction*) opisaną w pracach [126]÷[128] oraz jej zmodyfikowaną wersję (RASTA-PLP) opisaną w pracy [127], metody falkowej filtracji opisane w pracach [129],[130] oraz metody przetwarzania przez sieci neuronowe opisane w pracach [131]÷[133]. W systemach rozpoznawana mowy najczęściej stosowana jest metoda MFCC.

Dźwięk po przetworzeniu reprezentowany jest najczęściej przez 36 wymiarowe wektory cech reprezentujących przybliżone oraz wygładzone widmo dźwięku w oknie oraz jego dynamikę w czasie. Ponieważ wyznaczanie prawdopodobieństwa dla wektorów o dużej liczbie składowych jest kosztowne obliczeniowo oraz nie zawsze reprezentacja przez tak dużą liczbę składowych jest konieczna stosuje się metody redukcji ilości składowych.

W celu redukcji ilości składowych wektora cech stosuje się metody algebry liniowej takie jak: (PCA) (*Principal Component Analysis)*, (LDA) (*Linear Dyscriminant Analysis*) oraz jej zmodyfikowaną wersja HLDA (*Heteroscadestic* LDA). Metody te pozwalają zredukować ilość składowych wektora cech dźwięku. Zastosowanie metody (PCA) można znaleźć np. w pracy [145]. Ze względu na to, że wektory reprezentujące dźwięk można podzielić na zbiory, dla których powinna zostać zachowana dyskryminacja po przekształceniu liniowym, dlatego metody (LDA) oraz (HLDA) opisane w pracach [134]÷[144] są stosowane z większym powodzeniem.

Modele HMM poza rozpoznawaniem mowy są z powodzeniem wykorzystywane w wielu innych zagadnieniach. Obecnie można znaleźć tysiące prac związanych zastosowaniami modeli HMM.

Użycie modeli HMM związane z rozpoznawaniu obrazów można znaleźć w pracy [150] dotyczącej rozpoznawania gestykulacji rąk. W pracach [152]÷[160] modele HMM zostały

zastosowane do rozpoznawania pisma odręcznego. Modele HMM wykorzystywane są również w innych zagadnieniach takich jak: rozpoznawanie stopnia zużycia elementów obrabiarek w procesie technologicznym na podstawie obserwacji przyłożonego momenty obrotowego praca [146], w celu oceny jakości ruchu w sieciach komputerowych praca [148], modelowania protein praca [150], modelowania DNA prace [147], [149]. Oprogramowanie pozwalające modelować dane ekonomiczne można znaleźć np. w pracy [9].

W niniejszej pracy został wprowadzony model sieciowy HMM o odmiennej strukturze niż spotykane w literaturze. Zaproponowany model sieciowy HMM składa się z sieci probabilistycznej dyskretnych modeli HMM. Do wprowadzonego modelu sieciowego HMM został zaprojektowany algorytm wyznaczania prawdopodobieństwa wygenerowania obserwacji **O** przez sieć oraz algorytm optymalizacji parametrów sieci oparty na kryterium (ML). Wprowadzony model sieciowy jest uogólnieniem teorii modeli HMM i sprowadza się w granicznych przypadkach do modelu podstawowego HMM omówionego w podrozdziale (3.3) oraz do modelu z jawną gęstością rozkładu prawdopodobieństwa pozostawania w tym samym stanie omówionego w podrozdziale (3.4). W pracy zostały opracowane algorytmy dla modelu z dyskretnymi rozkładami prawdopodobieństw w stanach.

Jeżeli zaproponowany model zostanie potraktowany jako sieć lingwistyczna, w której modele w węzłach sieci odpowiadają fonemom, wówczas dzięki wprowadzonemu uogólnieniu naturalne wydaje się wykorzystanie w systemie rozpoznawania mowy dodatkowego rozkładu prawdopodobieństwa związanego z długością fonemów, która może być uzależniona od ich następstwa.

W przypadku zastosowania modelu sieciowego nie jest konieczne odrębne rozpatrywanie modli HMM stanowiących modele „*fonemów*" oraz modelu lingwistycznego.

Na podstawie analizy wprowadzonego modelu sieciowego został zaproponowany algorytmu podany w podrozdziale (11.2) pozwalający na przeprowadzenie automatycznego podziały językowego. Po przeprowadzeniu automatycznego podziału konieczne będzie zbudowanie słownika wyrazów złożonego z alternatyw stanów modelu sieciowego odpowiadającym różnej wymowie danego słowa.

W pracy zwrócono szczególną uwagę na aspekty matematyczne, praktyczne zastosowanie modeli HMM w systemach rozpoznawaniu mowy oraz wpływ metod obróbki sygnału na jakość działania systemu rozpoznawania. Zostały zamieszczone wyniki rozpoznawania izolowanych sylab oraz liczb przez podstawowe modele HMM z dyskretnymi rozkładami prawdopodobieństw w stanach oraz zaproponowany model sieciowy. Rozpoznawany sygnał dźwięku został poddany obróbce metodami LPC oraz MFCC a następnie poddany kwantyzacji wektorowej.

W rozdziale trzecim przedstawione są metody związane z teorią oraz zastosowaniami modeli HMM.

W podrozdziale pierwszym i drugim (3.1), (3.2) zostały przedstawione algorytmy wyznaczania prawdopodobieństwa obserwacji oraz algorytmy optymalizacji oparte na kryterium (ML) oraz (MAP) dla modelu procesu Markowa. Podrozdziały zostały opracowane na podstawie prac [6], [8].

W podrozdziale trzecim (3.3) zostały przedstawione algorytmy wyznaczania prawdopodobieństwa obserwacji, algorytm Viterbiego szukania optymalnej ścieżki oraz algorytmy optymalizacji oparte na kryteriach: (ML), (*segmental* ML), (MAP) oraz (*segmental* MAP) dla modeli z dyskretnymi rozkładami w stanach. Podane metody zostały opisane na podstawie prac [3]÷[11].

W podrozdziale czwartym (3.4) zostały omówione modele HMM z jawnym rozkładem prawdopodobieństwa pozostawania w tym samym stanie na podstawie pracy [3].

W kolejnym podrozdziale piątym (3.5) zostały przedstawione algorytmy poprawiające, dla modeli HMM z ciągłymi rozkładami prawdopodobieństw w stanach. Podane metody zostały opisane na podstawie prac [44]÷[77]. W podrozdziale piątym została również przedstawiona opracowana przez autora pracy metoda wyznaczania rozkładu „*prior*" w metodzie został wykorzystany podział zbioru wektorów treningowych na dwa podzbiory. W wyniku zastosowania metody został wyznaczony rozkład „*prior*" dla rozkładu normalnego z diagonalną macierzą kowariancji. Wyznaczony rozkład jest to rozkład warunkowy rozkładu normalnego-Gamma. W wyniku zastosowania rozkładu warunkowego została zredukowana ilość zapamiętywanych parametrów w algorytmie adaptacji Bayesa oraz zostały zredukowane niepotrzebne działania algorytmu w porównaniu z algorytmami adaptacji Bayesa podanymi w pracach [9], [73], [74], [76].

W podrozdziale szóstym (3.6) został przedstawiony model niestacjonarny HMM. Rozdział ten został opracowany na podstawie prac [59], [63], [77]. Podobnie jak dla modelu z ciągłymi rozkładami w stanach dla modelu niestacjonarnego została wykorzystana zaproponowana przez autora metoda wyznaczenia rozkładu „*prior*". W wyniku zastosowania metody został uzyskany identyczny estymator parametrów rozkładu normalnego niestacjonarnego jak opisany w pracy [77].

W podrozdziale siódmym (3.7) zostały omówione metody skalowania składowych macierzy $\alpha_t(i)$, $\beta_t(i)$ wykorzystywanych w pozostałych algorytmach. Podrozdział został napisany na podstawie pracy [3].

W podrozdziale ósmym (3.8) zostały omówione metody estymacji modeli HMM do wielokrotnych sekwencji obserwacji.

W podrozdziale dziewiątym (3.9) została przedstawiona metoda MLLR adaptacji do mówcy na podstawie prac [113]÷[121].

W podrozdziale dziesiątym zostały omówione metody (LDA) (*Linear Dyscriminant Analysis*) oraz jej zmodyfikowaną wersja HLDA (*Heteroscadestic LDA*) na podstawie prac [134]÷[144].

W rozdziale czwartym został przedstawiony opracowany przez autora model sieciowy HMM.

W podrozdziale pierwszym (4.1) został podany algorytm obliczania prawdopodobieństwa obserwacji. W podrozdziale drugim (4.2) został podany algorytm optymalizacji modelu oparty na kryterium (ML). W kolejnych podrozdziałach trzecim (4.3) oraz czwartym (4.4) zostało pokazane, że zaproponowane metody wyznaczania prawdopodobieństwa oraz estymator parametrów modelu sieciowego sprowadzają się w granicznych przypadkach do estymatorów modelu podstawowego HMM oraz do modelu z jawną gęstością prawdopodobieństwem pozostawania w danym stanie. W podrozdziałach pięć oraz sześć (4.5), (4.6) zostały podane odpowiednio metoda skalowania dla modelu sieciowego oraz algorytm optymalizacji tego modelu dla wielokrotnych sekwencji obserwacji.

W rozdziale piątym zostało przedstawione oszacowanie złożoności obliczeniowej oraz wymagane zasoby pamięci operacyjnej komputera dla algorytmów wyznaczania prawdopodobieństwa obserwacji oraz algorytmu adaptacji opartego na kryterium (ML). Przedstawione oszacowania zostały zrobione dla modelu podstawowego, modelu z jawnym rozkładem prawdopodobieństwa pozostawania w tym samym stanie oraz modelu sieciowego HMM.

W rozdziale szóstym zostały opisane dwa zastosowane algorytmy sztucznej inteligencji tj. algorytm genetyczny, który został zaprojektowany do optymalizacji modeli podstawowych HMM oraz algorytm ILS (*Iterated Local Search*) zaprojektowany do optymalizacji modeli HMM z jawną gęstością prawdopodobieństwa pozostawania w danym stanie oraz modelu sieciowego.

W rozdziale siódmym podane są zależności wykorzystywane do wyznaczenia algorytmu Bauma-Welcha. Wyprowadzenia zależności można znaleźć w pracy [6].

W rozdziale ósmym przedstawione są wykorzystane w pracy metody tworzenia sekwencji uczących z nagranego dźwięku. Zostały wykorzystane sekwencje uczące w postaci próbek oraz wartości reprezentujących wektory cech po przetworzeniu metodami LPC oraz MFCC.

W celu adaptacji systemu rozpoznawania mowy do mówcy stosuję się algorytm Adaptacji Bayesa oraz metodą MLLR. W wyniku zastosowania algorytmu adaptacji Bayesa modyfikowane są parametry modeli HMM, podobnie jak w przypadku zastosowania metody MLLR. W podrozdziale (8.6) została przedstawiona metoda adaptacji do mówcy polegająca na modyfikacji parametrów sygnału podczas przetwarzania. Na podstawie wyników zawartych w pracy doktorskiej [174]. W podrozdziale (8.6) została zaproponowana metoda adaptacji polegająca na wprowadzeniu dodatkowej funkcji skalującej $K(f)$ bezpośrednio do metod przetwarzania sygnały tj. do metody LPC oraz MFCC. Została zaproponowana sklejana funkcja skalująca $K(f)$ wyznaczana na podstawie dwóch formantów mówcy bazowego oraz mówcy, do którego przeprowadzana jest adaptacja.

W rozdziale dziewiątym został opisany zaimplementowany w C++ przez autora pracy system wykorzystujący modele HMM. W doborze początkowych parametrów modeli HMM zostały wykorzystane wyniki eksperymentów zawarte w pracy [3].

W rozdziale dziesiątym przedstawione są wyniki z przeprowadzonych badań skuteczności zaimplementowanych algorytmów. Zostały przeprowadzone eksperymenty rozpoznawania izolowanych sylab oraz liczb z wykorzystaniem wzorców w postaci modeli podstawowych HMM, oraz zaproponowanego przez autora pracy modelu sieciowego. Zostały wykorzystane modele z dyskretnymi rozkładami prawdopodobieństw w stanach.

Ponieważ interesujące jest zagadnienie wpływu reprezentacji danych na wyniki rozpoznawania, badania zostały przeprowadzone zarówno dla dźwięku w postaci próbek oraz poddanego obróbce metodami LPC oraz MFCC.

2. Cele i tezy pracy

1. Tezy rozprawy doktorskiej:

Przedmiotem pracy są zastosowania niejawnych modeli Markowa HMM do rozpoznawania mowy oraz wykorzystanie sztucznej inteligencji do optymalizacji modeli HMM.

W szczególności spodziewane są następujące wyniki:

- uzyskanie wyników pozwalających określić skuteczność zastosowania modeli HMM do rozpoznawania izolowanych sylab w języku polskim.
- zastosowanie uogólnionego modelu HMM do rozpoznawania izolowanych sylab, słów.
- opracowanie algorytmów przybliżonych (*algorytm genetyczny oraz iteracyjny*) optymalizacji modeli HMM oraz zbadanie skuteczności tych algorytmów.

2. Uzasadnienie ważności i celowości tez:

Rozpoznawanie mowy jest jednym z bardziej złożonych zagadnień, nad którym prowadzone są badania w wielu ośrodkach naukowo badawczych na całym świecie. Pomimo trwających od wielu lat badań zagadnienie to nie jest rozwiązane w stopniu pozwalającym powszechnie stosować niezawodne systemy automatycznego rozpoznawania mowy.

Jedną z bardziej efektywnych metod wykorzystywanych w rozpoznawaniu mowy jest metoda niejawnych modeli Markowa HMM. Modele HMM są wykorzystywane jako wzorce dźwięku. W zagadnieniach związanych z modelami HMM wykorzystywany jest złożony aparat matematyczny. Najbardziej złożonym problemem związanym z zastosowania modeli HMM jest opracowanie metod ich optymalizacji.

Należy zaznaczyć, że modelowany przy pomocy modeli HMM sygnał mowy poddawany jest obróbce metodami takimi jak LPC (*Linear Predictive Coding*), MFCC (*Mel Frequency Cepstral Coefficients*), (*Filter-Bank Analysis*) w ostatniej metodzie można zastosować np.: metody filtracji falkowej, co jest kolejnym zagadnieniem badawczym, choć metody filtracji były już stosowane, ale z mniejszym powodzeniem niż metoda MFCC.

Modele HMM znajdują również zastosowania w rozpoznawaniu obrazów, modelowaniu DNA, modelowaniu danych ekonomicznych. Ogólnie modele HMM mogą być wykorzystywane do modelowania sygnałów, które mają charakter niedeterministyczny.

3. Wprowadzenie do teorii niejawnych modeli Markowa HMM

W tym została przedstawiony jedna z bardzo znanych oraz szeroko rozpowszechnionych w użyciu metoda statystyczna. Metoda ta służy do scharakteryzowania własności sygnału za pomocą wzorców probabilistycznych. Jako wzorce sygnału wykorzystywane są niejawne modele Markowa HMM (*Hidden Markov Model*).

Najważniejszym zadaniem związanym z zastosowaniem modeli HMM jest estymacja parametrów tych modeli. W algorytmach estymacji parametrów modeli HMM tj: algorytmie Bauma-Welcha, segmental *k*-mens oraz algorytmie Bayesa stosuje się algorytm (EM) polegający na maksymalizacji wprowadzonej funkcji pomocniczej. W rozdziale tym zostały szczegółowo omówione stosowane algorytmy poprawiające.

Algorytm Bauma-Welcha jest to algorytm bazujący na kryterium (ML) maksymalizacji prawdopodobieństwa obserwacji $P(\mathbf{O}|M)$ jak pokazano w równaniu (3.1). W przypadku niejawnego modelu Markowa nie ma metod wyznaczenie optymalnych parametrów modelu *M,* dla których $P(\mathbf{O}|M)$ przyjmuje wartość maksymalną, dlatego do doboru parametrów wykorzystuje się algorytm (EM) (*Expectation Maximization*), dzięki któremu znajdowane jest maksimum lokalne.

$$\max_{M} P(\mathbf{O}/M) \tag{3.1}$$

W algorytmie (EM) maksymalizuje się funkcje pomocniczą $Q(M,\widehat{M})$. Maksymalizację przeprowadza się po parametrach nowego modeli $\widehat{M}$ przy znanych parametrach starego modeli *M*. Proces maksymalizacji powtarzany jest krokowo. W kolejnych krokach algorytmu parametry starego modelu HMM zastępowane są przez parametry modelu wyznaczone w poprzednim kroku.

W algorytm segmental *k*-means wykorzystuje się optymalną ścieżkę wyznaczaną z wykorzystaniem algorytmu Viterbiego. Algorytm opiera się na kryterium (*segmental* ML) opisanym zależnością (3.2).

$$\max_{M}(\max_{S} P(\mathbf{O},S/M)) \tag{3.2}$$

W przypadku algorytmu segmental *k*-means do wyznaczenia parametrów modelu może zostać wykorzystywany algorytm (EM), co zostało pokazane w podrozdziale (3.5.4).

Cechą algorytmu (ML) jest to, że w jednym kroku algorytmu następuje znaczące dopasowanie modelu do sekwencji (*zbioru sekwencji treningowych*) treningowej wykorzystanej podczas optymalizacji, natomiast zostają zapomniane informacje modelu związane z jego wcześniejszą optymalizacją.

W algorytmie adaptacji Bayesa uwzględnia się informację zawartą w modelu HMM przed poddaniem go optymalizacji.

W celu uwzględnienia informacji o stanie modelu HMM wprowadza się dodatkowy rozkład prawdopodobieństwa modelu „*prior*" $g(M)$. Algorytm adaptacji Bayesa opiera się na kryterium (MAP) podanym w równaniu (3.3). Parametry rozkład prawdopodobieństwa $g(M)$ dobierane są na podstawie wcześniejszych sekwencji treningowych natomiast parametry

początkowe posiadają takie wartości, dla których rozkład nie zawiera informacji na temat modelu. Dla parametrów bez informacji na temat modelu zadanie maksymalizacji (MAP) sprowadza się do zadania (ML).

$$\max_M P(M/\mathbf{O}) = \max_M \frac{P(\mathbf{O}/M)g(M)}{P(\mathbf{O})} \Leftrightarrow \max_M P(\mathbf{O}/M)g(M) \tag{3.3}$$

Do rozwiązania zagadnienia maksymalizacji (MAP) „*Maximum a Priori*" podanego równaniem (3.3) wykorzystuje się algorytm (EM), w wyniku maksymalizacji wyznacza się nowy model $\hat{M}$ oraz rozkład prawdopodobieństwa „*posterior*" $\hat{g}(M)$ według zależności (3.4). W rozkładzie „*posterior*" uwzględniona jest dodatkowa informacja związana z nową sekwencją treningową **O**. Podczas optymalizacji modelu HMM do kolejnych sekwencji jako nowy rozkład „*prior*" modelu $\hat{M}$ wykorzystywany jest wyznaczony wcześniej rozkład „*posterior*".

$$\hat{g}(M) = \frac{P(\mathbf{O}/M)g(M)}{P(\mathbf{O})}, \qquad \hat{M} = \arg\max_M \frac{P(\mathbf{O}/M)g(M)}{P(\mathbf{O})} \tag{3.4}$$

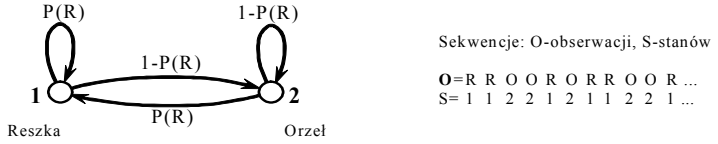
W algorytmie segmental MAP podobnie jak w algorytmie segmental *k*-means wykorzystuje się optymalną ścieżkę uzyskaną z wykorzystaniem algorytmu Viterbiego. Algorytm opiera się na kryterium opisanym zależnością (3.5).

$$\max_M(\max_S P(M \mid \mathbf{O}, S)) = \max_M(\max_S \frac{P(\mathbf{O}, S \mid M)g(M)}{P(\mathbf{O}, S)}) \Leftrightarrow \max_M(\max_S P(\mathbf{O}, S \mid M)g(M)) \tag{3.5}$$

W przypadku ostatniego algorytmu może zostać zastosowany algorytm (EM), co zostało pokazane w podrozdziale (3.5.5).

3.1. Model procesu rzucania monetą

W celu przedstawienia koncepcji algorytmów opartych na kryterium (ML) oraz (MAP) można rozpatrzyć model $M=(P(R))$ procesu rzucania monetą pokazany na rysunku (3.1). W przypadku, gdy model pracuje jako generator obserwacji należy wprowadzić prawdopodobieństwa początkowe wypadnięcia reszki δ_R, oraz wypadnięcia orła δ_O. Prawdopodobieństwa początkowe stanów powinny być równe prawdopodobieństwom dla stanu stacjonarnego: $\delta_R=P(R)$, $\delta_O=1-P(R)$. Omawiany model jest scharakteryzowany przez jedną wartość prawdopodobieństwa wypadnięcia reszki.



Rysunek 3.1. Model Markowa procesu rzucania monetą.

Jeżeli dana jest sekwencja obserwacji **O** pochodząca z procesu rzucania monetą, w której występuje η reszek oraz ν orłów, wówczas stosując kryterium (ML) maksymalizacji prawdopodobieństwa $P(\mathbf{O}|P(R)$ parametr $P(R)$ można dobrać korzystając z zależności (3.6). Kryterium maksymalizacji zostało podane dla przypadku, gdy nie jest znane uszeregowanie wypadających symboli.

$$\hat{P}(R) = \arg\max_{P(R)} P(\mathbf{O}/P(R)) = \binom{\eta+\nu}{\eta} \arg\max_{P(R)} \left[P(R)^{\eta}(1-P(R))^{\nu}\right] = \frac{\eta}{\eta+\nu} \tag{3.6}$$

Stosując zależność (3.6) zapominana jest informacja o poprzedniej wartości parametru modelu.

W celu zastosowania kryterium (MAP) na początku dobiera się rozkład „*prior*" bez informacji na temat modelu np. rozkład $g(P(R))$ o wartości równaj jeden w przedziale [0,1] jak w zależności (3.7).

$$\begin{aligned} &g(P(R)) = K\left[P(R)^{\eta-1}(1-P(R))^{\nu-1}\right] = 1, \quad dla \quad K=1, \eta=1, \nu=1, \quad 0 \le P(R) \le 1 \\ &g(P(R)) = 0, \quad dla, 0 > P(R) > 1 \end{aligned} \tag{3.7}$$

Po zastosowaniu kryterium (MAP) uzyskuje się zależność (3.8). Z zależności tej widać, że zagadnienie (MAP) sprowadza się do zagadnienia (ML).

$$\hat{P}(R) = \arg\max_{P(R)} \frac{P(\mathbf{O}/P(R))1}{\int_0^1 P(\mathbf{O}/P(R))dP(R)} = \frac{1}{\frac{(\eta)!(\nu)!}{(\eta+\nu+1)!}} \arg\max_{P(R)} \left[P(R)^{\eta}(1-P(R))^{\nu}\right] = \frac{\eta}{\eta+\nu} \tag{3.8}$$

Nowy rozkład prawdopodobieństwa „*posterior*" wyznacza się z równania (3.9). Stałą nowego rozkładu „*prior*" można pominąć, ponieważ nie ma ona wpływu na proces dalszej optymalizacji modelu.

$$\begin{aligned} &\hat{g}(P(R)) = \frac{1}{\frac{(\eta)!(\nu)!}{(\eta+\nu+1)!}} \left[P(R)^{\eta}(1-P(R))^{\nu}\right] = \hat{K}\left[P(R)^{\hat{\eta}-1}(1-P(R))^{\hat{\nu}-1}\right] \\ &\hat{\eta} = \eta+1 \\ &\hat{\nu} = \nu+1 \end{aligned} \tag{3.9}$$

Jeżeli zostanie uwzględniona nowa sekwencja treningowa **O**, w której występuje *i* reszek oraz *j* orłów wówczas stosując kryterium (MAP) parametr modelu $\hat{P}(R)$ oraz rozkład $\hat{g}(P(R))$ modyfikuje się według zależności (3.10).

$$\begin{aligned} \hat{P}(R) &= \frac{\eta+i-1}{\eta+i+\nu+j-2} = \arg\max_{P(R)} \frac{P(\mathbf{O}/P(R))g(P(R))}{\int_0^1 P(\mathbf{O}/P(R))g(P(R))dP(R)} = \\ &= \frac{1}{\frac{(\eta+i-1)!(\nu+j-1)!}{(\eta+i+\nu+j-1)!}} \arg\max_{P(R)} \left[P(R)^{\eta+i-1}(1-P(R))^{\nu+j-1}\right] \\ \hat{g}(P(R)) &= \frac{1}{\frac{(\eta+i-1)!(\nu+j-1)!}{(\eta+i+\nu+j-1)!}} \left[P(R)^{\eta+i-1}(1-P(R))^{\nu+j-1}\right] = \hat{K}\left[P(R)^{\eta+i-1}(1-P(R))^{\nu+j-1}\right] \end{aligned} \tag{3.10}$$

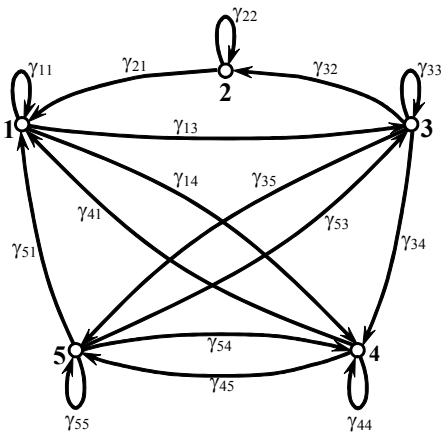
W wyniku zastosowania kryterium (MAP) estymator modelu można zapisać jak w zależnościach (3.11). Estymator pozwala wyznaczyć prawdopodobieństwo wypadnięcia reszki oraz parametry rozkładu „*posterior*".

$$\hat{P}(R) = \frac{\eta + i - 1}{\eta + \nu + i + j - 2} \qquad (3.11)$$
$$\hat{\eta} = \eta + i, \qquad \hat{\nu} = \nu + j$$

Podczas stosowania kryterium (MAP) uwzględniane są poprzednie sekwencje treningowe. Jeżeli w rozkładzie początkowym $g(P(R))$ zawarta jest istotna wiedza na temat modelu tj. suma wykładników η, ν ma dużą wartość wówczas nowe dane nie wpływają w znaczący sposób na zmianę wartość parametru $P(R)$. Stała rozkładu w zależnościach została zastąpiona przez K, stała ta nie ma wpływu na wyniki procesu optymalizacji.

3.2. Model procesu Markowa

Przykładowy model procesu Markowa został przedstawiony na rysunku (3.2). Model zbudowany jest z m stanów ponumerowanych do 1 do m. Stany modelu zazwyczaj odpowiadają pewnym fizycznym zdarzeniom modelowanego zjawiska. W regularnych odstępach czasu w modelu następują zmiany stanów (*model może powrócić do tego samego stanu*). Czas związany ze zmianami stanów są oznaczone przez $t=1,2,...,T$ aktualny stan w czasie t jest oznaczony przez S_t. Prawdopodobieństwo początkowe związane z czasem początkowym t=1 zostało oznaczone przez δ_i.



Rysunek 3.2. Model procesu Markowa z pięcioma stanami.

Dla modelu pierwszego rzędu probabilistyczna zależność następstwa stanów jest skrócona do poprzedniego stanu i opisana równaniem (3.12).

$$P(S_t = j \mid S_{t-1} = i, S_{t-2} = k, ...) = P(S_t = j \mid S_{t-1} = i) \qquad (3.12)$$

Zostaną rozważone tylko te procesy, dla których prawa strona równania (3.12) jest niezależna od czasu, wówczas prawdopodobieństwo zmiany stanów można zapisać w postaci macierzy prawdopodobieństw przejść $\Gamma=\{\gamma_{ij}\}$ dla wszystkich stanów i, j oraz wszystkich czasów t jak w równaniu (3.13).

$$\gamma_{ij} = P(S_t = j \mid S_{t-1} = i) \qquad (3.13)$$

Dla parametrów γ_{ij} zachodzą ograniczenia (3.14).

$$\gamma_{ij} \geq 0 \quad \forall i,j$$
$$\sum_{j=1}^{m} \gamma_{ij} = 1 \quad \forall i \tag{3.14}$$

Prawdopodobieństwo początkowe zdefiniowane jest jak w równaniem (3.15).

$$\delta_i = P(S_1 = i) \tag{3.15}$$

Podobnie jak dla prawdopodobieństw przejść dla prawdopodobieństw początkowych zachodzą ograniczenia (3.16).

$$\delta_i \geq 0 \quad \forall i$$
$$\sum_{i=1}^{m} \delta_i = 1 \tag{3.16}$$

Zdefiniowany model probabilistyczny zostanie oznaczany przez $M=(\delta,\Gamma)$.

Opisany model może działać jako generator sekwencji obserwacji złożonej z kolejnych stanów. Dlatego można postawić pytanie, jakie jest prawdopodobieństwo wygenerowania przez model M danej obserwacji $\mathbf{O}=(S_1,S_2,\ldots,S_T)$. Łatwo sprawdzić, że prawdopodobieństwo to można wyznaczyć korzystając z równania (3.17).

$$P(\mathbf{O} \mid M) = P(S_1, S_2, ..., S_T) = \delta_{s_1} \prod_{i=1}^{m} \prod_{j=1}^{m} \gamma ij^{f_{ij}} \tag{3.17}$$

gdzie: f_{ij} – jest to liczba przejść ze stanu i do stanu j w rozpatrywanej sekwencji stanów.

Kolejnym problemem jest wyznaczenie parametrów modelu, dla których prawdopodobieństwo (3.17) przyjmuje wartość maksymalną dla obserwacji $\mathbf{O}=(S_1,S_2,\ldots,S_T)$. Kryterium (ML) dla tego modelu przyjmuje postać podaną w równaniu (3.18).

$$\max_{\delta_{S_i},\gamma_{ij}} \delta_{s_1} \prod_{i=1}^{m} \prod_{j=1}^{m} \gamma_{ij}^{\ f_{ij}}, \qquad i,j = 1,2,...,m \tag{3.18}$$

Do rozwiązania zadania maksymalizacji (3.18) zostanie wykorzystana metoda mnożników Lagrange'a. Zamiast maksymalizować prawdopodobieństwo sekwencji należy maksymalizować logarytm prawdopodobieństwa tej sekwencji, jest to wygodne, ponieważ w wyniku logarytmowania uzyskuje się sumę składników zamiast iloczynu jak w zależności (3.19). Pierwsza część sumy związana jest z prawdopodobieństwem początkowym natomiast druga z prawdopodobieństwem przejść, dlatego obie części sumy można maksymalizować oddzielnie.

$$\max_{\delta_{S_1},\gamma_{ij}} \delta_{s_1} \prod_{i=1}^{m} \prod_{j=1}^{m} \gamma_{ij}^{\ f_{ij}} \Leftrightarrow \max_{\delta_{S_1},\gamma_{ij}} \left(\log \delta_{s_1} + \sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} \log(\gamma_{ij}) \right), \qquad i,j = 1,2,...,m \tag{3.19}$$

W wyniku maksymalizacji pierwszej część uzyskuje się składowe wektora początkowego. Można zauważyć, że rozwiązanie leży na ograniczeniach i jest postaci (3.20).

$$\delta_{S_1} = 1, \quad \delta_i = 0, \; dla\; i \neq S_1 \tag{3.20}$$

Funkcja Lagrange'a dla drugiej części podana jest w równaniu (3.21).

$$L(\log(\prod_{i=1}^{m}\prod_{j=1}^{m}\gamma_{ij}^{\ f_{ij}})) = \sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij}\log(\gamma_{ij}) - \sum_{i=1}^{m}\lambda_i\left(\sum_{j=1}^{m}\gamma_{ij} - 1\right) \tag{3.21}$$

Po zróżniczkowaniu i przyrównaniu pochodnych funkcji Lagrange'a do zera uzyskuje się układ równań (3.22). Układ ten jest złożony z *m*m* równań dla współczynników modelu oraz *m* równań dla ograniczeń.

$$\begin{cases} \underset{i,j}{\forall} & \dfrac{\partial L(\log P)}{\partial \gamma_{ij}} = f_{ij}\dfrac{1}{\gamma_{ij}} - \lambda_i = 0 \\ \underset{i}{\forall} & \dfrac{\partial L(\log P)}{\partial \lambda_i} = \left(\sum_{j=1}^{m}\gamma_{ij} - 1\right) = 0 \end{cases} \tag{3.22}$$

Z układu (3.22) można wyznaczyć optymalne parametry modelu korzystając z przekształceń (3.23).

$$\begin{aligned} &\underset{i,j}{\forall} \quad f_{ij}\frac{1}{\gamma_{ij}} - \lambda_i = 0 \\ &\underset{i,j}{\forall} \quad \gamma_{ij} = \frac{f_{ij}}{\lambda_i}, \qquad \text{oraz:} \quad \sum_{j=1}^{m}\gamma_{ij} = 1 = \sum_{j=1}^{m}\frac{f_{ij}}{\lambda_i} = \frac{1}{\lambda_i}\sum_{j=1}^{m} f_{ij} \quad \Rightarrow \quad \lambda_i = \sum_{j=1}^{m} f_{ij} \end{aligned} \tag{3.23}$$

W wyniku końcowym uzyskuje się estymator opisany zależnością (3.24).

$$\gamma_{ij} = \frac{f_{ij}}{\sum_{k=1}^{m} f_{ik}} \tag{3.24}$$

Zależności (3.20) oraz (3.24) pozwalają na wyznaczenie optymalnych parametrów modelu.

W celu zastosowania kryterium (MAP) należy wprowadzić rozkład parametrów modelu $g(M)=g(\delta_j,\gamma_{ij})$, jak w równaniu (3.25).

$$P(M) = K\prod_{i=1}^{m}\delta_i^{\ \eta_i - 1}\prod_{i=1}^{m}\prod_{j=1}^{m}\gamma ij^{\eta_{ij}-1} \tag{3.25}$$

Kryterium (MAP) przyjmuje postać podaną w równaniach (3.26).

$$\begin{aligned} &\max_{\delta_i,\gamma_{ij}} P(M/\mathbf{O}) = \max_{\delta_i,\gamma_{ij}}\frac{P(\mathbf{O}/M)g(M)}{P(\mathbf{O})} \Leftrightarrow \max_{\delta_i,\gamma_{ij}} P(\mathbf{O}/M)g(M), \quad gdzie: \\ &P(\mathbf{O}/M)g(M) = K\delta_{s_1}\prod_{i=1}^{m}\prod_{j=1}^{m}\gamma ij^{f_{ij}}\prod_{i=1}^{m}\delta_i^{\ \eta_i-1}\prod_{i=1}^{m}\prod_{j=1}^{m}\gamma ij^{\eta_{ij}-1} = K\delta_{S_1}^{\eta_{S_1}}\prod_{i\neq S_1}\delta_i^{\ \eta_i-1}\prod_{i=1}^{m}\prod_{j=1}^{m}\gamma ij^{f_{ij}+\eta_{ij}-1} \end{aligned} \tag{3.26}$$

W wyniku maksymalizacji (3.26) metodą mnożników Lagrange'a uzyskuje się estymator opisany zależnościami (3.27). Estymator pozwala wyznaczyć parametry nowego modelu $\hat{M}$ oraz parametry zmodyfikowanego rozkładu „*posterior*" $g(M)$.

$$\hat{\delta}_{S_1} = \frac{\eta_{S_1}}{1+\sum_{j=1}^{m}\eta_j - m}, \quad \hat{\delta}_i = \frac{\eta_i - 1}{1+\sum_{j=1}^{m}\eta_j - m}, \quad dla\ i \neq S_1,$$

$$\hat{\gamma}ij = \frac{f_{ij} + \eta_{ij} - 1}{\sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m}\eta_{ij} - m} \tag{3.27}$$

$$\hat{\eta}_{S_1} = \eta_{S_1} + 1, \quad \hat{\eta}_i = \eta_i, \quad dla\ i \neq S_1$$

$$\hat{\eta}_{ij} = \eta_{ij} + f_{ij}$$

Dala parametrów początkowych rozkładu „*prior*" $g(M)$ bez informacji na temat modelu odpowiednio równych η_i=1, η_{ij}=1 i,j=1,…,m, zależności estymatora (3.27) sprowadzają się do estymatora (3.20), (3.24) opartego na kryterium (ML). W rozkładzie „*prior*" może zostać pominięte działanie odejmowania jedynki w wykładniku, wówczas nastąpi uproszczenie zależności na estymator (3.27).

3.3. Niejawne modele Markowa HMM

W rozważanych w poprzednich podrozdziałach modelach każdy stan odpowiadał obserwowanemu zdarzeniu, wyjście (*obserwacja*) w takim modelu w danym stanie nie jest wartością losową, dlatego modele takie są zbyt proste, aby zastosować je do modelowania rzeczywistych procesów. W tym rozdziale zostanie pokazane rozszerzenie omawianego wcześniej modelu do niejawnego model Markowa, w którym obserwacje są probabilistyczną funkcją stanów. Schemat pięciostanowego modelu HMM przedstawiony na rysunku (3.2) jest taki sam jak dla procesu Markowa.

3.3.1. Elementy modeli HMM

Model HMM dla obserwacji przyjmujących dyskretne wartości w czasie jest scharakteryzowany przez następujące parametry:

1. m - liczba stanów w modelu. Stany są połączone w ten sposób, że każdy stan może być osiągnięty z innego stanu (*Model Ergodyczny*). Często stosuje się również modele nie ergodyczne o specyficznej strukturze. Stany są ponumerowane od *1* do m oraz stan w czasie t jest oznaczany przez S_t.
2. n - liczba obserwacji w stanie. Symbole występujące na wyjściu modelu oznaczane są przez $o \in \{1,2,...n\}$, natomiast sekwencja obserwacji, czyli uszeregowany zbiór kolejno występujących po sobie symboli oznaczana jest przez $\mathbf{O}=(o_1,o_2,..,o_T)$. Przez o_t oznaczona jest obserwacje w chwili t.
3. Macierz rozkładu prawdopodobieństw przejść $\Gamma=\{\gamma_{ij}\}$, $\gamma_{ij}=P(S_{t+1}=j|S_t=i)$, $1\leq i,j\leq m$.

$$\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1m} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2m} \\ \cdots & \cdots & & \cdots \\ \gamma_{m1} & \gamma_{m2} & \cdots & \gamma_{mm} \end{bmatrix} \tag{3.28}$$

Wartość γ_{ij} w macierzy (3.28) jest prawdopodobieństwem tego, że w kroku t+1 model znajdzie się w stanie j, przy założeniu, że w kroku t model znajdował się w stanie i. Jeżeli założy się, że każdy stan może być osiągalny z każdego innego stanu w jednym kroku wówczas $\gamma_{ij}>0$ dla wszystkich i,j.

4. Macierz rozkładu prawdopodobieństwa obserwacji $\pi=\{\pi_{ki}\}$, $\pi_{ki}=P(o_t=k|S_t=i)$, $1\le i\le m$, $1\le t\le T$.

$$\pi=\begin{bmatrix}\pi_{11} & \pi_{12} & \cdots & \pi_{1m}\\ \pi_{22} & \pi_{22} & \cdots & \pi_{2m}\\ \cdots & \cdots & & \cdots\\ \pi_{n1} & \pi_{n2} & \cdots & \pi_{nm}\end{bmatrix} \tag{3.29}$$

Macierz zawiera w kolumnach rozkłady prawdopodobieństwa obserwacji dla stanów j=1,2,...,m. Wartość π_{ki} w macierzy (3.29) jest prawdopodobieństwem tego, że model wygeneruje obserwacje $k\in\{1,2,...,n\}$ przy założeniu, że znajduje się w stanie i.

5. Wektor rozkładu prawdopodobieństw początkowych oznaczony przez $\delta=\{\delta_i\}$, $\delta_i=P(S_1=i)$, $1\le i\le m$. Składowe δ_i rozkładu początkowego określają prawdopodobieństwa, z jakimi model rozpocznie pracę w stanie i.

W powyższych podpunktach został zdefiniowaliśmy model HMM. Dla modelu będzie wykorzystywane zbiorowe oznaczenie $M=(\Gamma,\pi,\delta)$.

Prawdopodobieństwo obserwacji $\mathbf{O}=(o_1,o_2,...,o_T)$ zostanie oznaczone przez $P(\mathbf{O}|M)$.

Model M może być wykorzystany jako generator sekwencji obserwacji $\mathbf{O}=(o_1,o_2,...,o_T)$.

3.3.2. Trzy podstawowe problemy dotyczące HMM (*Wyniki analizy niejawnych modeli Markowa*)

Dany jest model HMM scharakteryzowany w poprzednim podrozdziale. Należy rozwiązać trzy problemy, aby model ten mógł być zastosowany w aplikacji rozpoznawania mowy. Problemy są następujące:

Zadanie 1) Dana jest sekwencja obserwacji $\mathbf{O}=(o_1,o_2,...,o_T)$ oraz model $M=(\Gamma,\pi,\delta)$. Należy znaleźć prawdopodobieństwo zaobserwowania tej sekwencji $P(\mathbf{O}|M)$ dla danego modelu M.

Zadanie 2) Dana jest sekwencja obserwacji $\mathbf{O}=(o_1,o_2,...,o_T)$ oraz model $M=(\Gamma,\pi,\delta)$. Należy znaleźć optymalną sekwencję stanów $S=(S_1,S_2,...,S_T)$.

Zadanie 3) Dana jest sekwencja obserwacji $\mathbf{O}=(o_1,o_2,...,o_T)$. Należy dobrać parametry modelu $M=(\Gamma,\pi,\delta)$ tak, aby maksymalizować prawdopodobieństwo $P(\mathbf{O}|M)$ zaobserwowania sekwencji $\mathbf{O}$.

Zadanie 1 jest zagadnieniem oceny. Mając dany model i sekwencje obserwacji, należy obliczyć prawdopodobieństwo, że ta sekwencja obserwacji została wygenerowana przez ten model.

Zadanie 2 jest zadaniem, w którym należy znaleźć ukryte sekwencje stanów. W zadaniu tym możliwe są dwa podejścia: wybierania stanów, które są indywidualnie najbardziej prawdopodobne lub wybieranie stanów, które tworzą najbardziej prawdopodobną sekwencje stanów. W praktyce wybiera się drugie kryterium optymalizacji.

Zadanie 3 jest zadaniem, w którym optymalizowane są parametry modelu tak, aby model z jak największym prawdopodobieństwem generował daną sekwencje obserwacji. Sekwencja obserwacji służąca do dobrania parametrów modelu jest nazywana sekwencją treningową.

W celu podsumowania można rozważyć prosty system rozpoznający odizolowane słowa. Dla wszystkich V słów ponumerowanych przez 1 do V ze słownika należy zaprojektować oddzielny m-stanowy model HMM M_v v=1,2,...,V. Sygnał mowy odpowiadający słowom reprezentowany jest przez czasową sekwencję **O** wartości z pewnego zakresu do 1 do n. Modele M_v dla poszczególnych słów można utworzyć korzystając z wyników rozwiązania zadania 3. Rozpoznanie nieznanego słowa można uzyskać korzystając z teorii Bayesa oraz rozwiązania zadania 1, rozwiązując zadanie maksymalizacji pokazane w równaniu (3.30). Jeżeli przyjmie się, że każdy model jest równie prawdopodobny $P(M_i)$=1/V, wówczas zadanie maksymalizacji (3.30) sprowadza się do znalezienia modelu, dla którego prawdopodobieństwo wygenerowanie sekwencji **O** jest największe.

$$v = \max_v P(M_v / \mathbf{O}) = \max_v \frac{P(\mathbf{O} \mid M_v)P(M_v)}{P(\mathbf{O})} = \frac{P(M_v)}{P(\mathbf{O})} \max_v P(\mathbf{O} \mid M_v) \tag{3.30}$$

Zadanie 2 jest wykorzystywane na kilka sposobów, po pierwsze rozwiązanie tego zadania wykorzystywane jest w algorytmie segmental k-means oraz segmental MAP, po drugie może ono zostać wykorzystane w celu zmniejszenia wielkości modelu reprezentującego pojedyncze słowo, jednak najważniejszym zastosowaniem rozwiązania zadania 2 jest rozpoznawanie mowy ciągłej. W przypadku rozpoznawania mowy ciągłej (pewnego fragmentu mowy ciągłej) szuka się optymalnej ścieżki w modelu lingwistycznym wykorzystując rozwiązanie zadania 2.

W dalszej części zostaną przedstawione formalne metody rozwiązania każdego z fundamentalnych problemów związanych z niejawnymi modelami Markowa.

3.3.3. Rozwiązanie problemu 1 (*Wyznaczenie prawdopodobieństwa obserwacji*)

Najprostsza metoda wyznaczenie $P(\mathbf{O}|M)$ polega na wyliczeniu wszystkich możliwych sekwencji stanów (*ścieżek*) o długości T *(długość obserwacji*), jest m^T takich sekwencji stanów. Jedna sekwencja stanów podana jest równaniem (3.31).

$$S = (S_1, S_2, ..., S_T) \tag{3.31}$$

gdzie S_1 jest stanem początkowym. Prawdopodobieństwo zaobserwowania sekwencji obserwacji **O** pod warunkiem, że znana jest sekwencja stanów S podane jest w zależności (3.32).

$$P(\mathbf{O} \mid S, M) = \prod_{i=1}^{T} P(o_t \mid S_t) \tag{3.32}$$

Obserwacje są zależne tylko od stanów stąd (3.32) można zapisać w postaci (3.33).

$$P(\mathbf{O} \mid S, M) = \pi_{o_1 S_1} \pi_{o_2 S_2} ... \pi_{o_T S_T} \tag{3.33}$$

Prawdopodobieństwo sekwencji stanów podane jest w zależności (3.34).

$$P(S \mid M) = \delta_{S_1} \gamma_{S_1 S_2} \gamma_{S_2 S_3} ... \gamma_{S_{T-1} S_T} \tag{3.34}$$

Korzystając z własności (3.33), (3.34) prawdopodobieństwo dołączone można zapisać w postaci (3.35).

$$P(\mathbf{O}, S \mid M) = P(\mathbf{O} \mid S, M) P(S \mid M) \tag{3.35}$$

Korzystając z prawdopodobieństwa warunkowego poszukiwane prawdopodobieństwo obserwacji można zapisać jak w zależności (3.36).

$$P(\mathbf{O} \mid M) = \sum_{S_1, S_2, ..., S_T} P(\mathbf{O} \mid S, M) P(S \mid M) = \sum_{S_1, S_2, ..., S_T} \delta_{S_1} \pi_{o_1 S_1} \gamma_{S_1 S_2} \pi_{o_2 S_2} \gamma_{S_2 S_3} ... \gamma_{S_{T-1} S_T} \pi_{o_T S_T} \tag{3.36}$$

Obliczenie powyższego wyrażenia wymaga $2T \cdot m^T$ działań. Jest m^T sekwencji stanów a dla każdej trzeba wykonać $2T$-1 działań mnożenia oraz wykonać sumowanie dla każdej sekwencji stanów. Dokładnie należy wykonać $(2m\text{-}1)m^T$ mnożeń oraz m^T-1 dodawań. To zadanie jest niewykonalne nawet dla niewielkich wartości parametrów np.: m=5, T=100 trzeba wykonać $2 \cdot 100 \cdot 5^{100} \approx 10^{72}$ działań. Ale istnieje lepsza procedura rozwiązująca to zagadnienie nazwana procedurą *forward-backward*.

W celu wyznaczenia algorytmu definiujemy się parametry $\alpha_t(i)$, $\beta_t(i)$ dla wszystkich stanów i oraz wszystkich t od 1 do T, jak w zależnościach (3.37).

$$\begin{aligned} \alpha_t(i) &:= P(o_1, o_2 ..., o_t, S_t = i \mid M) \\ \beta_t(i) &:= P(o_{t+1}, o_{t+2}, ..., o_T \mid S_t = i, M) \end{aligned} \tag{3.37}$$

Wykorzystując własność (2.1) z rozdziału (7) można wyznaczyć zależność (3.38).

$$\begin{aligned} \alpha_t(i)\beta_t(i) &= P(S_t = i) P(o_1, o_2, ..., o_t \mid S_t = i) P(o_{t+1}, o_{t+2}, ..., o_T \mid S_t = i) = \\ &= P(S_t = i) P(o_1, o_2, ..., o_T \mid S_t = i) = \\ &= P(o_1, o_2, ..., o_T \wedge S_t = i) \end{aligned} \tag{3.38}$$

Korzystając z prawdopodobieństwa warunkowego oraz własności prawdopodobieństwa całkowitego uzyskuje się zależność (3.39).

$$\sum_{i=1}^{m} \alpha_t(i) \beta_t(i) = P(o_1, o_2, ..., o_T) = P(\mathbf{O} \mid M) \tag{3.39}$$

Stąd prawdopodobieństwo $P(\mathbf{O}|M)$ można obliczyć dla każdego czasu t. Dla t=T uzyskuje się zależność (3.40).

$$L_T = \sum_{i=1}^{m} \alpha_T(i). \tag{3.40}$$

W celu znalezienia współczynników $\alpha_t(i)$ oraz $\beta_t(i)$ przyjmuje się, że: $\beta_T(i)$=1 oraz, że $\alpha_1(i)$ opisane jest zależnością (3.41).

$$\alpha_1(i) = P(S_1 = i) P(o_1 \mid S_1 = i) = \delta_i \pi_{o_1 i}, \tag{3.41}$$

Wartości $\alpha_1(i)$, $\beta_T(i)$ są to wartości startowe przy wyliczaniu rekurencyjnych wzorów na te parametry. Po zastosowaniu własności (2.2) z rozdziału (7) oraz własności prawdopodobieństwa warunkowego uzyskuje się zależności (3.42).

$$
\begin{aligned}
&\alpha_{t+1}(j) \\
&= \sum_{i=1}^{m} P(o_1,...,o_{t+1} \wedge S_t = i, S_{t+1} = j) \\
&= \sum P(S_t = i, S_{t+1} = j) P(o_1,...,o_{t+1} \mid S_t = i, S_{t+1} = j) \\
&= \sum P(S_t = i) \gamma_{ij} P(o_1,...,o_t \mid S_t = i) P(o_{t+1} \mid S_{t+1} = j) \\
&= \sum P(o_1,...,o_t \wedge S_t = i) \gamma_{ij} \pi_{o_{t+1} j} \\
&= \left(\sum_{i=1}^{m} \alpha_t(i) \gamma_{ij} \right) \pi_{o_{t+1} j}
\end{aligned}
\tag{3.42}
$$

Podobnie po zastosowaniu własności (2.3), dla której (l=t+1) oraz własności (2.4) z rozdziału (7) uzyskuje się zależności (3.43).

$$
\begin{aligned}
&\beta_t(i) \\
&= \sum_{j=1}^{m} P(o_{t+1},...,o_T \wedge S_t = i, S_{t+1} = j) / P(S_t = i) \\
&= \sum P(o_{t+1},...,o_T \mid S_t = i, S_{t+1} = j) P(S_t = i, S_{t+1} = j) / P(S_t = i) \\
&= \sum P(o_{t+1},...,o_T \mid S_{t+1} = j) \gamma_{ij} \\
&= \sum P(o_{t+1} \mid S_{t+1} = j) P(o_{t+2},..,o_T \mid S_{t+1} = j) \gamma_{ij} \\
&= \sum_{j=1}^{m} \pi_{o_{t+1} j} \beta_{t+1}(j) \gamma_{ij}
\end{aligned}
\tag{3.43}
$$

3.3.4. Rozwiązanie problemu 2 (*Wyznaczenie optymalnej sekwencji stanów*)

W odróżnieniu do zadania 1 w tym przypadku istnieje wiele możliwych rozwiązań. Rozwiązanie zależy od definicji optymalnej sekwencji stanów. Jednym z rozwiązań jest wybranie stanów S_t, które są indywidualnie najbardziej prawdopodobne w kolejnych chwilach czasowych. W celu wyznaczenia stanów indywidualnie najbardziej wprowadza się parametr $\chi_t(i)$ opisany równaniem (3.44).

$$
\chi_t(i) = P(S_t = i \mid \mathbf{O}) = \frac{P(\mathbf{O} \wedge S_t = i)}{P(\mathbf{O})} = \frac{P(\mathbf{O} \wedge S_t = i)}{\sum_{i=1}^{m} P(\mathbf{O} \wedge S_t = i)} = \frac{\alpha_t(i) \beta_t(i)}{\sum_{i=1}^{m} \alpha_t(i) \beta_t(i)} \tag{3.44}
$$

Korzystając z wprowadzonego parametru $\chi_t(i)$ opisanego równaniem (3.44) można znaleźć najbardziej prawdopodobny stan S_t^* w chwili t korzystając z zależności (3.45).

$$
S_t^* = \arg \max_{1 \le i \le N} \left[\chi_t(i) \right] \tag{3.45}
$$

Pomimo, że metodą tą zostaną wyznaczone stany, może się zdarzyć, że wyznaczona sekwencja będzie nieodpowiednia, jeśli HMM ma prawdopodobieństwo pomiędzy dwoma kolejnymi stanami S_t, S_{t+1} równe zero tj. $\gamma_{S_t S_{t+1}}$=0. Jedynym możliwym sposobem rozwiązania powyższego problemu jest zmodyfikowanie kryterium optymalności tak, aby maksymalizować sekwencje kilku kolejnych stanów np. dwóch stanów (S_t,S_{t+1}) lub trzech stanów (S_t,S_{t+1},S_{t+2}) i tak

dalej. W aplikacjach stosowane jest kryterium, w którym szukana jest sekwencja stanów, dla której $P(\mathbf{O}|S,M)$ przyjmuje wartość największą. Metoda znalezienia takiej sekwencji bazuje na metodzie programowania dynamicznego i jest nazywana algorytmem Viterbiego.

Algorytmem Viterbiego:

W celu znalezienia sekwencji definiuje się parametr $\eta_t(i)$ jak w równaniu (3.46).

$$\eta_t(i) = \max_{S_1,S_2,...,S_T} P(S_1,S_2,...,S_T = i \wedge o_1,o_2,...,o_t) \tag{3.46}$$

Parametr $\eta_t(i)$ jest największym prawdopodobieństwem wzdłuż pojedynczej ścieżki. Ścieżka zaczyna się w stanie S_1 a kończy w stanie S_t. Wzór indukcyjny wyznaczania parametru podany jest równaniem (3.47).

$$\eta_{t+1}(j) := [\max_i \eta_t(i)\gamma_{ij}]\pi_{o_{t+1}j} \tag{3.47}$$

Aby uzyskać ścieżkę konieczne jest zapamiętywanie argumentu maksymalnego w kolejnych krokach algorytmu. Zadanie to realizuje się wprowadzając wektor $\varphi_t(j)$. Korzystając z wprowadzonego parametru $\eta_t(i)$ algorytm szukania optymalnej ścieżki można zapisać jak w zależnościach (3.48).

$$\begin{cases} \eta_1(i) = \delta_1 \cdot \pi_{o_1 i}, & \varphi_1(i) = 0 \qquad 1 \le i \le m \\ \eta_t(j) = \left[\max\limits_{1\le i\le m}[\eta_{t-1}(i)\cdot\gamma_{ij}]\right]\cdot\pi_{o_t j}, & \varphi_t(j) = \arg\max\limits_{1\le i\le m}[\eta_{t-1}(i)\cdot\gamma_{ij}] \\ & \qquad 1 \le j \le m, \quad 2 \le t \le T \\ P^* = \max\limits_{1\le i\le m}[\eta_T(i)] & \\ S_T^* = \arg\max\limits_{1\le i\le m}[\eta_T(i)], & S_t^* = \varphi_{t+1}(S_{t+1}^*) \\ & \qquad t = T-1, T-2, ..., 1 \end{cases} \tag{3.48}$$

Główna różnica pomiędzy algorytmem na współczynniki $\alpha_t(i)$, $\beta_t(i)$ polega na tym, że w algorytmie Viterbiego wybierana jest tylko ścieżka dająca największe prawdopodobieństwo (*liczone jest maksimum*). W zależnościach (3.48) parametr P^* oznacza wartość prawdopodobieństwo na optymalnej ścieżce $S^*=(S_1^*,S_2^*,...,S_T^*)$.

Ścieżkę można również wyznaczyć z wzorów rekurencyjnych korzystając z zapamiętanych wartości w macierzy $\eta_t(i)$ jak w zależności (3.49).

$$\begin{aligned} S_T^* &= \arg\max_{1\le i\le m} \eta_T(i) \\ S_t^* &= \arg\max_{1\le i\le m}[\eta_t(i)\cdot\gamma_{iS_{t+1}^*}] \qquad dla \;\; t = T-1, T-2, ..., 1 \end{aligned} \tag{3.49}$$

Alternatywny Algorytm Viterbiego:

W alternatywnym algorytmie Viterbiego najczęściej stosuje się logarytmy parametrów modelu podane w zależnościach (3.50). Po wprowadzeniu logarytmów parametrów nie jest konieczne wykonywanie działań mnożenia.

$$\begin{aligned}
&\widehat{\delta}_i = \log(\delta_i) && 1 \le i \le m \\
&\widehat{\pi}_{o_t i} = \log(\pi_{o_t i}) && 1 \le i \le m, \quad 1 \le t \le T \\
&\widehat{\gamma}_{ij} = \log(\gamma_{ij}) && 1 \le i, j \le m
\end{aligned} \tag{3.50}$$

Po wprowadzeniu logarytmów (3.50) algorytm Viterbiego przyjmuje postać podaną w zależnościach (3.51).

$$\begin{cases}
\widehat{\eta}_1(i) = \log(\eta_1(i)) = \widehat{\delta}_1 + \widehat{\pi}_{o_1 i}, & \varphi_1(i) = 0 \qquad 1 \le i \le m \\
\widehat{\eta}_t(j) = \left[\max_{1 \le i \le m}[\widehat{\eta}_{t-1}(i) + \widehat{\gamma}_{ij}]\right] + \widehat{\pi}_{o_t j}, & \varphi_t(j) = \arg\max_{1 \le i \le m}[\widehat{\eta}_{t-1}(i) + \widehat{\gamma}_{ij}] \\
& \qquad\qquad 1 \le j \le m, \quad 2 \le t \le T \\
\widehat{P}^* = \max_{1 \le i \le m}[\widehat{\eta}_T(i)] & \\
S_T^* = \arg\max_{1 \le i \le m}\left[\widehat{\eta}_T(i)\right], & S_t^* = \varphi_{t+1}(S_{t+1}^*) \\
& \qquad\qquad t = T-1, T-2, \ldots, 1
\end{cases} \tag{3.51}$$

Po zastosowaniu alternatywnego algorytmu Viterbiego uzyskuje się taką samą ścieżkę oraz wartość $\widehat{P}^*$ logarytmu prawdopodobieństwo na optymalnej ścieżce. Alternatywny algorytm wymaga m^2T dodawań oraz przeliczenia logarytmów. Przeliczenie logarytmów może być wykonane tylko raz, czego koszt obliczeniowy jest nieistotny.

3.3.5. Rozwiązanie problemu 3 (*Estymacja parametrów modelu*)

Najtrudniejszym problemem jest dobór parametrów modelu probabilistycznego. Nie istnieją metody analityczne w zamkniętej formie pozwalające na dobór parametrów modelu, dla których $P(\mathbf{O}|M)$ jest maksymalne jak to miało miejsce dla modelu procesu Markowa.

Istnieją metody pozwalające wyznaczyć parametry modelu HMM, dla których prawdopodobieństwo $P(\mathbf{O}|M)$ osiąga lokalne minimum. W celu znalezienia lokalnego minimum funkcji prawdopodobieństwa stosuje się metody iteracyjne.

W celu zapisania procedury estymacji parametrów modelu HMM wygodnie jest zdefiniować parametr $\xi_t(i,j)$ podany w równaniu (3.52). Parametr ten oznacza prawdopodobieństwo, że model był w stanie i w czasie t oraz stanie j w czasie $t+1$ pod warunkiem, że została wygenerowana sekwencja obserwacja O.

$$\begin{aligned}
\xi_t(i,j) := p(S_t = i, S_{t+1} = j \mid \mathbf{O}) &= \frac{p(S_t = i, S_{t+1} = j \wedge \mathbf{O})}{P(\mathbf{O})} \\
&= \frac{\alpha_t(i)\gamma_{ij}\pi_{o_{t+1}j}\beta_{t+1}(j)}{\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_t(i)\gamma_{ij}\pi_{o_{t+1}j}\beta_{t+1}(j)}
\end{aligned} \tag{3.52}$$

Wcześniej podczas wyznaczania optymalnej sekwencji stanów został zdefiniowany parametr $\chi_t(i)$. W zależności (3.53) zostało podane powiązanie pomiędzy parametrami $\xi_t(i,j)$ oraz wcześniej zdefiniowanym parametrem $\chi_t(i)$.

$$\chi_t(i)=\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{m}\alpha_{\hat{t}}(i)\beta_{\hat{t}}(i)}=\frac{\alpha_t(i)\cdot\sum_{j=1}^{m}\left[\gamma_{ij}\pi_{o_{t+1}j}\beta_{t+1}(j)\right]}{\sum_{i=1}^{m}\alpha_{\hat{t}}(i)\beta_{\hat{t}}(i)}=\frac{\left[\sum_{j=1}^{m}\alpha_{t-1}(j)\gamma_{ji}\right]\pi_{o_t i}\beta_t(i)}{\sum_{i=1}^{m}\alpha_{\hat{t}}(i)\beta_{\hat{t}}(i)}$$
$$=\sum_{j=1}^{m}\frac{\alpha_t(i)\gamma_{ij}\pi_{o_{t+1}j}\beta_{t+1}(j)}{\sum_{i=1}^{m}\alpha_{\hat{t}}(i)\beta_{\hat{t}}(i)}=\sum_{j=1}^{m}\frac{\alpha_{t-1}(j)\gamma_{ji}\pi_{o_t i}\beta_t(i)}{\sum_{i=1}^{m}\alpha_{\hat{t}}(i)\beta_{\hat{t}}(i)}=\sum_{j=1}^{m}\xi_t(i,j)=\sum_{j=1}^{m}\xi_{t-1}(j,i) \tag{3.53}$$

3.3.5.1. Algorytm Bauma-Welcha estymacji parametrów dla modelu podstawowego DHMM

Algorytm Bauma-Welcha bazuje na kryterium (ML) opisanym zależnością (3.1). W celu doboru parametrów modelu wykorzystuje się algorytm (EM). W algorytmie (EM) maksymalizuje się funkcję pomocniczą $Q(M|\widehat{M})$. Wprowadzana funkcja pomocnicza jest opisana równaniem (3.54).

$$Q(M,\widehat{M})=\sum_{S_1,S_2,\ldots,S_T}P(\mathbf{O},S\mid M)\log P(\mathbf{O},S\mid\widehat{\mathrm{M}}) \tag{3.54}$$

Suma po stanach $S=(S_1,S_2,...,S_T)$ oznacza sumę po wszystkich sekwencjach stanów. Dla wprowadzonej funkcji Q prawdziwa jest własność (3.55). Dowód własności można znaleźć w pracy [47].

$$Q(M,\widehat{M})\geq Q(M,M)\Rightarrow P(\mathbf{O}\mid\widehat{M})\geq P(\mathbf{O}\mid M) \tag{3.55}$$

Dowód:

$$\begin{aligned}\log(P(\mathbf{O}\mid\widehat{M})/P(\mathbf{O}\mid M))&=\log\{\sum_{S_1,S_2,..,S_T}P(\mathbf{O},S\mid\widehat{M})/P(\mathbf{O}\mid M)\}=\\&=\log\{\sum_{S_1,S_2,..,S_T}[P(\mathbf{O},S\mid M)/P(\mathbf{O}\mid M)]P(\mathbf{O},S\mid\widehat{M})/P(\mathbf{O},S\mid M)\}\geq\\&\geq\sum_{S_1,S_2,..,S_T}[P(\mathbf{O},S\mid M)/P(\mathbf{O}\mid M)]\log[P(\mathbf{O},S\mid\widehat{M})/P(\mathbf{O},S\mid M)]=\\&=P(\mathbf{O}\mid M)^{-1}[Q(M\mid\widehat{M})-Q(M\mid M)]\geq 0\end{aligned} \tag{3.56}$$

W nierówności (3.56) korzysta się z nierówności Jensena (3.57), gdzie f jest to funkcja wklęsła natomiast $p_1+p_2+...+p_K=1$ oraz $x_i\in[a,b]$.

$$Ef(X)=\sum_{i=1}^{K}p_i f(x_i)\leq f(\sum_{i=1}^{K}p_i x_i)=f(EX) \tag{3.57}$$

W wyniku maksymalizacji funkcji pomocniczej $Q(M|\widehat{M})$ uzyskuje się nowy model $\widehat{M}$. Dla znalezionego modelu spełniona jest zależność (3.55). W praktyce stosuje się metodę wielokrotnej maksymalizacji funkcji $Q(M_i|\widehat{M}_i)$, z tym, że w kolejnych iteracjach parametry znanego modelu M_i zastępuje się parametrami modelu $\widehat{M}_{i-1}$ wyznaczonymi w poprzedniej iteracji. Powtarzając powyższą operacje wielokrotnie wartość prawdopodobieństwa $P(\mathbf{O}|M)$ dochodzi do punktu krytycznego wówczas nie następuje jego dalsza poprawa.

W celu przeprowadzenia maksymalizacji funkcji Q elementy sumy (3.54) należy zapisać w postaci (3.58).

$$P(\mathbf{O},S\mid M)=\delta_{S_1}\left(\prod_{t=1}^{T-1}\gamma_{S_tS_{t+1}}\pi_{o_tS_t}\right)\pi_{o_TS_T}$$
$$\log P(\mathbf{O},S\mid \hat{M})=\log\hat{\delta}_{S_1}+\sum_{t=1}^{T-1}\log\hat{\gamma}_{S_tS_{t+1}}+\sum_{t=1}^{T}\log\hat{\pi}_{o_tS_t} \tag{3.58}$$

Po wprowadzeniu zależności (3.58) funkcję $Q(M|\widehat{M})$ można zapisać jak w równaniu (3.59).

$$\begin{aligned}Q(M\mid\widehat{M})=&\\ \sum_{S_1,S_2,\ldots,S_T}&\left\{P(\mathbf{O},S\mid M)\log\widehat{\delta}_{S_1}+P(\mathbf{O},S\mid M)\sum_{t=1}^{T-1}\log\widehat{\gamma}_{S_tS_{t+1}}+P(\mathbf{O},S\mid M)\sum_{t=1}^{T}\log\widehat{\pi}_{o_tS_t}\right\}=\\ &=\sum_{i=1}^{m}P(\mathbf{O},S_1=i\mid M)\log\widehat{\delta}_i+\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{t=1}^{T-1}P(\mathbf{O},S_t=i,S_{t+1}=j\mid M)\log\widehat{\gamma}_{ij}+\\ &\qquad+\sum_{i=1}^{m}\sum_{k=1}^{n}\sum_{(po\ \ t\ \ t.\dot{z}.\,o_t=k)}P(\mathbf{O},S_t=i\mid M)\log\widehat{\pi}_{ki}\end{aligned} \tag{3.59}$$

Funkcja $Q(M|\widehat{M})$ została wyrażona w postaci sumy trzech elementów. Ponieważ każdy z elementów sumy jest funkcją innych parametrów nowego modelu $\widehat{M}$, dlatego można maksymalizować oddzielnie każdą składową. Parametry nowego modelu muszą spełniać warunki normalizacji (3.60).

$$\sum_{j=1}^{m}\widehat{\delta}_j=1,\qquad\sum_{j=1}^{m}\widehat{\gamma}_{ij}=1\qquad\forall j,\qquad\sum_{k=1}^{n}\widehat{\pi}_{ki}=1\qquad\forall i, \tag{3.60}$$

Elementy funkcji Q mają postać (3.61) podobnie jak w przypadku modelu procesu Markowa.

$$\sum_{j=1}^{m}w_j\log y_j \tag{3.61}$$

$$\sum_{j=1}^{m}y_j=1,\qquad y_j\geq 0 \tag{3.62}$$

Rozwiązanie globalnego maksimum dla wyrażenia (3.61) przy ograniczeniach (3.62) przyjmuje postać (3.63).

$$y_j=\frac{w_j}{\sum_{j=1}^{m}w_j},\qquad j=1,2,\ldots,m \tag{3.63}$$

Wykorzystując (3.63) w (3.59) oraz wprowadzając parametry (3.44), (3.52) uzyskuje się estymator parametrów modelu podany w zależnościach (3.64).

$$\widehat{\delta}_i = \frac{P(\mathbf{O}, S_1 = i)}{P(\mathbf{O})} = \frac{\alpha_1(i)\beta_1(i)}{\sum_{j=1}^{m}\alpha_T(i)} = \chi_1(i)$$

$$\widehat{\gamma}_{ij} = \frac{\sum_{t=1}^{T-1} P(\mathbf{O}, S_t = i, S_{t+1} = j)}{\sum_{t=1}^{T-1} P(\mathbf{O}, S_t = i)} = \frac{\sum_{t=1}^{T-1} \alpha_t(i)\gamma_{ij}\pi_{o_t j}\beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \chi_t(i)} \tag{3.64}$$

$$\widehat{\pi}_{ki} = \frac{\sum_{(po\ \ t\ \ t.\dot{z}.o_t=k)} P(\mathbf{O}, S_t = i)}{\sum_{k=1}^{n}\sum_{(po\ \ t\ \ t.\dot{z}.o_t=k)} P(\mathbf{O}, S_t = i)} = \frac{\sum_{(po\ \ t\ \ t.\dot{z}.o_t=k)} \alpha_t(i)\beta_t(i)}{\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)} = \frac{\sum_{(po\ \ t\ \ t.\dot{z}.o_t=k)} \chi_t(i)}{\sum_{t=1}^{T}\chi_t(i)}$$

Opisany algorytm jest stosowany iteracyjnie. Końcowa wartość uzyskanego prawdopodobieństwa $P(\mathbf{O}|M)$ zależy od doboru początkowych parametrów modelu.

Nie istnieją metody wyznaczania parametrów optymalnych globalnie dla modeli HMM, dlatego celowe wydaje się zbadanie wyników zastosowana algorytmów przybliżonych do optymalizacji modeli HMM.

3.3.5.2. Algorytm Bayesa estymacji parametrów modelu podstawowego DHMM

Algorytm Bayesa bazuje na kryterium (MAP) opisanym zależnością (3.3). Jako rozkład „*prior*" w kryterium (MAP) wprowadza się rozkład Dirichleta $g(M)=g(\delta,\gamma,\Gamma)$ podany równaniem (3.65).

$$g(M) = K\prod_{i=1}^{m}\delta_i^{\eta_i - 1}\prod_{i=1}^{m}\prod_{j=1}^{m}\gamma_{ij}^{\eta_{ij}-1}\prod_{i=1}^{m}\prod_{k=1}^{n}\pi_{kj}^{\nu_{ki}-1} \tag{3.65}$$

W celu doboru parametrów modelu wykorzystuje się algorytm (EM). Zmodyfikowana funkcja pomocnicza R podana jest w równaniu (3.66).

$$\begin{aligned} R(M,\widehat{M}) &= \frac{1}{P(\mathbf{O}\mid M)}Q(M,\widehat{M}) + \log g(\hat{M}) = \\ &= \frac{1}{P(\mathbf{O}\mid M)}\sum_{S_1,S_2,\ldots,S_T} P(\mathbf{O}, S \mid M)\log P(\mathbf{O}, S \mid \widehat{\mathrm{M}}) + \log g(\hat{M}) \end{aligned} \tag{3.66}$$

Funkcja pomocnicza Q jest zdefiniowana w taki sam sposób jak funkcja Q w Algorytmie Bauma-Welcha. Można pokazać, że podobnie jak dla funkcji Q dla funkcji R spełniona jest zależność (3.67).

$$R(M,\widehat{M}) \geq R(M,M) \Rightarrow P(\mathbf{O}\mid\widehat{M}) \geq P(\mathbf{O}\mid M) \tag{3.67}$$

Po przekształceniu funkcji R uzyskuje się zależność (3.68).

$$
\begin{aligned}
R(M \mid \widehat{M}) &= \frac{1}{P(\mathbf{O} \mid M)} \sum_{i=1}^{m} P(\mathbf{O}, S_1 = i \mid M) \log \widehat{\delta}_i + \\
&+ \frac{1}{P(\mathbf{O} \mid M)} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{t=1}^{T-1} P(\mathbf{O}, S_t = i, S_{t+1} = j \mid M) \log \widehat{\gamma}_{ij} + \\
&+ \frac{1}{P(\mathbf{O} \mid M)} \sum_{i=1}^{m} \sum_{k=1}^{n} \sum_{(po \; t \; t.\dot{z}.\, o_t = k)} P(\mathbf{O}, S_t = i \mid M) \log \widehat{\pi}_{ki} + \\
&+ \sum_{i=1}^{m} (\eta_i - 1) \log \widehat{\delta}_i + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{t=1}^{T-1} (\eta_{ij} - 1) \log \widehat{\gamma}_{ij} + \sum_{i=1}^{m} \sum_{k=1}^{n} P(\nu_{kj} - 1) \log \widehat{\pi}_{o_t i} + \log K = \\
&= \frac{1}{P(\mathbf{O})} \sum_{i=1}^{m} c_i \log \widehat{\delta}_i + \frac{1}{P(\mathbf{O})} \sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij} \log \widehat{\gamma}_{ij} + \frac{1}{P(\mathbf{O})} \sum_{i=1}^{m} \sum_{k=1}^{n} d_{ki} \log \widehat{\pi}_{ki} + \\
&+ \sum_{i=1}^{m} (\eta_i - 1) \log \widehat{\delta}_i + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{t=1}^{T-1} (\eta_{ij} - 1) \log \widehat{\gamma}_{ij} + \sum_{i=1}^{m} \sum_{k=1}^{n} (\nu_{kj} - 1) \log \widehat{\pi}_{ki} + \log K
\end{aligned}
\tag{3.68}
$$

Współczynniki $c_i = \alpha_1(i)\beta_1(i)$, $c_{ij} = \sum_{t=1}^{T-1} \alpha_t(i) \gamma_{ij} \pi_{o_t j} \beta_{t+1}(j)$, $d_{ki} \sum_{(po \; t \; t.\dot{z}.\, o_t = k)} \alpha_t(i) \beta_t(i)$ można wyznaczyć przy pomocy algorytmu *forward-backward*. W wyniku maksymalizacji funkcji R metodą mnożników Lagrange'a uzyskuje się estymator Bayesa podany w zależnościach (3.69).

$$
\begin{aligned}
\widehat{\delta}_i &= \frac{\frac{1}{P(\mathbf{O})} c_i + \eta_i - 1}{1 + \sum_{i=1}^{m} \eta_i - m}, \\
\widehat{\gamma}_{ij} &= \frac{\frac{1}{P(\mathbf{O})} c_{ij} + \eta_{ij} - 1}{\frac{1}{P(\mathbf{O})} \sum_{j=1}^{m} c_{ij} + \sum_{j=1}^{m} \eta_{ij} - m} \\
\widehat{\pi}_{ki} &= \frac{\frac{1}{P(\mathbf{O})} d_{ki} + \nu_{ki} - 1}{\frac{1}{P(\mathbf{O})} \sum_{k=1}^{n} d_{ki} + \sum_{k=1}^{n} \nu_{ki} - n}
\end{aligned}
\tag{3.69}
$$

Współczynniki rozkładu „*posterior*" powinny być modyfikowane po ustalonej liczbie kroków algorytmu (3.69) według zależności (3.70). Wartości początkowe współczynników rozkładu modelu $g(M)$ powinny być równe jeden. Jeżeli wartości współczynników są równe jeden wówczas algorytm (3.69) sprowadza się do algorytmu Bauma-Welcha.

$$
\begin{aligned}
\hat{\eta}_i &= \frac{1}{P(\mathbf{O})} c_i + \eta_i, \\
\hat{\eta}_{ij} &= \frac{1}{P(\mathbf{O})} c_{ij} + \eta_{ij}, \\
\hat{\nu}_{ki} &= \frac{1}{P(\mathbf{O})} d_{ki} + \nu_{ki}.
\end{aligned}
\tag{3.70}
$$

Do funkcji pomocniczej R został wprowadzony dodatkowy czynnik $1/P(\mathbf{O}|M)$ nie uwzględniany w odpowiednikach zależności (3.69), (3.70) opisanych np. w pracy [74] przy takiej samej definicji współczynników c_i, c_{ij}, d_{ki}. Czynnik ten ma istotny wpływ na postać algorytmu. Uzasadnione jest porównanie wyników uzyskanych dzięki zastosowaniu algorytmu bez współczynnika oraz zaproponowanego w tym podrozdziale. W wyniku wprowadzenie dodatkowego czynnika sekwencje, dla których prawdopodobieństwo $P(\mathbf{O}|M)$ ma małą wartość mają większy wpływ na modyfikację parametrów modelu.

3.3.5.3. Algorytm segmental *k*-means estymacji parametrów modelu podstawowego DHMM

Algorytm segmental *k*-means bazuje na kryterium (*segmental* ML) opisanym zależnością (3.2). W algorytmie segmental *k*-means wykorzystuje się optymalną ścieżkę uzyskaną dzięki algorytmowi Viterbiego. Prawdopodobieństwo na optymalnej ścieżce $S^*=S_1^*,S_2^*,\ldots,S_T^*$ wyrażone jest zależnością (3.71).

$$P(\mathbf{O},S^*\mid M)=\delta_{S_1^*}\left(\prod_{t=1}^{T-1}\gamma_{S_t^*S_{t+1}^*}\pi_{o_tS_t^*}\right)\pi_{o_TS_T^*}=\delta_{S_1^*}\left(\prod_{i=1}^{m}\prod_{j=1}^{m}\gamma_{ij}^{f_{ij}}\prod_{i=1}^{m}\prod_{k=1}^{n}\pi_{ki}^{g_{ki}}\right) \tag{3.71}$$

gdzie: f_{ij} – ilość przejść ze stanu i do stanu j, g_{ki} - ilość obserwacji k w stanie i na optymalnej ścieżce.

W celu znalezienia parametrów modelu należy maksymalizować logarytm prawdopodobieństwa na optymalnej ścieżce. Po wprowadzeniu logarytmu prawdopodobieństwa (3.71) kryterium maksymalizacji przyjmuje postać podaną w równaniu (3.72).

$$\max_{\delta_i,\gamma_{ij},\pi_{ki}}\left(\log\delta_{S_1^*}+\sum_{i=1}^{m}\sum_{j=1}^{m}f_{ij}\log\gamma_{ij}+\sum_{i=1}^{m}\sum_{k=1}^{n}g_{ki}\log\pi_{ki}\right) \tag{3.72}$$

Korzystając z metody mnożników Lagrange'a z zależności (3.72) można wyznaczyć estymator podany w zależnościach (3.73).

$$\begin{aligned}&\widehat{\delta}_{S_1^*}=1,\quad \delta_i=0\quad dla\ i\neq S_1^*\\ &\widehat{\gamma}_{ij}=\frac{f_{ij}}{\sum_{j=1}^{m}f_{ij}}\\ &\widehat{\pi}_{ki}=\frac{g_{ki}}{\sum_{k=1}^{n}g_{ki}}\end{aligned} \tag{3.73}$$

3.3.5.4. Algorytmu segmental MAP estymacji parametrów modelu podstawowego DHMM

Algorytm segmental MAP bazuje na kryterium (*segmental* MAP) opisanego zależnością (3.5). W celu zastosowania kryterium wprowadza się rozkład Dirichleta $g(M)=g(\delta,\gamma,\Gamma)$ podany równaniem (3.65) jako rozkład „*prior*" dla rozkładów w modelu DHMM. W algorytmie segmental MAP wykorzystuje się optymalną ścieżkę uzyskaną dzięki algorytmowi Viterbiego. Prawdopodobieństwo na optymalnej ścieżce wyrażone jest zależnością (3.71).

W celu znalezienia parametrów przeprowadza się zadanie maksymalizacji podane równaniem (3.74).

$$\max_{\delta_i,\gamma_{ij},\pi_{kj}} (\log(P(\mathbf{O},S^* \mid M)g(M))) \tag{3.74}$$

Maksymalizowaną funkcję można zapisać w postaci podanej w zależności (3.75).

$$\begin{aligned}
&\log(P(\mathbf{O},S^* \mid M)g(M)) = \\
&= \log(\delta_{S_1^*}\left(\prod_{i=1}^{m}\prod_{j=1}^{m}\gamma_{ij}^{f_{ij}}\prod_{i=1}^{m}\prod_{k=1}^{n}\pi_{ki}^{g_{ki}}\right)K\prod_{i=1}^{m}\delta_i^{\eta_i-1}\prod_{i=1}^{m}\prod_{j=1}^{m}\gamma_{ij}^{\eta_{ij}-1}\prod_{i=1}^{m}\prod_{k=1}^{n}\pi_{ki}^{\nu_{ki}-1}) = \\
&= \log\delta_{S_1^*}^{\eta_{S_1}} + \sum_{i\neq S_1}\log\delta_i^{\eta_i-1} + \sum_{i=1}^{m}\sum_{j=1}^{m}\log\gamma_{ij}^{f_{ij}+\eta_{ij}-1} + \sum_{i=1}^{m}\sum_{k=1}^{n}\log\pi_{ki}^{g_{ki}+\nu_{ki}-1} = \\
&= \eta_{S_1^*}\log\delta_{S_1^*} + \sum_{i\neq S_1^*}(\eta_i-1)\log\delta_i + \sum_{i=1}^{m}\sum_{j=1}^{m}(f_{ij}+\eta_{ij}-1)\log\gamma_{ij} + \sum_{i=1}^{m}\sum_{k=1}^{n}(g_{ki}+\nu_{ki}-1)\log\pi_{ki}
\end{aligned} \tag{3.75}$$

Po zastosowaniu metody mnożników Lagrange'a do maksymalizacji funkcji (3.75) uzyskuje się estymator segmental MAP podany w zależnościach (3.76).

$$\begin{aligned}
&\hat{\delta}_{S_1^*} = \frac{\eta_{S_1^*}}{1+\sum_{i=1}^{m}\eta_i - m}, \quad \hat{\delta}_i = \frac{\eta_i - 1}{1+\sum_{i=1}^{m}\eta_i - m}, \quad dla \quad i \neq S_1^* \\
&\hat{\gamma}_{ij} = \frac{f_{ij}+\eta_{ij}-1}{\sum_{j=1}^{m}f_{ij} + \sum_{j=1}^{m}\eta_{ij} - m} \\
&\hat{\pi}_{ki} = \frac{g_{ki}+\nu_{ki}-1}{\sum_{k=1}^{n}g_{ki} + \sum_{k=1}^{n}\nu_{ki} - n}
\end{aligned} \tag{3.76}$$

Współczynniki rozkładu „*posterior*" powinny być modyfikowane po każdym kroku algorytmu według zależności (3.77).

$$\begin{aligned}
&\hat{\eta}_{S_1^*} = \eta_{S_1^*} + 1, \quad \hat{\eta}_i = \eta_i, \quad dla\ i \neq S_1^* \\
&\hat{\eta}_{ij} = f_{ij} + \eta_{ij} \\
&\hat{\nu}_{ki} = g_{ki} + \nu_{ki}.
\end{aligned} \tag{3.77}$$

3.4. Zastosowanie jawnej gęstości rozkładu prawdopodobieństwa pozostawania w tym samym stanie

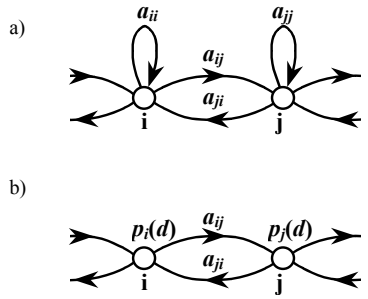
W tym rozdziale został opisany model HMM z jawnym rozkładem prawdopodobieństwa pozostawania w tym samym stanie. Prawdopodobieństwo pozostawania w stanie S_i przez okres d dla modelu podstawowego można wyrazić równaniem (3.78).

$$p_i(d) = \gamma_{ii}^{d-1}(1-\gamma_{ii}) \tag{3.78}$$

Po uwzględnieniu tego, że $\gamma_{ii}<1$ łatwo zauważyć, że prawdopodobieństwo $p_i(d)$ bardzo szybko maleje. Przez autorów pracy [3] zostało zaproponowane rozwiązaniem tego problemu

poprzez wprowadzenie rozkładu czasu pozostawania w danym stanie. Zakłada się, że w modelu nie istnieje możliwość przejścia do tego samego stanu (γ_{ii}=0).

Na rysunku (3.3) pokazana jest różnica między klasycznym modelem, a modelem z jawną funkcją rozkładu określającą prawdopodobieństwo pozostania w tym samym stanie, model taki będzie nazywany modelem z czasem przebywania.



Rysunek 3.3. Różnica pomiędzy modelem klasycznym rysunek a), a modelem z czasem przebywania rysunek b)

W opisanym modelu wyjście ze stanu *i* następuje po czasie *d* (*od momentu wejścia do tego stanu*) zgodnie z danym rozkładem $p_i(d)$.

W celu wyznaczenia dla zdefiniowanego modelu prawdopodobieństwa $P(\mathbf{O}|M)$ należy wprowadzić szereg zmian do procedur dla modelu podstawowego. Wprowadza się parametr $\alpha_t(i)$ podany w zależności (3.79).

$$\alpha_t(i) = P(o_1, o_2, ..., o_t, \ \textit{przebywanie w stanie } i \textit{ kończy się w chwili } t) \tag{3.79}$$

Ponadto zakłada się, że model był w *r* stanach podczas pierwszych *t* obserwacji oraz jeśli stany te zostaną oznaczone przez $S_1,S_2,...,S_r$, oraz czas pozostawania w każdym stanie zostanie oznaczony przez $d_1,d_2,...,d_r$, wówczas zachodzi zależność (3.80).

$$\begin{aligned} &S_r = i \\ &\sum_{k=1}^{r} d_k = t \end{aligned} \tag{3.80}$$

Równanie (3.79) może być zapisane jako w zależności (3.81).

$$\begin{aligned} \alpha_t(i) = \sum_S \sum_d & \delta_{S_1} \cdot p_{S_1}(d_1) \cdot P(o_1, o_2, ..., o_{d_1} \mid S_1) \\ & \cdot \gamma_{S_1 S_2} p_{S_2}(d_2) \cdot P(o_{d_1+1}, o_2, ..., o_{d_2} \mid S_2) \\ & \cdot \gamma_{S_2 S_3} p_{S_3}(d_3) \cdot P(o_{d_1+d_2+1}, o_2, ..., o_{d_3} \mid S_3) \quad ... \\ & \cdot \gamma_{S_{r-1} S_r} p_{S_r}(d_r) \cdot P(o_{d_1+d_2+...+d_{r-1}+1}, o_{d_1+d_2+...+d_{r-1}+2}, ..., o_t \mid S_r) \end{aligned} \tag{3.81}$$

Suma w zależności (3.81) jest wyznaczana dla wszystkich możliwych kombinacjach stanów *S* oraz czasów pozostawania *d*. Stosując indukcje zależność (3.81) można zapisać jak w równaniu (3.82).

$$\alpha_t(i) = \delta_i p_i(t) \prod_{t'=1}^{t} \pi_{o_{t'} i} + \sum_{j=1}^{m} \sum_{d=1}^{D} \alpha_{t-d}(j) \gamma_{ji} p_i(d) \prod_{t'=t-d+1}^{t} \pi_{o_{t'} i} \tag{3.82}$$

Parametr *D* oznacza maksymalny czas pozostawania w danym stanie. Jest to stała wspólna dla wszystkich stanów. Równanie (3.82) może być wykorzystane dla $t > D$, wówczas należy w nim opuścić pierwszą część sumy związaną z prawdopodobieństwem początkowym δ_i.

W zależnościach (3.83) wypisane zostały zależności na pierwsze trzy współczynniki $\alpha_t(i)$.

$$\begin{aligned}
\alpha_1(i) &= \delta_i p_i(1)\pi_{o_1 i} \\
\alpha_2(i) &= \delta_i p_i(2)\prod_{t'=1}^{2}\pi_{o_{t'} i} + \sum_{\substack{j=1 \\ j\neq i}}^{m}\alpha_1(j)\gamma_{ji}p_i(1)\pi_{o_2 i} \\
\alpha_3(i) &= \delta_i p_i(3)\prod_{t'=1}^{3}\pi_{o_{t'} i} + \sum_{d=1}^{2}\sum_{\substack{j=1 \\ j\neq i}}^{m}\alpha_{3-d}(j)\gamma_{ji}p_i(d)\prod_{t'=4-d}^{3}\pi_{o_{t'} i}
\end{aligned} \tag{3.83}$$

Podobnie jak dla modelu tradycyjnego prawdopodobieństwo $P(\mathbf{O}|M)$ można obliczyć z równania (3.84).

$$P(\mathbf{O}/M) = \sum_{i=1}^{m}\alpha_T(i) \tag{3.84}$$

W celu wyznaczenia estymatora parametrów należy dodatkowo zdefiniować parametry (3.85).

$$\begin{aligned}
\alpha_t^*(i) &= P(o_1, o_2, ..., o_t,\ \textit{następuje przejscie do stanu } i \textit{ w chwili } t+1) \\
\beta_t(i) &= P(o_{t+1}, o_{t+2}, ..., o_T \mid \textit{przebywanie w stanie } i \textit{ kończy się w chwili } t) \\
\beta_t^*(i) &= P(o_{t+1}, o_{t+2}, ..., o_T \mid \textit{następuje przejscie do stanu } i \textit{ w chwili } t+1)
\end{aligned} \tag{3.85}$$

Parametr $\beta_t^*(i)$ wyznacza się rekurencyjnie podobnie jak parametr $\alpha_t(i)$. W zależnościach (3.86) zostały wypisane zależności na pierwsze cztery współczynniki $\beta_t^*(i)$.

$$\begin{aligned}
&\beta_T^*(i) = 1, \qquad \beta_{T-1}^*(i) = p_i(1)\pi_{o_T i} \\
&\beta_{T-2}^*(i) = p_i(2)\prod_{t'=T-1}^{T}\pi_{o_{t'} i} + p_i(1)\pi_{o_{T-1} i}\sum_{\substack{j=1 \\ j\neq i}}^{m}\gamma_{ij}\beta_{T-1}^*(j) \\
&\beta_{T-3}^*(i) = p_i(3)\prod_{t'=T-2}^{T}\pi_{o_{t'} i} + \sum_{d=1}^{2}p_i(d)\prod_{t'=T-2}^{T+d-3}\pi_{o_{t'} i}\sum_{\substack{j=1 \\ j\neq i}}^{m}\gamma_{ij}\beta_{T+d-3}^*(j)
\end{aligned} \tag{3.86}$$

Pomiędzy parametrami $\alpha_t(i)$, $\alpha_t^*(i)$, $\beta_t(i)$ oraz $\beta_t^*(i)$ zachodzą zależności (3.87).

$$\begin{aligned}
\alpha_t^*(i) &= \sum_{i=1}^{m}\alpha_t(i)\gamma_{ij} \\
\alpha_t(i) &= \sum_{d=1}^{D}\alpha_{t-d}^*(i)p_i(d)\prod_{t'=t-d+1}^{t}\pi_{o_{t'} i} \\
\beta_t(i) &= \sum_{j=1}^{m}\gamma_{ij}\beta_j^*(j) \\
\beta_t^*(i) &= \sum_{d=1}^{D}\beta_{t+d}(i)p_i(d)\prod_{t'=t+1}^{t+d}\pi_{o_{t'} i}
\end{aligned} \tag{3.87}$$

W oparciu na zależnościach (3.87) nowe parametry modelu wyznaczane są według zależności (3.88).

$$\widehat{\delta}_i = \frac{\delta_i \beta_0^*(i)}{P(\mathbf{O} \mid M)}$$

$$\widehat{\gamma}_{ij} = \frac{\sum_{t=1}^{T} \alpha_t(i)\gamma_{ij}\beta_t^*(j)}{\sum_{j=1}^{m}\sum_{t=1}^{T} \alpha_t(i)\gamma_{ij}\beta_t^*(j)}$$

$$\widehat{\pi}_{ki} = \frac{\sum_{(po\ \ t\ \ t.\dot{z}.o_t=k)} \left[\sum_{\tau<t} \alpha_\tau^*(i)\cdot\beta_\tau^*(i) - \sum_{\tau<t}\alpha_\tau(i)\cdot\beta_\tau(i)\right]}{\sum_{k=1}^{n}\sum_{(po\ \ t\ \ t.\dot{z}.o_t=k)} \left[\sum_{\tau<t} \alpha_\tau^*(i)\cdot\beta_\tau^*(i) - \sum_{\tau<t}\alpha_\tau(i)\cdot\beta_\tau(i)\right]} \tag{3.88}$$

$$\widehat{p}_i(d) = \frac{\sum_{t=1}^{T} \alpha_t^*(i)\, p_i(d)\, \beta_{t+d}(i) \prod_{t'=t+1}^{t+d} \pi_{o_{t'},i}}{\sum_{d=1}^{D}\sum_{t=1}^{T} \alpha_t^*(i)\, p_i(d)\, \beta_{t+d}(i) \prod_{t'=t+1}^{t+d} \pi_{o_{t'},i}}$$

Interpretacja formuł estymacji podanych w zależnościach (3.88) jest następująca: formuła estymacji $\widehat{\delta}_i$ jest prawdopodobieństwem tego, że stan *i* był pierwszym stanem dla danej obserwacji **O** do prawdopodobieństwa obserwacji, estymator parametru $\widehat{\gamma}_{ij}$ można zinterpretować jako oczekiwaną liczbę przejść ze stanu *i* do stanu *j* do oczekiwanej liczby przejść ze stanu *i* do dowolnego stanu, estymator parametru $\widehat{\pi}_{ki}$ można zinterpretować jako oczekiwaną liczbę pojawienia się obserwacji *k* w stanie *i* do oczekiwanej liczby przypadków pojawienia się jakiejkolwiek obserwacji w stanie *i*, estymator parametru $\widehat{p}_i(d)$ można zinterpretować jako oczekiwaną liczbę przypadków pozostawania w stanie *i* przez okres *d* do oczekiwanej liczby przypadków pozostawania w stanie *i* przez dowolny okres.

Modele z jawną gęstością rozkładu prawdopodobieństwa pozostawania w danym stanie dla pewnych problemów wykazuję znaczną poprawę jakości uzyskiwanych wyników. Jednym z problemów zastosowania omówionego modelu jest duża złożoność obliczeniowa. Złożoność obliczeniowa rośnie wykładniczo z parametrem *D*, dlatego jedynie dla niewielkich wartości parametru *D* możliwe jest zastosowanie tych modeli w praktycznych aplikacjach. Kolejnym problemem jest to, że podczas estymacji parametrów należy wyznaczyć dodatkowe *D* parametrów dla każdego stanu. Z powodów dużej złożoności obliczeniowej estymacji parametrów modelu nie może być wykorzystana taka sama ilość danych jak dla modelu podstawowego HMM.

3.5. Model HMM z ciągła gęstością rozkładu prawdopodobieństwa obserwacji w stanach CDHMM

W poprzednich rozdziałach pracy omawiane były modelach HMM z dyskretnymi rozkładami prawdopodobieństwa w stanach. W przypadku omówionych modeli obserwacje charakteryzują się tym, że przyjmują dyskretne wartości ze skończonego zbioru.

W przypadku modeli z ciągłymi rozkładami w stanach CDHMM (*Continuous Density* HMM) składowe $\mathbf{o}_t$ sekwencji obserwacji $\mathbf{O}=(\mathbf{o}_1,\ldots,\mathbf{o}_T)$ dla poszczególnych chwil czasowych $t=1,2,\ldots T$ są to wektory o składowych przyjmujących wartości ciągłe. Pierwsze składowe

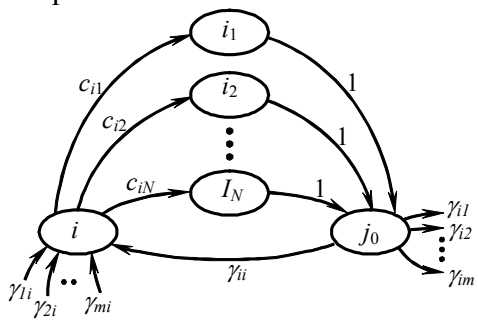
wektora obserwacji **o** odpowiadają wektorom (*cepstrum*) $c_1,\ldots,c_L$ dźwięku, natomiast kolejne składowe odpowiadają odpowiednio pierwszym i drugim pochodnym tego wektora $\mathbf{o}=(c_1,\ldots c_L, \Delta c_1,\ldots,\Delta c_L, \Delta^2 c_1,\ldots,\Delta^2 c_L)=(o_1,\ldots,o_D)$. Wymiar wektora obserwacji najczęściej wynosi $D=3*L$, gdzie L jest to ilość wyznaczanych współczynników (*cepstrum*) dla sygnału.

W celu modelowania wielowymiarowych wektorów o wartościach ciągłych w stanie wprowadza się do każdego stanu modelu HMM funkcję gęstości prawdopodobieństwa (*pdf*) (*probability density function*). Najczęściej jako funkcję gęstości w stanie wykorzystuje się kombinację liniową wielowymiarowych rozkładów normalnych z diagonalną macierzą kowariancji.

Parametry modelu zdefiniowane są w identyczny sposób jak dla modelu podstawowego HMM opisanego w podrozdziale (3.3.1), różnica polega na odmiennej definicji prawdopodobieństwa wygenerowania obserwacji w stanie $\pi_{\mathbf{o}j}=P(\mathbf{o}|S=j)$. Rozkład ten jest zdefiniowany w postaci kombinacji liniowej N rozkładów normalnych $\mathcal{N}$ jak w równaniu (3.89).

$$\pi_{\mathbf{o}j} = \sum_{k=1}^{N} c_{jk}\pi_{\mathbf{o}jk}, \qquad 1 \le j \le N$$
$$\pi_{\mathbf{o}jk} = \mathcal{N}(\mathbf{o}, \boldsymbol{\mu}_{jk}, \boldsymbol{U}_{jk}) \tag{3.89}$$

Na rysunku (3.4) pokazana jest graficzna reprezentacja pojedynczego stanu w modelu HMM z ciągłymi rozkładami prawdopodobieństw w stanach.



Rysunek 3.4. Graficzna interpretacja stanu modelu z ciągłymi rozkładami w stanach CDHMM.

Rozkład normalny $\mathcal{N}$ jest opisany równaniem (3.90), jest to m wymiarowy rozkład z wektorem średniej $\boldsymbol{\mu}_{jk}$ oraz macierzą kowariancji $\boldsymbol{U}_{jk}$. Parametry rozkładu są określone dla każdego z N elementów sumy (3.89) w każdym z m stanów. Składowe wektora obserwacji przyjmują wartości rzeczywiste $\mathbf{o}\in\Re^D$.

$$\mathcal{N}(\mathbf{o}, \boldsymbol{\mu}_{jk}, \boldsymbol{U}_{jk}) = \frac{1}{(2\pi)^{D/2}\,|\boldsymbol{U}_{jk}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{o}-\boldsymbol{\mu}_{jk})^T \boldsymbol{U}_{jk}^{-1}(\mathbf{o}-\boldsymbol{\mu}_{jk})\right] =$$
$$= \frac{1}{(2\pi)^{D/2}\prod_{d=1}^{D} u_{jkd}^{1/2}} \exp\left[-\frac{1}{2}\sum_{d=1}^{D}\frac{(o_d-\mu_{jkd})^2}{u_{jkd}^2}\right], \tag{3.90}$$

Zakłada się, że parametry rozkładu są nie-skorelowane, stąd macierz $\boldsymbol{U}_{jk}$ jest macierzą diagonalną, dla której bardzo ławo można wyznaczyć $|\boldsymbol{U}_{jk}|$ oraz macierz odwrotną $\boldsymbol{U}_{jk}^{-1}$ jak w zależnościach (3.91).

$$\boldsymbol{U}_{jk}=\begin{bmatrix} u_{jk1}^2 & & & 0 \\ & u_{jk2}^2 & & \\ & & \ddots & \\ 0 & & & u_{jkD}^2 \end{bmatrix},\quad |\boldsymbol{U}_{jk}|=\prod_{d=1}^{D} u_{jkd},\quad \boldsymbol{U}_{jk}^{-1}=\begin{bmatrix} \frac{1}{u_{jk1}^2} & & & 0 \\ & \frac{1}{u_{jk2}^2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{u_{jkD}^2} \end{bmatrix} \tag{3.91}$$

$$\boldsymbol{\mu}_{jk}^{T}=[\mu_{jk1}\quad \mu_{jk2}\quad \ldots\quad \mu_{jkD}]$$

Funkcja rozkładu prawdopodobieństwa $\pi_{\mathbf{o}j}$ oraz współczynniki c_{jk} z równania (3.89) spełniają własności (3.92).

$$\begin{aligned}\int_{-\infty}^{\infty}\pi_{\mathbf{o}j}d\mathbf{o}&=\int_{-\infty}^{\infty}\sum_{k=1}^{N}c_{jk}\mathcal{N}(\mathbf{o},\boldsymbol{\mu}_{jk},\boldsymbol{U}_{jk})d\mathbf{o}\\&=\sum_{k=1}^{N}c_{jk}\int_{-\infty}^{\infty}\mathcal{N}(\mathbf{o},\boldsymbol{\mu}_{jk},\boldsymbol{U}_{jk})d\mathbf{o}=\sum_{k=1}^{N}c_{jk}=\mathbf{1},\qquad 1\le j\le N\end{aligned} \tag{3.92}$$

Model HMM z ciągłymi rozkładami w stanach może pracować jako generator sekwencji obserwacji **O**. W wyniku pracy modelu przez T chwil czasowych zostanie wygenerowana sekwencja obserwacji $\mathbf{O}=(\mathbf{o}_1,\ldots,\mathbf{o}_T)$. Składowe sekwencji obserwacji $\mathbf{o}_t$ t=1,…,T są to D wymiarowe wektory $\mathbf{o}_t\in\mathcal{R}^D$.

Algorytm wyznaczania prawdopodobieństwa oraz algorytm Viterbiego jest identyczny jak dla modelu podstawowego, jedyna różnica polega na sposobie wyznaczania prawdopodobieństwa $\pi_{\mathbf{o}j}$. W alternatywnym algorytmie Viterbiego konieczne jest wyznaczanie logarytmów $\pi_{\mathbf{o}j}$.

3.5.1. Algorytm Bauma-Welcha estymacji parametrów modelu HMM z rozkładami ciągłymi

Podobnie jak dla modelu z rozkładami dyskretnymi zależności na estymator Bauma-Welcha wyznacza się przy pomocy algorytmu (EM). Parametry modelu wyznacza się przez maksymalizację funkcji pomocniczej $Q(M|\widehat{M})$ opisanej zależnością (3.93).

$$Q(M,\widehat{M})=\sum_{S_1,S_2,\ldots,S_T,l_1,l_2,\ldots,l_T}P(\mathbf{O},S,l/M)\log P(\mathbf{O},S,l/\widehat{M}) \tag{3.93}$$

W przypadku rozkładów ciągłych została wprowadzona modyfikacja do ścieżki tj. poza stanami S_i=1,2,…,m w ścieżce uwzględnione są poszczególne podstany l_i=1,2,…,N związane z wielokrotnymi rozkładami normalnymi w stanie.

W celu przeprowadzenia maksymalizacji funkcji Q elementy sumy (3.93) zapisuje się jak w zależności (3.94).

$$\begin{aligned}P(\mathbf{O},S,l/M)&=\delta_{S_1}c_{S_1l_1}\mathcal{N}(\mathbf{o}_1,\boldsymbol{\mu}_{S_1l_1},\boldsymbol{U}_{S_1l_1})\prod_{t=1}^{T-1}\gamma_{S_tS_{t+1}}c_{S_{t+1}l_{t+1}}\mathcal{N}(\mathbf{o}_{t+1},\boldsymbol{\mu}_{S_tl_{t+1}},\boldsymbol{U}_{S_tl_{t+1}})\\\log P(\mathbf{O},S,l/\widehat{M})&=\log\hat{\delta}_{S_1}+\sum_{t=1}^{T-1}\log\hat{\gamma}_{S_tS_{t+1}}+\sum_{t=1}^{T}\log\widehat{c}_{S_tl_t}+\sum_{t=1}^{T}\log\mathcal{N}(\mathbf{o}_t,\widehat{\boldsymbol{\mu}}_{S_tl_t},\widehat{\boldsymbol{U}}_{S_tl_t})\end{aligned} \tag{3.94}$$

Po wprowadzeniu zależności (3.94) funkcje $Q(M|\widehat{M})$ można zapisać jak w równaniu (3.95).

$$
\begin{aligned}
&Q(M/\widehat{M}) = \\
&\sum_{S_1,S_2,\ldots,S_T,l_1,l_2,\ldots,l_t} \Big\{ P(\mathbf{O},S,l/M)\log\widehat{\delta}_{S_1} + P(\mathbf{O},S,l/M)\sum_{t=1}^{T-1}\log\widehat{\gamma}_{S_tS_{t+1}} + \\
&\quad + P(\mathbf{O},S,l/M)\sum_{t=1}^{T}\log c_{S_tl_t} + P(\mathbf{O},S,l/M)\sum_{t=1}^{T}\log\mathcal{N}(\mathbf{o}_t,\boldsymbol{\mu}_{S_tl_t},\boldsymbol{U}_{S_tl_t}) \Big\} = \\
&= \sum_{i=1}^{m} P(\mathbf{O},S_1=i/M)\log\widehat{\delta}_i + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{t=1}^{T-1} P(\mathbf{O},S_t=i,S_{t+1}=j/M)\log\widehat{\gamma}_{ij} + \\
&\quad + \sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{t=1}^{T} P(\mathbf{O},S_t=i,l_t=k/M)\log\widehat{c}_{ik} + \sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{t=1}^{T} P(\mathbf{O},S_t=i,l_t=k/M)\log\mathcal{N}(\mathbf{o}_t,\widehat{\boldsymbol{\mu}}_{ik},\widehat{\boldsymbol{U}}_{ik})
\end{aligned}
\tag{3.95}
$$

Funkcję $Q(M|\widehat{M})$ została wyrażona w postaci sumy czterech elementów. Ponieważ każdy z wyszczególnionych elementów sumy jest funkcją innych parametrów nowego modelu $\widehat{M}$, dlatego można maksymalizować oddzielnie każdą składową podobnie jak dla modelu podstawowego. Trzy pierwsze elementy funkcji Q mają postać (3.61) podobnie jak dla modelu procesu Markowa, dlatego korzystając z metody mnożników Lagrange'a można wyznaczyć estymator parametrów δ_i, γ_{ij}, c_{ik}, podany w zależnościach (3.96).

$$
\begin{aligned}
\widehat{\delta}_i &= \frac{P(\mathbf{O},S_1=i)}{\sum_{j=1}^{m}P(\mathbf{O},S_1=i)} = \frac{P(\mathbf{O},S_1=i)}{P(\mathbf{O})} = \frac{\alpha_1(i)\beta_1(i)}{\sum_{j=1}^{m}\alpha_T(i)} = \chi_1(i), \\
\widehat{\gamma}_{ij} &= \frac{\sum_{t=1}^{T-1}P(\mathbf{O},S_t=i,S_{t+1}=j)}{\sum_{t=1}^{T-1}P(\mathbf{O},S_t=i)} = \frac{\sum_{t=1}^{T-1}\alpha_t(i)\gamma_{ij}\pi_{\mathbf{o}_t,i}\beta_{t+1}(j)}{\sum_{t=1}^{T-1}\alpha_t(i)\beta_t(i)} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\chi_t(i)} \\
\widehat{c}_{ik} &= \frac{\sum_{t=1}^{T}P(\mathbf{O},S_t=i,l_t=k\,|\,M)}{\sum_{k=1}^{m}\sum_{t=1}^{T}P(\mathbf{O},S_t=i,l_t=k\,|\,M)} = c_{ik}\,\frac{\sum_{t=1}^{T}\pi_{\mathbf{o}_t ik}v_t(i)}{\sum_{k=1}^{m}c_{ik}\sum_{t=1}^{T}\pi_{\mathbf{o}_t ik}v_t(i)}
\end{aligned}
\tag{3.96}
$$

Sposób wyznaczenia parametru $v_t(i)$ został pokazany w równaniach (3.97).

$$
\begin{aligned}
v_1(i) &= P(\mathbf{O},S_1=i,l_1=k/M)/c_{ik} = \delta_1\beta_1(i) \\
v_t(i) &= P(\mathbf{O},S_t=i,l_t=k/M)/c_{ik} = \beta_t(i)\sum_{j=1}^{m}\gamma_{ji}\alpha_{t-1}(j), \quad t=2,\ldots,T
\end{aligned}
\tag{3.97}
$$

Parametry $\widehat{\boldsymbol{\mu}}_{ik}$, $\widehat{\boldsymbol{U}}_{ik}$ rozkładów w stanach można obliczyć maksymalizując czwartą część funkcji Q różniczkując ją i przyrównując składowe do zera jak w równaniu (3.98).

$$\sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{t=1}^{T} v_t(i) c_{ik} \mathcal{N}(\mathbf{o}_t, \boldsymbol{\mu}_{ik}, \boldsymbol{U}_{ik}) \frac{\partial}{\partial \hat{\boldsymbol{\mu}}_{ik}} \log \mathcal{N}(\mathbf{o}_t, \hat{\boldsymbol{\mu}}_{ik}, \hat{\boldsymbol{U}}_{ik}) = 0$$
$$\sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{t=1}^{T} v_t(i) c_{ik} \mathcal{N}(\mathbf{o}_t, \boldsymbol{\mu}_{ik}, \boldsymbol{U}_{ik}) \frac{\partial}{\partial \hat{\boldsymbol{U}}_{ik}} \log \mathcal{N}(\mathbf{o}_t, \hat{\boldsymbol{\mu}}_{ik}, \hat{\boldsymbol{U}}_{ik}) = 0 \tag{3.98}$$

Pochodne z równań (3.98) można zapisać w postaci (3.99).

$$\frac{\partial}{\partial \hat{\boldsymbol{\mu}}_{ik}} \log \mathcal{N}(\mathbf{o}_t, \hat{\boldsymbol{\mu}}_{ik}, \hat{\boldsymbol{U}}_{ik}) = \hat{\boldsymbol{U}}_{ik}^{-1}(\mathbf{o}_t - \hat{\boldsymbol{\mu}}_{ik})$$
$$\frac{\partial}{\partial \hat{\boldsymbol{U}}_{ik}} \log \mathcal{N}(\mathbf{o}_t, \hat{\boldsymbol{\mu}}_{ik}, \hat{\boldsymbol{U}}_{ik}) = \hat{\boldsymbol{U}}_{ik}^{-3/2}[\hat{\boldsymbol{U}}_{ik} - (\mathbf{o}_t - \hat{\boldsymbol{\mu}}_{ik})(\mathbf{o}_t - \hat{\boldsymbol{\mu}}_{ik})^T] \tag{3.99}$$

Po wprowadzeniu (3.99) do (3.98) uzyskuje się zależności (3.100).

$$\sum_{i=1}^{m}\sum_{k=1}^{N} \hat{\boldsymbol{U}}_{ik}^{-1} \sum_{t=1}^{T} v_t(i) c_{ik} \mathcal{N}(\mathbf{o}_t, \boldsymbol{\mu}_{ik}, \boldsymbol{U}_{ik})(\mathbf{o}_t - \hat{\boldsymbol{\mu}}_{ik}) = 0$$
$$\sum_{i=1}^{m}\sum_{k=1}^{N} \hat{\boldsymbol{U}}_{ik}^{-3/2} \sum_{t=1}^{T} v_t(i) c_{ik} \mathcal{N}(\mathbf{o}_t, \boldsymbol{\mu}_{ik}, \boldsymbol{U}_{ik})[\hat{\boldsymbol{U}}_{ik} - (\mathbf{o}_t - \hat{\boldsymbol{\mu}}_{ik})(\mathbf{o}_t - \hat{\boldsymbol{\mu}}_{ik})^T] = 0 \tag{3.100}$$

Z równań (3.100) można wyznaczyć estymator parametrów rozkładów w stanach $\hat{\boldsymbol{\mu}}_{ik}$, $\hat{\boldsymbol{U}}_{ik}$, podany w równaniach (3.101).

$$\hat{\boldsymbol{\mu}}_{ik} = \frac{\sum_{t=1}^{T} \pi_{\mathbf{o}_t ik} v_t(i) \mathbf{o}_t}{\sum_{t=1}^{T} \pi_{\mathbf{o}_t ik} v_t(i)}$$
$$\hat{\boldsymbol{U}}_{ik} = \frac{\sum_{t=1}^{T} \pi_{\mathbf{o}_t ik} v_t(i)(\mathbf{o}_t - \hat{\boldsymbol{\mu}}_{ik})(\mathbf{o}_t - \hat{\boldsymbol{\mu}}_{ik})^T}{\sum_{t=1}^{T} \pi_{\mathbf{o}_t ik} v_t(i)} \tag{3.101}$$

3.5.2. Wyznaczenie rozkładu „*prior*" dla rozkładu normalnego z diagonalną macierzą kowariancji

W estymatorze Bayesa wykorzystuje się rozkład „*prior*" modelu. Ponieważ w literaturze nie zostało wyjaśnione, w jaki sposób został wyznaczony rozkład „*prior*" w podrozdziale tym została podana metoda wyznaczenia tego rozkładu dla rozkładu normalnego z diagonalną macierzą kowariancji.

Należy założyć, że są dwa zbiory wektorów obserwacji, $\{\boldsymbol{x}_t\}$ o liczności τ, oraz $\{\boldsymbol{y}_t\}$ o liczności T. Można postawić pytanie, w jaki sposób dobrać parametry $\boldsymbol{\mu}$, $\boldsymbol{U}$ rozkładu normalnego $\mathcal{N}$, żeby maksymalizować prawdopodobieństwo obserwacji podane równaniem (3.102), oraz w jaki sposób dobrać rozkład „*prior*" dla rozkładu normalnego. Rozpatrywany rozkład normalny $\mathcal{N}$ jest to rozkład z diagonalną macierzą kowariancji.

$$P(\boldsymbol{x}_t, \boldsymbol{y}_t) = \prod_{t=1}^{\tau} \mathcal{N}(\boldsymbol{x}_t, \boldsymbol{\mu}, \boldsymbol{U}) \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{y}_t, \boldsymbol{\mu}, \boldsymbol{U}) \tag{3.102}$$

Parametry rozkładu można dobrać rozwiązując układ równań (3.103) powstały w wyniku przyrównania pochodnej logarytmu równania (3.102) do zera. Pochodne (3.103) zostały policzone dla diagonalnej macierzy kowariancji $\boldsymbol{U}$. Nie jest konieczne sprawdzanie znaku drugiej pochodnej, ponieważ rozkład normalny ma tylko jedno maksimum.

$$\begin{aligned}
&\frac{\partial \log P(\boldsymbol{x}_t, \boldsymbol{y}_t)}{\partial \boldsymbol{\mu}} = \boldsymbol{U}^{-1} \sum_{t=1}^{\tau} (\boldsymbol{x}_t - \boldsymbol{\mu}) + \boldsymbol{U}^{-1} \sum_{t=1}^{T} (\boldsymbol{y}_t - \boldsymbol{\mu}) = 0 \\
&\frac{\partial \log P(\boldsymbol{x}_t, \boldsymbol{y}_t)}{\partial \boldsymbol{U}} = -\boldsymbol{U}^{-3/2} \sum_{t=1}^{\tau} [\boldsymbol{U} - (\boldsymbol{x}_t - \boldsymbol{\mu})(\boldsymbol{x}_t - \boldsymbol{\mu})^T] - \boldsymbol{U}^{-3/2} \sum_{t=1}^{T} [\boldsymbol{U} - (\boldsymbol{y}_t - \boldsymbol{\mu})(\boldsymbol{y}_t - \boldsymbol{\mu})^T] = 0
\end{aligned} \tag{3.103}$$

Po rozwiązaniu (3.103) uzyskuje się zależności na estymator parametrów rozkładu podany w zależnościach (3.104).

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{t=1}^{\tau} \boldsymbol{x}_t + \sum_{t=1}^{T} \boldsymbol{y}_t}{\tau + T}, \qquad \hat{\boldsymbol{U}} = \frac{\sum_{t=1}^{\tau} (\boldsymbol{x}_t - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}})^T + \sum_{t=1}^{T} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}})^T}{\tau + T} \tag{3.104}$$

Jeżeli zostaną uwzględnione tylko obserwacje $\{\boldsymbol{x}_t\}$, wówczas estymator przyjmuje postać podaną w równaniu (3.105).

$$\boldsymbol{m} = \frac{\sum_{t=1}^{\tau} \boldsymbol{x}_t}{\tau}, \qquad \boldsymbol{U} = \frac{\sum_{t=1}^{\tau} (\boldsymbol{x}_t - \boldsymbol{m})(\boldsymbol{x}_t - \boldsymbol{m})^T}{\tau} \tag{3.105}$$

Po wprowadzeniu (3.105) do (3.104) zależności na wektor średniej $\hat{\boldsymbol{\mu}}$ oraz macierz kowariancji $\hat{\boldsymbol{U}}$ przyjmują postać podaną w równaniach (3.106).

$$\begin{aligned}
\hat{\boldsymbol{\mu}} &= \frac{\boldsymbol{m}\tau + \sum_{t=1}^{T} \boldsymbol{y}_t}{\tau + T} \\
\hat{\boldsymbol{U}} &= \frac{\sum_{t=1}^{\tau} \boldsymbol{x}_t \boldsymbol{x}_t^T - 2\sum_{t=1}^{\tau} \boldsymbol{x}_t \hat{\boldsymbol{\mu}}^T + \sum_{t=1}^{\tau} \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^T + \sum_{t=1}^{T} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}})^T}{\tau + T} = \\
&= \frac{\sum_{t=1}^{\tau} \boldsymbol{x}_t \boldsymbol{x}_t^T - \sum_{t=1}^{\tau} \boldsymbol{m}\boldsymbol{m}^T + \sum_{t=1}^{\tau} \boldsymbol{m}\boldsymbol{m}^T - 2\sum_{t=1}^{\tau} \boldsymbol{x}_t \hat{\boldsymbol{\mu}}^T + \sum_{t=1}^{\tau} \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^T + \sum_{t=1}^{T} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}})^T}{\tau + T} = \\
&= \frac{\sum_{t=1}^{\tau} (\boldsymbol{x}_t \boldsymbol{x}_t^T - \boldsymbol{m}\boldsymbol{m}^T) + \boldsymbol{m}\boldsymbol{m}^T\tau - 2\boldsymbol{m}\hat{\boldsymbol{\mu}}^T\tau + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T\tau + \sum_{t=1}^{T} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}})^T}{\tau + T} = \\
&= \frac{\sum_{t=1}^{\tau} (\boldsymbol{x}_t \boldsymbol{x}_t^T - \boldsymbol{m}\boldsymbol{m}^T) + \tau(\boldsymbol{m} - \hat{\boldsymbol{\mu}})(\boldsymbol{m} - \hat{\boldsymbol{\mu}})^T + \sum_{t=1}^{T} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}})^T}{\tau + T} = \\
&= \frac{2\boldsymbol{\beta} + \tau(\boldsymbol{m} - \hat{\boldsymbol{\mu}})(\boldsymbol{m} - \hat{\boldsymbol{\mu}})^T + \sum_{t=1}^{T} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}})^T}{\tau + T}
\end{aligned} \tag{3.106}$$

gdzie: $\boldsymbol{\beta}$ - jest to macierz, dla której wyznaczane są wartości na diagonalnej. Po przeprowadzeniu elementarnych przekształceń (3.106) uzyskuje się zależności (3.107).

$$\begin{cases} \hat{\boldsymbol{\mu}}\tau + \sum_{t=1}^{T}\hat{\boldsymbol{\mu}} = \boldsymbol{m}\tau + \sum_{t=1}^{T}\boldsymbol{y}_t \\ \hat{\boldsymbol{U}}\tau + \sum_{t=1}^{T}\hat{\boldsymbol{U}} = 2\boldsymbol{\beta} + \tau(\boldsymbol{m}-\hat{\boldsymbol{\mu}})(\boldsymbol{m}-\hat{\boldsymbol{\mu}})^T + \sum_{t=1}^{T}(\boldsymbol{y}_t-\hat{\boldsymbol{\mu}})(\boldsymbol{y}_t-\hat{\boldsymbol{\mu}})^T \end{cases}$$
$$\begin{cases} \hat{\boldsymbol{U}}^{-1}\sum_{t=1}^{T}(\boldsymbol{y}_t-\hat{\boldsymbol{\mu}}) + \hat{\boldsymbol{U}}^{-1}\tau(\boldsymbol{m}-\hat{\boldsymbol{\mu}}) = 0 \\ -\hat{\boldsymbol{U}}^{-3/2}\sum_{t=1}^{T}[\hat{\boldsymbol{U}} - (\boldsymbol{y}_t-\hat{\boldsymbol{\mu}})(\boldsymbol{y}_t-\hat{\boldsymbol{\mu}})^T] - \hat{\boldsymbol{U}}^{-3/2}[\hat{\boldsymbol{U}}\tau - \tau(\boldsymbol{m}-\hat{\boldsymbol{\mu}})(\boldsymbol{m}-\hat{\boldsymbol{\mu}})^T - 2\boldsymbol{\beta}] = 0 \end{cases} \quad (3.107)$$

Korzystając z zależności (3.107) można łatwo zauważyć, że rozkład „*prior*" powinien spełniać układ równań (3.108).

$$\begin{aligned} \frac{\partial \log g(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{U}})}{\partial \hat{\boldsymbol{\mu}}} &= \tau\boldsymbol{U}^{-1}(\boldsymbol{m}-\hat{\boldsymbol{\mu}}) \\ \frac{\partial \log g(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{U}})}{\partial \hat{\boldsymbol{U}}} &= -\hat{\boldsymbol{U}}^{-3/2}[\boldsymbol{U}\tau - \tau(\boldsymbol{m}-\hat{\boldsymbol{\mu}})(\boldsymbol{m}-\hat{\boldsymbol{\mu}})^T - 2\boldsymbol{\beta}] \end{aligned} \quad (3.108)$$

Rozkład warunkowy rozkładu normalnego-Gamma podany w równaniu (3.109) dla parametru 2α-1=τ spełnia własności (3.108).

$$g(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{U}}) = K\frac{1}{\left|\hat{\boldsymbol{U}}\right|^{\tau/2}}\exp\left(-\frac{\tau}{2}(\boldsymbol{m}-\hat{\boldsymbol{\mu}})^T\hat{\boldsymbol{U}}^{-1}(\boldsymbol{m}-\hat{\boldsymbol{\mu}}) - \frac{1}{2}\mathbf{1}^T\boldsymbol{\beta}\hat{\boldsymbol{U}}^{-1}\mathbf{1}\right) \quad (3.109)$$

Parametry rozkładu „*prior*" po uwzględnieniu nowych danych można wyznaczyć z zależności (3.110). Parametr τ oraz składowe macierzy $\boldsymbol{\beta}$ dla początkowego rozkładu bez informacji na temat modelu powinny być równe zero.

$$\begin{aligned} \hat{\boldsymbol{m}} &= \hat{\boldsymbol{\mu}} \\ \hat{\tau} &= \tau + T \\ \hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + \frac{\tau(\boldsymbol{m}-\hat{\boldsymbol{\mu}})(\boldsymbol{m}-\hat{\boldsymbol{\mu}})^T}{2} + \frac{\sum_{t=1}^{T}(\boldsymbol{y}_t\boldsymbol{y}_t^T - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T)}{2} \end{aligned} \quad (3.110)$$

W literaturze parametr τ występujący w mianowniku przed exponentą w równaniu (3.109) jest zastępowany jest przez 2α-1, prawdopodobnie ze względu na analogie z rozkładem Dirichleta, w którym również odejmowana jest wartość 1, w przypadku rozkładów dyskretnych nie ma to wpływu na ilość zapamiętywanych parametrów, natomiast po wprowadzeniu α w równaniu (3.109) konieczne jest zapamiętywanie czwartego parametru.

Zależności (3.104), (3.105) są to zależności estymatora parametrów rozkładu normalnego opartego na kryterium (ML), zależności (3.106) są to zależności estymatora opartego na kryterium (MAP), natomiast zależności (3.110) pozwalają na wyznaczać nowych parametrów rozkładu „*prior*".

Wniosek 1.

Nie jest konieczna znajomość jawnej postaci rozkładu „*prior*" (3.109) w celu wyznaczenia i zastosowania estymatora opartego na kryterium (MAP). W przypadku modelu HMM w celu wyznaczenia estymatora opartego na kryterium (MAP) konieczna jest jedynie znajomość pochodnych (3.108) tego rozkładu.

Wyznaczenia zależności na parametr β rozkładu „*posterior*".

Korzystając z zależności (3.106) wartość parametru $\boldsymbol{\beta}$ oraz $\hat{\boldsymbol{\beta}}$ można zapisać jak w (3.111).

$$\begin{aligned} 2\boldsymbol{\beta} &= \sum_{t=1}^{\tau}(\boldsymbol{x}_t\boldsymbol{x}_t^{\,T} - \boldsymbol{m}\boldsymbol{m}^T) \\ 2\hat{\boldsymbol{\beta}} &= \sum_{t=1}^{\tau}(\boldsymbol{x}_t\boldsymbol{x}_t^{\,T} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T) + \sum_{t=1}^{T}(\boldsymbol{y}_t\boldsymbol{y}_t^{\,T} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T) \end{aligned} \tag{3.111}$$

Po elementarnych przekształceniach uzyskuje się zależność (3.112) na parametr $\hat{\boldsymbol{\beta}}$ w funkcji nowych danych $\boldsymbol{y}_t$ oraz znanych parametrów $\boldsymbol{\beta}$, $\boldsymbol{m}$, τ rozkładu „*prior*".

$$\begin{aligned} 2\hat{\boldsymbol{\beta}} &= \sum_{t=1}^{\tau}(\boldsymbol{x}_t\boldsymbol{x}_t^{\,T} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T) + \sum_{t=1}^{\tau}(-\boldsymbol{m}\boldsymbol{m}^T) + \sum_{t=1}^{\tau}(\boldsymbol{m}\boldsymbol{m}^T) + \sum_{t=1}^{T}(\boldsymbol{y}_t\boldsymbol{y}_t^{\,T} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T) = \\ &= \sum_{t=1}^{\tau}(\boldsymbol{x}_t\boldsymbol{x}_t^{\,T} - \boldsymbol{m}\boldsymbol{m}^T) + \sum_{t=1}^{\tau}(\boldsymbol{m}\boldsymbol{m}^T - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T) + \sum_{t=1}^{T}(\boldsymbol{y}_t\boldsymbol{y}_t^{\,T} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T) = \\ &= 2\boldsymbol{\beta} + \tau(\boldsymbol{m}\boldsymbol{m}^T - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T) + \sum_{t=1}^{T}(\boldsymbol{y}_t\boldsymbol{y}_t^{\,T} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T) \end{aligned} \tag{3.112}$$

Zależność na parametr $\hat{\boldsymbol{\beta}}$ przyjmuje postać podaną w (3.110).

W wyniku zastosowania omówionej metody został wyznaczony rozkład „*prior*" dla rozkładu normalnego z diagonalną macierzą kowariancji w postaci rozkładu warunkowego rozkładu noramlnego-Gamma.

W wyniku wprowadzenia rozkładu warunkowego została zredukowana ilość zapamiętywanych parametrów wykorzystywanych w podstawowym algorytmie optymalizacji modeli HMM tj. algorytmie Bayesa oraz zostały zredukowane niepotrzebne działania w porównaniu z pracami [9], [73]-[74], [76].

W pracy [9] w procedurach estymatora zależność (5.42) str. 223 oraz w procedurach modyfikacji parametrów rozkładu „*prior*" zależność (5.49) str. 224 wprowadzony jest niepotrzebny dodatkowy parametr α, parametry początkowe rozkładu opisane są zależnością (5.59) str. 226. Powodem wprowadzenia parametru na pewno nie jest wyznaczenie estymatora nie obciążonego, ponieważ łatwo zauważyć, że parametry α oraz τ są powiązane, co zostało pokazane w poniższej zależnościach (3.113).

$$\begin{cases} \alpha_i = \alpha_{i-1} + \dfrac{1}{2}z_{i-1}, \alpha_1 = 1/2 \\ \tau_i = \tau_{i-1} + z_{i-1}, \quad \tau_1 = 0 \end{cases} \Rightarrow \begin{cases} \alpha_i = \dfrac{1}{2}(1 + \sum_{k=1}^{i-1} z_k) \\ \tau_i = \sum_{k=1}^{i-1} z_k \end{cases} \Rightarrow \alpha_i = \frac{1}{2}(1+\tau_i) \Rightarrow \begin{matrix} 2\alpha_i - 1 = \tau_i \\ i = 2,3,... \end{matrix} \tag{3.113}$$

3.5.3. Algorytm Bayesa estymacji parametrów modelu HMM z rozkładami ciągłymi

Podobnie jak dla modelu z dyskretnymi rozkładami prawdopodobieństw do doboru parametrów modelu wykorzystuje się algorytm (EM). Jako rozkład „*prior*" w kryterium (MAP) wprowadzony się rozkład Dirichleta jako rozkład „*prior*" dla rozkładów dyskretnych podany równaniem (3.114), oraz rozkład „*prior*" dla rozkładu normalnego $\mathcal{N}$ wyznaczony w podrozdziale (3.5.2) podany równaniem (3.115).

$$g(\hat{\delta},\hat{\gamma},\hat{c}) = K\prod_{i=1}^{m}\hat{\delta}_i^{\eta_i}\prod_{i=1}^{m}\prod_{j=1}^{m}\hat{\gamma}_{ij}^{\eta_{ij}}\prod_{i=1}^{m}\prod_{k=1}^{N}\hat{c}_{ik}^{\nu_{ik}} \tag{3.114}$$

$$g(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{U}}) = G\prod_{i=1}^{m}\prod_{k=1}^{N}\frac{1}{\left|\hat{\boldsymbol{U}}_{ik}\right|^{\tau_{ik}/2}}\exp(-\frac{\tau_{ik}}{2}(\boldsymbol{m}_{ik}-\hat{\boldsymbol{\mu}}_{ik})^T\hat{\boldsymbol{U}}_{ik}^{-1}(\boldsymbol{m}_{ik}-\hat{\boldsymbol{\mu}}_{ik})-\frac{1}{2}\mathbf{1}^T\boldsymbol{\beta}_{ik}\hat{\boldsymbol{U}}_{ik}^{-1}\mathbf{1}) \tag{3.115}$$

Rozkład „*prior*" modelu CDHMM $g(\hat{M})$ można zapisać jak w równaniu (3.116).

$$g(\hat{M}) = g(\hat{\delta},\hat{\gamma},\hat{c})g(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{U}}) \tag{3.116}$$

W celu doboru parametrów modelu wykorzystuje się algorytm (EM). Zmodyfikowana funkcja pomocnicza $R(M|\hat{M})$ podana jest w równaniu (3.117).

$$\begin{aligned} R(M,\hat{M}) &= \frac{1}{P(\mathbf{O}\,|\,M)}Q(M,\hat{M}) + \log g(\hat{M}) = \\ &= \frac{1}{P(\mathbf{O}\,|\,M)}\sum_{S_1,S_2,\ldots,S_T,l_1,l_2,\ldots,l_T}P(\mathbf{O},S,l\,|\,M)\log P(\mathbf{O},S,l\,|\,\hat{\mathrm{M}}) + \log g(\hat{M}) \end{aligned} \tag{3.117}$$

Zależności na estymator parametrów wyznacza się przez maksymalizację wprowadzonej funkcji pomocniczej $R(M|\hat{M})$. Elementy funkcji pomocniczej można zapisać jak w zależnościach (3.118).

$$\begin{aligned} P(\mathbf{O},S,l\,|\,M) &= \delta_{S_1}c_{S_1l_1}\pi_{\mathbf{o}_1S_1l_1}\prod_{t=1}^{T-1}\gamma_{S_tS_{t+1}}c_{S_{t+1}l_{t+1}}\pi_{\mathbf{o}_tS_{t+1}l_{t+1}} \\ \log P(\mathbf{O},S,l\,|\,\hat{M}) &= \log\hat{\delta}_{S_1} + \sum_{t=1}^{T-1}\log\hat{\gamma}_{S_tS_{t+1}} + \sum_{t=1}^{T}\log\hat{c}_{S_tl_t} + \sum_{t=1}^{T}\log\hat{\pi}_{\mathbf{o}_tS_tl_t} \\ \pi_{\mathbf{o}ik} &= \mathcal{N}(\mathbf{o},\boldsymbol{\mu}_{ik},\boldsymbol{U}_{ik}),\quad i=1,2,\ldots,m,\quad k=1,2,\ldots,N \end{aligned} \tag{3.118}$$

Po przekształceniu funkcji pomocniczej R uzyskuje się równanie (3.119).

$$\begin{aligned} &R(M\,|\,\hat{M}) = \\ &\frac{1}{P(\mathbf{O}\,|\,M)}\sum_{S_1,S_2,\ldots,S_T,l_1,l_2,\ldots,l_t}\Bigg\{P(\mathbf{O},S,l\,|\,M)\log\hat{\delta}_{S_1} + P(\mathbf{O},S,l\,|\,M)\sum_{t=1}^{T-1}\log\hat{\gamma}_{S_tS_{t+1}} + \\ &+P(\mathbf{O},S,l\,|\,M)\sum_{t=1}^{T}\log\hat{c}_{S_tl_t} + P(\mathbf{O},S,l\,|\,M)\sum_{t=1}^{T}\log\hat{\pi}_{\mathbf{o}_tS_tl_t}\Bigg\} + \log g(\hat{M}) = \end{aligned} \tag{3.119}$$

$$= \frac{1}{P(\mathbf{O}\,|\,M)} \sum_{i=1}^{m} P(\mathbf{O}, S_1 = i\,|\,M) \log \widehat{\delta}_i + \frac{1}{P(\mathbf{O}\,|\,M)} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{t=1}^{T-1} P(\mathbf{O}, S_t = i, S_{t+1} = j\,|\,M) \log \widehat{\gamma}_{ij} +$$
$$+ \frac{1}{P(\mathbf{O}\,|\,M)} \sum_{i=1}^{k} \sum_{k=1}^{N} \sum_{t=1}^{T} P(\mathbf{O}, S_t = i, l_t = k\,|\,M) \log \widehat{c}_{ik} + \frac{1}{P(\mathbf{O}\,|\,M)} \sum_{i=1}^{m} \sum_{k=1}^{N} \sum_{t=1}^{T} P(\mathbf{O}, S_t = i, l_t = k\,|\,M) \log \widehat{\pi}_{\mathbf{o}_t S_t l_t} +$$
$$+ \sum_{i=1}^{m} \eta_i \log \widehat{\delta}_i + \sum_{i=1}^{m} \sum_{j=1}^{m} \eta_{ij} \log \widehat{\gamma}_{ij} +$$
$$+ \sum_{i=1}^{m} \sum_{k=1}^{N} \nu_{ik} \log \widehat{c}_{ik} + \sum_{i=1}^{m} \sum_{k=1}^{N} \log g(\widehat{\boldsymbol{\mu}}_{\boldsymbol{ik}}, \widehat{\boldsymbol{U}}_{\boldsymbol{ik}}) + \log(KG) \tag{3.119}$$

Z zależności (3.119) można wyznaczyć estymator parametrów δ_i, γ_{ij}, c_{ik}, podany w zależnościach (3.120).

$$\widehat{\delta}_i = \frac{\frac{1}{P(\mathbf{O}\,|\,M)} P(\mathbf{O}, S_1 = i) + \eta_i}{\frac{1}{P(\mathbf{O}\,|\,M)} \sum_{i=1}^{m} P(\mathbf{O}, S_1 = i) + \sum_{i=1}^{m} \eta_i} = \frac{\frac{1}{P(\mathbf{O})} \alpha_1(i) \beta_1(i) + \eta_i}{1 + \sum_{i=1}^{m} \eta_i}$$

$$\widehat{\gamma}_{ij} = \frac{\frac{1}{P(\mathbf{O}\,|\,M)} \sum_{t=1}^{T-1} P(\mathbf{O}, S_t = i, S_{t+1} = j) + \eta_{ij}}{\frac{1}{P(\mathbf{O}\,|\,M)} \sum_{t=1}^{T-1} P(\mathbf{O}, S_t = i) + \sum_{j=1}^{m} \eta_{ij}} = \frac{\frac{1}{P(\mathbf{O})} \sum_{t=1}^{T-1} \alpha_t(i) \gamma_{ij} \pi_{\mathbf{o}_t i} \beta_{t+1}(j) + \eta_{ij}}{\frac{1}{P(\mathbf{O})} \sum_{t=1}^{T-1} \alpha_t(i) \beta_t(i) + \sum_{j=1}^{m} \eta_{ij}}$$

$$\widehat{c}_{ik} = \frac{\frac{1}{P(\mathbf{O}\,|\,M)} \sum_{t=1}^{T} P(\mathbf{O}, S_t = i, l_t = k\,|\,M) + \nu_{ik}}{\frac{1}{P(\mathbf{O}\,|\,M)} \sum_{k=1}^{N} \sum_{t=1}^{T} P(\mathbf{O}, S_t = i, l_t = k\,|\,M) + \sum_{k=1}^{N} \nu_{ik}} =$$
$$= c_{ik} \frac{\frac{1}{P(\mathbf{O})} \sum_{t=1}^{T} \pi_{\mathbf{o}_t ik} \nu_t(i) + \nu_{ik}}{\frac{1}{P(\mathbf{O})} \sum_{k=1}^{N} c_{ik} \sum_{t=1}^{T} \pi_{\mathbf{o}_t ik} \nu_t(i) + \sum_{k=1}^{N} \nu_{ik}} \tag{3.120}$$

Estymator parametrów $\widehat{\boldsymbol{\mu}}_{ik}$, $\widehat{\boldsymbol{U}}_{ik}$ można obliczyć rozwiązując układ równań (3.121) postały przez zróżniczkowanie czwartej części funkcji R i przyrównanie pochodnych do zera.

$$\begin{aligned} &\frac{\partial}{\partial \widehat{\boldsymbol{\mu}}_{\boldsymbol{ik}}} \left(\sum_{i=1}^{m} \sum_{k=1}^{N} \frac{1}{P(\mathbf{O}\,|\,M)} \sum_{t=1}^{T} P(\mathbf{O}, S_t = i, l_t = k\,|\,M) \pi_{\mathbf{o}_t S_t l_t} + \sum_{i=1}^{m} \sum_{k=1}^{N} \log g(\widehat{\boldsymbol{\mu}}_{\boldsymbol{ik}}, \widehat{\boldsymbol{U}}_{\boldsymbol{ik}}) \right) = 0 \\ &\frac{\partial}{\partial \widehat{\boldsymbol{U}}_{\boldsymbol{ik}}} \left(\frac{1}{P(\mathbf{O}\,|\,M)} \sum_{i=1}^{m} \sum_{k=1}^{N} \sum_{t=1}^{T} P(\mathbf{O}, S_t = i, l_t = k\,|\,M) \pi_{\mathbf{o}_t S_t l_t} + \sum_{i=1}^{m} \sum_{k=1}^{N} \log g(\widehat{\boldsymbol{\mu}}_{\boldsymbol{ik}}, \widehat{\boldsymbol{U}}_{\boldsymbol{ik}}) \right) = 0 \end{aligned} \tag{3.121}$$

Po zróżniczkowaniu (3.121) uzyskuje się zależności (3.122).

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{k=1}^{N}\hat{\boldsymbol{U}}_{ik}^{-1}[\frac{1}{P(\mathbf{O}\,|\,M)}\sum_{t=1}^{T}v_t(i)c_{ik}\pi_{\mathbf{o}_t ik}(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})+\tau_{ik}(\boldsymbol{m}_{ik}-\hat{\boldsymbol{\mu}}_{ik})]=0\\
&\sum_{i=1}^{m}\sum_{k=1}^{N}-\hat{\boldsymbol{U}}_{ik}^{-3/2}\{\frac{1}{P(\mathbf{O}\,|\,M)}\sum_{t=1}^{T}v_t(i)c_{ik}\pi_{\mathbf{o}_t ik}[\hat{\boldsymbol{U}}_{ik}-(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})^T]+\\
&\qquad+[\hat{\boldsymbol{U}}_{ik}\tau_{ik}-\tau_{ik}(\boldsymbol{m}_{ik}-\hat{\boldsymbol{\mu}}_{ik})(\boldsymbol{m}_{ik}-\hat{\boldsymbol{\mu}}_{ik})^T-2\boldsymbol{\beta}_{ik}]\}=0
\end{aligned}\tag{3.122}$$

Z zależności (3.122) można wyznaczyć estymator parametrów rozkładów w stanach $\hat{\boldsymbol{\mu}}_{ik}$, $\hat{\boldsymbol{U}}_{ik}$, podany w równaniach (3.123).

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_{ik}&=\frac{\tau_{ik}\boldsymbol{m}_{ik}+\frac{1}{P(\mathbf{O}\,|\,M)}\sum_{t=1}^{T}\pi_{\mathbf{o}_t ik}v_t(i)\mathbf{o}_t}{\tau_{ik}+\frac{1}{P(\mathbf{O}\,|\,M)}\sum_{t=1}^{T}\pi_{\mathbf{o}_t ik}v_t(i)}\\
\hat{\boldsymbol{U}}_{ik}&=\frac{2\boldsymbol{\beta}_{ik}+\tau_{ik}(\boldsymbol{m}_{ik}-\hat{\boldsymbol{\mu}}_{ik})(\boldsymbol{m}_{ik}-\hat{\boldsymbol{\mu}}_{ik})^T+\frac{1}{P(\mathbf{O}\,|\,M)}\sum_{t=1}^{T}\pi_{\mathbf{o}_t ik}v_t(i)(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})^T}{\tau_{ik}+\frac{1}{P(\mathbf{O}\,|\,M)}\sum_{t=1}^{T}\pi_{\mathbf{o}_t ik}v_t(i)}
\end{aligned}\tag{3.123}$$

Parametry rozkładu „*posterior*" modyfikuje się według zależności (3.124).

$$\begin{aligned}
\hat{\eta}_i&=\eta_i+\frac{1}{P(\mathbf{O}\,|\,M)}\alpha_1(i)\beta_1(i),\\
\hat{\eta}_{ij}&=\eta_{ij}+\frac{1}{P(\mathbf{O}\,|\,M)}\sum_{t=1}^{T-1}\alpha_t(i)\gamma_{ij}\pi_{\mathbf{o}_t i}\beta_{t+1}(j),\\
\hat{\nu}_{ki}&=\nu_{ik}+\frac{1}{P(\mathbf{O}\,|\,M)}\sum_{t=1}^{T}\pi_{\mathbf{o}_t ik}v_t(i),\\
\hat{\boldsymbol{m}}_{ik}&=\hat{\boldsymbol{\mu}}_{ik},\\
\hat{\tau}_{ik}&=\tau_{ik}+\frac{1}{P(\mathbf{O}\,|\,M)}\sum_{t=1}^{T}\pi_{\mathbf{o}_t ik}v_t(i),\\
\hat{\boldsymbol{\beta}}_{ik}&=\boldsymbol{\beta}_{ik}+\frac{\tau_{ik}(\boldsymbol{m}_{ik}-\hat{\boldsymbol{\mu}}_{ik})(\boldsymbol{m}_{ik}-\hat{\boldsymbol{\mu}}_{ik})^T}{2}+\frac{\frac{1}{P(\mathbf{O}\,|\,M)}\sum_{t=1}^{T}\pi_{\mathbf{o}_t ik}v_t(i)(\mathbf{o}_t\mathbf{o}_t^T-\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T)}{2}
\end{aligned}\tag{3.124}$$

W wyznaczonych zależnościach na estymator została wprowadzona dodatkowa wartość $P(\mathbf{O}|M)$. W wyniku tego sekwencje dające małe prawdopodobieństwo mają większy wpływu na modyfikację parametrów modelu niż sekwencje o dużych wartościach prawdopodobieństwa. Uzasadnione jest porównanie wyników uzyskanych z zastosowania algorytmu z pracy [6] oraz zaproponowanego w tym podrozdziale.

3.5.4. Algorytm segmental *k*-means estymacji parametrów modelu HMM z rozkładami ciągłymi

W algorytmie segmental *k*-means wykorzystuje się optymalną ścieżkę $S^*=S_1{}^*,S_2{}^*,\ldots,S_T{}^*$ uzyskaną dzięki algorytmowi Viterbiego. Prawdopodobieństwo na ścieżce może zostać zapisane w postaci sumy (3.125), gdzie $l=\{l_1,l_2,\ldots,l_T\}$ dotyczy podstanów modelu CDHMM.

$$P(\mathbf{O},S^*\mid M)=\sum_{l_1,l_2,\ldots,l_T}P(\mathbf{O},S^*,l\mid M) \tag{3.125}$$

Ponieważ w omawianym algorytmie w znalezionej optymalnej ścieżce S^* uwzględnione są tylko stany modelu natomiast nie są uwzględnione podstany l konieczne jest zastosowanie algorytmu (EM). W celu zastosowania algorytmu (EM) należy wprowadzić funkcję pomocniczą (3.126).

$$Q(M,\widehat{M})=\sum_{l_1,l_2,\ldots,l_T}P(\mathbf{O},S^*,l\mid M)\log P(\mathbf{O},S^*,l\mid \widehat{M}) \tag{3.126}$$

Prawdopodobieństwo na pojedynczej ścieżce $S^*=\{S_1{}^*,S_2{}^*,\ldots,S_T{}^*\}$ $l=\{l_1,l_2,\ldots,l_T\}$ można zapisać jak w równaniu (3.127).

$$\begin{aligned}&P(\mathbf{O},S^*,l\mid M)=\delta_{S_1^*}\Big(\prod_{t=1}^{T-1}\gamma_{S_t^*S_{t+1}^*}c_{S_t^*l_t}\pi_{\mathbf{o}_tS_t^*l_i}\Big)c_{S_T^*l_T}\pi_{\mathbf{o}_TS_T^*l_T}\\&\log P(\mathbf{O},S^*,l\mid \widehat{M})=\log\widehat{\delta}_{S_1^*}+\sum_{t=1}^{T-1}\log\widehat{\gamma}_{S_t^*S_{t+1}^*}+\sum_{t=1}^{T}\log\widehat{c}_{S_t^*l_t}+\sum_{t=1}^{T}\log\mathcal{N}(\mathbf{o}_t,\widehat{\boldsymbol{\mu}}_{S_t^*l_t},\widehat{\boldsymbol{U}}_{S_t^*l_t})\end{aligned} \tag{3.127}$$

Po wykorzystaniu zależności (3.127) w równaniu (3.126) uzyskuje się zależność (3.128).

$$\begin{aligned}&Q(M,\widehat{M})=\sum_{l_1,l_2,\ldots,l_T}P(\mathbf{O},S^*,l\mid M)\log\widehat{\delta}_{S_1^*}+\sum_{l_1,l_2,\ldots,l_T}P(\mathbf{O},S^*,l\mid M)\sum_{t=1}^{T-1}\log\widehat{\gamma}_{S_t^*S_{t+1}^*}+\\&+\sum_{l_1,l_2,\ldots,l_T}P(\mathbf{O},S^*,l\mid M)\sum_{t=1}^{T}\log\widehat{c}_{S_t^*l_t}+\sum_{l_1,l_2,\ldots,l_T}P(\mathbf{O},S^*,l\mid M)\sum_{t=1}^{T}\log\mathcal{N}(\mathbf{o}_t,\widehat{\boldsymbol{\mu}}_{S_t^*l_t},\widehat{\boldsymbol{U}}_{S_t^*l_t})=\\&=P(\mathbf{O},S^*\mid M)\log\widehat{\delta}_{S_1^*}+P(\mathbf{O},S^*\mid M)\sum_{i=1}^{m}\sum_{j=1}^{m}f^{ij}\log\widehat{\gamma}_{ij}+\sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{po}\sum_{t:S_t^*=i}P(\mathbf{O},S^*,S_t=i,l_t=k\mid M)\log\widehat{c}_{ik}+\\&+\sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{po}\sum_{t:S_t^*=i}P(\mathbf{O},S^*,S_t=i,l_t=k\mid M)\log\pi_{\mathbf{o}_tik}=\\&=P(\mathbf{O},S^*\mid M)\log\widehat{\delta}_{S_1^*}+P(\mathbf{O},S^*\mid M)\sum_{i=1}^{m}\sum_{j=1}^{m}f^{ij}\log\widehat{\gamma}_{ij}+\sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{po}\sum_{t:S_t^*=i}v_t(i)c_{ik}\pi_{\mathbf{o}_tik}\log\widehat{c}_{ik}+\\&+\sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{po}\sum_{t:S_t^*=i}v_t(i)c_{ik}\pi_{\mathbf{o}_tik}\log\mathcal{N}(\mathbf{o}_t,\widehat{\boldsymbol{\mu}}_{ik},\widehat{\boldsymbol{U}}_{ik})\end{aligned} \tag{3.128}$$

Stosując metodę mnożników Lagrange'a można wyznaczyć estymator parametrów δ_i, γ_{ij}, c_{ik} podany w zależnościach (3.129).

$$\hat{\delta}_{S_1^*} = 1, \quad \delta_j = 0 \; dla \; j \neq S_1^*$$

$$\hat{\gamma}_{ij} = \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}}, \qquad f_{ij} - ilość\ przejść\ ze\ stanu\ i\ do\ j\ w\ ścieżce\ S^* = (S_1^*, S_2^*, ..., S_T^*) \quad (3.129)$$

$$\hat{c}_{ik} = c_{ik} \frac{\sum_{po} \sum_{t: S_1^* = i} v_t(i) \pi_{\mathbf{o}_t ik}}{\sum_{k=1}^{N} c_{ik} \sum_{po} \sum_{t: S_1^* = i} v_t(i) \pi_{\mathbf{o}_t ik}}$$

Estymator parametrów $\hat{\boldsymbol{\mu}}_{ik}, \hat{\boldsymbol{U}}_{ik}$ można wyznaczyć z czwartej części równania (3.128). W wyniku rozwiązania układu równań powstałego przez zróżniczkowanie czwartej części funkcji Q i przyrównanie pochodnych do zera uzyskuje się estymator podany w zależnościach (3.130).

$$\hat{\boldsymbol{\mu}}_{ik} = \frac{\sum_{po} \sum_{t: S_t^* = i} v_t(i) c_{ik} \pi_{\mathbf{o}_t ik} \mathbf{o}_t}{\sum_{k=1}^{N} c_{ik} \sum_{po} \sum_{t: S_t^* = i} v_t(i) c_{ik} \pi_{\mathbf{o}_t ik}}$$

$$\hat{\boldsymbol{U}}_{ik} = \frac{\sum_{po} \sum_{t: S_t^* = i} v_t(i) c_{ik} \pi_{\mathbf{o}_t ik} (\mathbf{o}_t - \hat{\boldsymbol{\mu}}_{ik})(\mathbf{o}_t - \hat{\boldsymbol{\mu}}_{ik})^T}{\sum_{k=1}^{N} c_{ik} \sum_{po} \sum_{t: S_t^* = i} v_t(i) c_{ik} \pi_{\mathbf{o}_t ik}} \quad (3.130)$$

Ponieważ została wprowadzona funkcja pomocnicza Q, dlatego podczas wyznaczania nowych parametrów konieczne jest iteracyjne stosowanie procedury estymacji.

W literaturze można znaleźć podobną procedurę, w której nie uwzględnia się czynnika $v_t(i)$, wówczas komponenty mikstury (*podstany* l) traktowane są jak oddzielne stany. Uzasadnione jest sprawdzenie wyników optymalizacji uzyskanych dzięki zastosowaniu obu algorytmów.

3.5.5. Algorytm segmental MAP estymacji parametrów modelu HMM z rozkładami ciągłymi

Podobnie jak w algorytmie omówionym w rozdziale (3.5.4) na optymalnej ścieżce S^* uwzględnione są tylko stany modelu natomiast nie są uwzględnione podstany l, dlatego konieczne jest zastosowanie algorytmu (EM). W celu zastosowania algorytmu (EM) należy wprowadzić zmodyfikowaną funkcję pomocniczą R opisaną równaniem (3.131).

$$R(M, \hat{M}) = \sum_{l_1, l_2, ..., l_T} P(\mathbf{O}, S^*, l \mid M) \log P(\mathbf{O}, S^*, l \mid \hat{M}) + \log g(\hat{M}) \quad (3.131)$$

Prawdopodobieństwo na ścieżce może zostać zapisana w postaci sumy (3.125), gdzie $l=\{l_1, l_2, \ldots, l_T\}$ dotyczy podstanów modelu CDHMM. Wykorzystany rozkład „*prior*" modelu $g(\hat{M})$ jest identyczny jak zastosowany w algorytmie Bayesa z podrozdziału (3.5.3). Funkcje R można zapisać jak w równaniu (3.132).

$$
\begin{aligned}
R(M,\widehat{M}) &= \log\widehat{\delta}_{S_1^*} + \sum_{i=1}^{m}\sum_{j=1}^{m} f^{ij}\log\widehat{\gamma}_{ij} + \\
&+ \sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{po\ t:S_t^*=i} v_t(i)c_{ik}\pi_{\mathbf{o}_t ik}\log\widehat{c}_{ik} + \\
&+ \sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{po\ t:S_t^*=i} v_t(i)c_{ik}\pi_{\mathbf{o}_t ik}\log\mathcal{N}(\mathbf{o}_t,\widehat{\boldsymbol{\mu}}_{ik},\widehat{\boldsymbol{U}}_{ik}) + \\
&+ \sum_{i=1}^{m}\eta_i\log\widehat{\delta}_i + \sum_{i=1}^{m}\sum_{j=1}^{m}\eta_{ij}\log\widehat{\gamma}_{ij} + \\
&+ \sum_{i=1}^{m}\sum_{k=1}^{N}\nu_{ik}\log\widehat{c}_{ik} + \sum_{i=1}^{m}\sum_{k=1}^{N}\log g(\widehat{\boldsymbol{\mu}}_{\boldsymbol{ik}},\widehat{\boldsymbol{U}}_{\boldsymbol{ik}}) + \log(KG)
\end{aligned}
\tag{3.132}
$$

Po minimalizacji (3.132) uzyskuje się estymator parametrów δ_i, γ_{ij}, c_{ik} podany w zależnościach (3.133).

$$
\begin{aligned}
&\hat{\delta}_{S_1^*} = \frac{\eta_{S_1^*}+1}{\sum_{i=1}^{m}\eta_i+1}, \quad \delta_j = \frac{\eta_j}{\sum_{i=1}^{m}\eta_i+1}\ dla\ j\neq S_{1\,t}^{**} \\
&\hat{\gamma}_{ij} = \frac{\eta_{ij}+f_{ij}}{\sum_{j=1}^{m}\eta_{ij}+\sum_{j=1}^{m}f_{ij}}, \quad f_{ij}-ilość\ przejść\ ze\ stanu\ i\ do\ j\ w\ ścieżce\ S^*=(S_1^*,S_2^*,...,S_T^*) \\
&\hat{c}_{ik} = \frac{\nu_{ik}+c_{ik}\sum_{po\ t:S_t^*=i} v_t(i)\pi_{\mathbf{o}_t ik}}{\sum_{k=1}^{N}\nu_{ik}+\sum_{k=1}^{N}c_{ik}\sum_{po\ t:S_t^*=i} v_t(i)\pi_{\mathbf{o}_t ik}}
\end{aligned}
\tag{3.133}
$$

Estymator parametrów $\widehat{\boldsymbol{\mu}}_{ik}, \widehat{\boldsymbol{U}}_{ik}$ można wyznaczyć z czwartej części równania (3.132). W wyniku rozwiązania układu równań powstałego przez zróżniczkowanie czwartej części funkcji R i przyrównanie pochodnych do zera uzyskuje się estymator podany w zależnościach (3.134).

$$
\begin{aligned}
\widehat{\boldsymbol{\mu}}_{ik} &= \frac{\tau_{ik}\boldsymbol{m}_{ik} + \sum_{po\ t:S_t^*=i} v_t(i)c_{ik}\pi_{\mathbf{o}_t ik}\mathbf{o}_t}{\tau_{ik}+\sum_{k=1}^{N}c_{ik}\sum_{po\ t:S_t^*=i} v_t(i)c_{ik}\pi_{\mathbf{o}_t ik}} \\
\widehat{\boldsymbol{U}}_{\boldsymbol{ik}} &= \frac{2\boldsymbol{\beta}_{\boldsymbol{ik}} + \tau_{\boldsymbol{ik}}(\boldsymbol{m}_{\boldsymbol{ik}}-\widehat{\boldsymbol{\mu}}_{\boldsymbol{ik}})(\boldsymbol{m}_{\boldsymbol{ik}}-\widehat{\boldsymbol{\mu}}_{\boldsymbol{ik}})^T + \sum_{po\ t:S_t^*=i} v_t(i)c_{ik}\pi_{\mathbf{o}_t ik}(\mathbf{o}_t-\widehat{\boldsymbol{\mu}}_{ik})(\mathbf{o}_t-\widehat{\boldsymbol{\mu}}_{ik})^T}{\tau_{\boldsymbol{ik}}+\sum_{k=1}^{N}c_{ik}\sum_{po\ t:S_t^*=i} v_t(i)c_{ik}\pi_{\mathbf{o}_t ik}}
\end{aligned}
\tag{3.134}
$$

Parametry rozkładu „*posterior*” powinny być modyfikowane według zależności (3.135).

$$\hat{\eta}_{S_1^*} = \eta_{S_1^*} + 1, \quad \hat{\eta}_i = \eta_i, \quad dla\ i \neq S_1^*$$
$$\hat{\eta}_{ij} = \eta_{ij} + f_{ij}$$
$$\hat{\nu}_{ki} = \nu_{ki} + c_{ik} \sum_{po \quad t:S_t^*=i} \nu_t(i)\pi_{\mathbf{o}_t ik}$$
$$\widehat{\boldsymbol{m}}_{\boldsymbol{ik}} = \widehat{\boldsymbol{\mu}}_{\boldsymbol{ik}}, \tag{3.135}$$
$$\widehat{\tau}_{\boldsymbol{ik}} = \tau_{\boldsymbol{ik}} + \sum_{k=1}^{N} c_{ik} \sum_{po \quad t:S_t^*=i} \nu_t(i) c_{ik} \pi_{\mathbf{o}_t ik},$$
$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{ik}} = \boldsymbol{\beta}_{\boldsymbol{ik}} + \frac{\tau_{\boldsymbol{ik}}(\boldsymbol{m}_{\boldsymbol{ik}} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{ik}})(\boldsymbol{m}_{\boldsymbol{ik}} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{ik}})^T}{2} + \frac{\sum_{po \quad t:S_t^*=i} \nu_t(i) c_{ik} \pi_{\mathbf{o}_t ik} (\mathbf{o}_t \mathbf{o}_t^T - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T)}{2}$$

3.6. Model niestacjonarny HMM „*Trended HMM*”

W tym podrozdziale zostaną omówione niestacjonarne modele HMM nazywane „*Trended HMM*” stosowane w systemach rozpoznawania mowy. W modelach niestacjonarnych została wprowadzona dodatkowa funkcja do rozkładów prawdopodobieństw w stanach modelu HMM. Wprowadzona funkcja ma postać wielomianu opisującego zmiany wartości średniej w czasie.

W publikacji [59] z roku 1992 został podany algorytm Bauma-Welcha optymalizacji modelu niestacjonarnego. W publikacji [63] z roku 1994 został podany algorytm segmental *k*-means optymalizacji oparty na maksymalizacji prawdopodobieństwa na optymalnej ścieżce, w drugim algorytmie optymalna ścieżka zostaje wyznaczona dzięki zastosowaniu zmodyfikowanego algorytmu Viterbiego. W obu wymienionych algorytmach zostały zastosowane odmienne wielomiany opisujące wartość średnią.

W pracy [107] z 1997 został podany algorytm optymalizacji oparty na kryterium (MCE). W opisanym algorytmie modyfikowane są tylko dwa modele tj. model zgodny z sekwencją treningową w kierunku przeciwnym do gradientu oraz model niezgodny z sekwencją treningową w kierunku zgodnym z gradientem przyjętej empirycznie funkcji oceny.

Najbardziej istotny w praktycznych zastosowaniach algorytm optymalizacji niestacjonarnych modeli HMM oparty na kryterium MAP został opisany w pracy [77] z 2001.

3.6.1. Algorytm Bauma-Welcha estymacji parametrów modelu niestacjonarnego

Model HMM może działać jako generator sekwencji obserwacji. W modelu niestacjonarnym obserwacja $\mathbf{o}_t$ w stanie jest generowana według zależności (3.136).

$$\mathbf{o}_t = \boldsymbol{F}_t(i,k) + \boldsymbol{R}(i,k), \quad \boldsymbol{F}_t(i,k) = \sum_{p=0}^{D} \boldsymbol{B}_{\boldsymbol{ik}}^{\boldsymbol{p}} t^p = \boldsymbol{B}_{ik}\boldsymbol{t}$$
$$\boldsymbol{B}_{ik} = [\boldsymbol{B}_{\boldsymbol{ik}}^{0} \cdots \boldsymbol{B}_{\boldsymbol{ik}}^{\boldsymbol{D}}] = \begin{bmatrix} b_{ik}^{10} & b_{ik}^{11} & \cdots & b_{ik}^{1D} \\ \vdots & \vdots & & \vdots \\ b_{ik}^{D0} & b_{ik}^{D0} & \cdots & b_{ik}^{DD} \end{bmatrix}, \quad \boldsymbol{t}^{\boldsymbol{T}} = [1 \quad t \cdots t^D] \tag{3.136}$$

Przyjęto, że rząd wielomianu opisującego wartość średnią jest równy wymiarowi D wektora obserwacji, stąd $\boldsymbol{F}_t(i,k)$ jest to wektor złożony z wielomianów rzędu D, opisujących dynamikę wartości średniej w czasie, natomiast $\boldsymbol{R}(i,k)$ jest to wektor losowany według D-wymiarowego rozkładu normalnego o wartości średniej równej zeru oraz macierzy kowariancji $\boldsymbol{U}_{ik}$. Dla powyżej opisanego modelu wartość prawdopodobieństwa obserwacji **o** w stanie i, k w chwili t wyznaczana jest z funkcji rozkładu podanej w zależności (3.137).

$$\mathcal{N}(\mathbf{o}_t, \boldsymbol{B}_{ik}\boldsymbol{t}, \boldsymbol{U}_{ik}) = \pi_{\mathbf{o}_t ik} = \frac{1}{(2\pi)^{D/2} \mid \boldsymbol{U}_{ik} \mid^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{o} - \boldsymbol{B}_{ik}\boldsymbol{t})^T \boldsymbol{U}_{ik}^{-1} (\mathbf{o} - \boldsymbol{B}_{ik}\boldsymbol{t}) \right] \tag{3.137}$$

Wartość prawdopodobieństwa obserwacji $P(\mathbf{O}|\mathrm{M})$ można wyznaczyć stosując algorytm *forward-backward*.

Estymator parametrów modelu wyznacza się przez minimalizacją funkcji pomocniczej Q (3.138).

$$\begin{aligned} Q(M \mid \widehat{M}) = &\sum_{i=1}^{m} P(\mathbf{O}, S_1 = i \mid M) \log \widehat{\delta}_i + \sum_{i=1}^{m} \sum_{j=1}^{N} \sum_{t=1}^{T-1} P(\mathbf{O}, S_t = i, S_{t+1} = j \mid M) \log \widehat{\gamma}_{ij} + \\ &+ \sum_{i=1}^{m} \sum_{k=1}^{N} \sum_{t=1}^{T} P(\mathbf{O}, S_t = i, l_t = k \mid M) \log \widehat{c}_{ik} + \sum_{i=1}^{m} \sum_{k=1}^{N} \sum_{t=1}^{T} P(\mathbf{O}, S_t = i, l_t = k \mid M) \log \mathcal{N}(\mathbf{o}_t, \widehat{\boldsymbol{B}}_{ik}\boldsymbol{t}, \widehat{\boldsymbol{U}}_{ik}) \end{aligned} \tag{3.138}$$

Z zależności widać, że estymator na parametry δ_i, γ_{ij}, c_{ik}, jest identyczny jak dla modelu podstawowego. Estymator składowych macierzy $\widehat{\boldsymbol{B}}_{ik}$, oraz diagonalnej macierzy $\widehat{\boldsymbol{U}}_{ik}$ wyznacza się z równań uzyskanych przez zróżniczkowanie czwartej części funkcji pomocniczej (3.138) po tych składowych i przyrównaniu pochodnych do zera.

$$\begin{aligned} &\sum_{i=1}^{m} \sum_{k=1}^{N} \sum_{t=1}^{T} \nu_t(i) c_{ik} \pi_{\mathbf{o}_t ik} \frac{\partial}{\partial \widehat{\boldsymbol{B}}_{ik}} \log \mathcal{N}_t(\mathbf{o}_t, \widehat{\boldsymbol{B}}_{ik}\boldsymbol{t}, \widehat{\boldsymbol{U}}_{ik}) = 0 \\ &\sum_{i=1}^{m} \sum_{k=1}^{N} \sum_{t=1}^{T} \nu_t(i) c_{ik} \pi_{\mathbf{o}_t ik} \frac{\partial}{\partial \widehat{\boldsymbol{U}}_{ik}} \log \mathcal{N}_t(\mathbf{o}_t, \widehat{\boldsymbol{B}}_{ik}\boldsymbol{t}, \widehat{\boldsymbol{U}}_{ik}) = 0 \end{aligned} \tag{3.139}$$

W wyniku zróżniczkowania (3.138) uzyskuje się zależności (3.140).

$$\begin{aligned} &\frac{\partial}{\partial \widehat{\boldsymbol{B}}_{ik}^{d}} \log \mathcal{N}_t(\mathbf{o}_t, \widehat{\boldsymbol{B}}_{ik}\boldsymbol{t}, \widehat{\boldsymbol{U}}_{ik}) = \widehat{\boldsymbol{U}}_{ik}^{-1} (\mathbf{o}_t - \sum_{l=0}^{D} \widehat{\boldsymbol{B}}_{ik}^{l} t^l) t^d, \qquad d = 0,1,...,D \\ &\frac{\partial}{\partial \widehat{\boldsymbol{U}}_{ik}} \log \mathcal{N}_t(\mathbf{o}_t, \widehat{\boldsymbol{B}}_{ik}\boldsymbol{t}, \widehat{\boldsymbol{U}}_{ik}) = \widehat{\boldsymbol{U}}_{ik}^{-3/2} [\widehat{\boldsymbol{U}}_{ik} - (\mathbf{o}_t - \widehat{\boldsymbol{B}}_{ik}\boldsymbol{t})(\mathbf{o}_t - \widehat{\boldsymbol{B}}_{ik}\boldsymbol{t})^T] \end{aligned} \tag{3.140}$$

Zależność na estymator macierzy kowariancji różni się od zależności dla modelu CDHMM jedynie sposobem wyznaczania estymowanej wartości średniej, natomiast estymator macierzy $\widehat{\boldsymbol{B}}_{ik}$ można wyznaczyć z równania (3.141).

$$\widehat{\boldsymbol{U}}_{ik}^{-1} \sum_{t=1}^{T} \nu_t(i) c_{ik} \pi_{\mathbf{o}_t ik} (\mathbf{o}_t - \sum_{l=0}^{D} \widehat{\boldsymbol{B}}_{ik}^{l} t^l) t^d = 0, \qquad d = 0,1,...,D \tag{3.141}$$

Równanie (3.141) można przekształcić do postaci (3.142). Współczynniki macierzy $C_{ik}(m)$ oraz wektor $\boldsymbol{V}_i(d)$ można wyznaczyć korzystając z algorytmu *forward-backward*.

$$\sum_{l=0}^{D} C_{ik}(l+d)\widehat{\boldsymbol{B}}_{ik}^{l} = \boldsymbol{V}_{ik}(d), \qquad d = 0,1,...,D,$$

$$C_{ik}(m) = \sum_{t=1}^{T} \nu_t(i) c_{ik} \pi_{\mathbf{o}_t ik} t^m, \quad m = 0,1,...,2D, \tag{3.142}$$

$$\boldsymbol{V}_{ik}(d) = \sum_{t=1}^{T} \nu_t(i) c_{ik} \pi_{\mathbf{o}_t ik} \mathbf{o}_t t^d, \quad d = 0,1,...,D,$$

Po dekompozycji i zapisaniu pierwszego równania (3.142) w notacji macierzowej uzyskuje się układ równań (3.143), z którego można wyznaczyć macierz $\widehat{\boldsymbol{B}}_{ik}$.

$$\begin{bmatrix} C_{ik}(0) & C_{ik}(1) & \cdots & C_{ik}(D) \\ C_{ik}(1) & C_{ik}(2) & \cdots & C_{ik}(D+1) \\ \vdots & \vdots & & \vdots \\ C_{ik}(D) & C_{ik}(D+1) & \cdots & C_{ik}(2D) \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{B}}_{\boldsymbol{ik}}^{\boldsymbol{0}\,T} \\ \widehat{\boldsymbol{B}}_{\boldsymbol{ik}}^{\boldsymbol{1}\,T} \\ \vdots \\ \widehat{\boldsymbol{B}}_{\boldsymbol{ik}}^{\boldsymbol{D}\,T} \end{bmatrix} = \begin{bmatrix} \boldsymbol{V}_{ik}^{\ T}(0) \\ \boldsymbol{V}_{ik}^{\ T}(1) \\ \vdots \\ \boldsymbol{V}_{ik}^{\ T}(D) \end{bmatrix}, \quad stąd: \tag{3.143}$$

$$\widehat{\boldsymbol{B}}_{\boldsymbol{ik}} = \left[\boldsymbol{C}_{ik}^{-1}\boldsymbol{V}_{\boldsymbol{ik}}^{\boldsymbol{T}}\right]^{T}$$

3.6.2. Algorytm segmental *k*-means estymacji parametrów modelu niestacjonarnego

W tym podrozdziale został opisany algorytm segmental *k*-means optymalizacji niestacjonarnego modelu HMM. W omówionym modelu obserwacja generowana jest według zależności (3.144).

$$\mathbf{o}_t = \boldsymbol{F}_{\boldsymbol{t}}(i,k) + \boldsymbol{R}(i,k) = \sum_{p=0}^{D} \boldsymbol{B}_{\boldsymbol{ik}}^{\boldsymbol{p}} (t-\tau_i)^p + \boldsymbol{R}(i,m) = \boldsymbol{B}_{ik}\boldsymbol{T} + \boldsymbol{R}(i,k),$$
$$\boldsymbol{T}^{\boldsymbol{T}} = \begin{bmatrix} 1 & t-\tau_i & \cdots & (t-\tau_i)^D \end{bmatrix} = \begin{bmatrix} 1 & d_t & \cdots & d_t^{\ D} \end{bmatrix} \tag{3.144}$$

W zależności (3.144) została wprowadzona dodatkowa wartość τ_i, oznaczająca czas, w którym model przeszedł do stanu *i*. Z powodu wprowadzenia dodatkowego parametru czas nie jest liczony od początku sekwencji jak w przypadku poprzednio omówionego modelu optymalizowanego algorytmem Bauma-Welcha.

Wartość prawdopodobieństwa obserwacji w stanie *i*, *k* w chwili *t* z uwzględnieniem czasu przebywania w stanie *i* od chwili τ_i wyznaczana jest z zależności (3.145).

$$\mathcal{N}(\mathbf{o}_t, \boldsymbol{B}_{ik}\boldsymbol{T}, \boldsymbol{U}_{ik}) = \pi_{\mathbf{o}_t ikd_t} = \frac{1}{(2\pi)^{D/2} |\boldsymbol{U}_{ik}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{o} - \boldsymbol{B}_{ik}\boldsymbol{T})^T \boldsymbol{U}_{ik}^{-1}(\mathbf{o} - \boldsymbol{B}_{ik}\boldsymbol{T})\right] \tag{3.145}$$

Optymalna sekwencja stanów $S^*=\{S_1^*,S_2^*,\ldots,S_T^*\}$, $l^*=\{l_1^*,l_2^*,\ldots,l_T^*\}$ niezbędna w celu zastosowani algorytmu segmental *k*-means wyznaczana jest z wykorzystaniem zmodyfikowanego algorytmu Viterbiego.

Zmodyfikowany Algorytm Viterbiego:

W celu wyznaczenia ścieżki należy zdefiniować parametr $\eta_t(i,k,d)$ opisany w równaniu (3.146). Wartość wprowadzonego parametry jest największym prawdopodobieństwem wzdłuż

ścieżki kończącej się w stanie $S_t=i$, dodatkowy parametr d określa czas przebywania w ostatnim stanie.

$$\eta_t(i,k,d) = \max_{S_1,S_2,...,S_t,l_1,l_2,...,l_t} P(S_1,S_2,...,S_t=i,l_t=k,d_t=d,\mathbf{o}_1,\mathbf{o}_2,...,\mathbf{o}_t), \quad 0 \le d < t \quad (3.146)$$

Rekurencyjna zależność wyznaczania parametru opisana jest zależnością (3.147). W celu zapamiętania optymalnej ścieżki wprowadza się parametr $\varphi_t(i,k,d)$, określający poprzedni stan na optymalnej ścieżce kończącej się w stanie i, k.

$$\begin{cases} \eta_1(i,k,0) = \delta_i c_{ik} \pi_{\mathbf{o}_i ik 0} & \varphi_1(i,k,0) = 0, \qquad 1 \le i \le m, \quad 1 \le k \le N \\ \eta_t(j,k,0) = & \varphi_t(j,k,0) = \arg\max_{i \ne j} \max_{1 \le l \le N} \max_{0 \le \tau < t-1} [\eta_{t-1}(i,m,\tau)\gamma_{ij} c_{jk}] \\ = \max_{i \ne j} \max_{1 \le l \le N} \max_{0 \le \tau < t-1} [\eta_{t-1}(i,l,\tau)\gamma_{ij} c_{jk}] \pi_{\mathbf{o}_t jk 0} & \\ & 1 \le j,i \le m, \quad 1 \le k,l \le N, \quad 2 \le t \le T \\ \eta_t(j,k,d) = \eta_{t-1}(j,k,d-1)\gamma_{jj} c_{jk} \pi_{\mathbf{o}_t jkd} & \varphi_t(j,k,d) = (j,k,d-1) \\ & 1 \le j \le m, \quad 1 \le k \le N, \quad 0 < d < t-1 \quad 2 \le t \le T \\ P^* = \max_{1 \le i \le m} \max_{1 \le k \le N} \max_{0 \le d < T} [\eta_T(i,k,d)] & \\ (S_T^*, l_T^*, d_T^*) = \arg\max_{1 \le i \le m} \max_{1 \le k \le N} \max_{0 \le d < T} [\eta_T(i,k,d)] & (S_t^*, l_t^*, d_t^*) = \varphi_{t+1}(S_{t+1}^*, l_{t+1}^*, d_{t+1}^*) \\ & t = T-1, T-2, ..., 1 \end{cases} \quad (3.147)$$

Podobnie jak dla podstawowego algorytmu Viterbiego ścieżkę można wyznaczyć korzystając z zapamiętanych parametrów $\varphi_t(j,k,d)$, lub przeszukując tablice $\eta_t(j,k,d)$. W wyniku zastosowania algorytmu (3.147) zostaje wyznaczona ścieżka postaci $S^*=\{S_1^*,S_2^*,\ldots,S_T^*\}$, $l^*=\{l_1^*,l_2^*,\ldots,l_T^*\}$ oraz wartości parametrów $d_1^*,d_2^*,\ldots,d_T^*$. Parametry d_t^* będą oznaczane przez d_t.

Ponieważ obserwacje przyporządkowane do stanów zostały zdeterminowane dzięki zastosowaniu algorytmu segmentacji, dlatego parametry modelu HMM w odróżnieniu od algorytmu Bauma-Welcha są modyfikowane dla oddzielnych wektorów obserwacji przyporządkowanych do stanu i, k.

Po znalezieniu optymalnej ścieżki $S^*=\{S_1^*,S_2^*,\ldots,S_T^*\}$, $l^*=\{l_1^*,l_2^*,\ldots,l_T^*\}$ można wyznaczyć prawdopodobieństwo na ścieżce oraz jego logarytm jak w zależności (3.148).

$$P(\mathbf{O},S^*,l^* \mid M)=\delta_{S_1^*}\left(\prod_{t=1}^{T-1}\gamma_{S_t^*S_{t+1}^*}\right)\prod_{t=1}^{T}c_{S_t^*l}\;\pi_{\mathbf{o}_t S_t^* l_t^* d_t}$$
$$\log P(\mathbf{O},S^*,l^* \mid M)=\log\hat{\delta}_{S_1^*}+\sum_{t=1}^{T-1}\log\hat{\gamma}_{S_t^*S_{t+1}^*}+\sum_{t=1}^{T}\log\hat{c}_{S_t^*l_t^*}+\sum_{t=1}^{T}\log\mathcal{N}(\mathbf{o}_t,\boldsymbol{B}_{S_t^*l_t^*}\boldsymbol{T},\hat{\boldsymbol{U}}_{S_t^*l_t^*}) \tag{3.148}$$

Stosując metodę mnożników Lagrange'a do pierwszych trzech elementów logarytmowanej zależności (3.148) można wyznaczyć estymator prawdopodobieństw początkowych δ_i, prawdopodobieństw przejść γ_{ij} oraz współczynników c_{ik} estymator dla tych współczynników jest taki jak dla modelu stacjonarnego.

Czwartą cześć równania (3.148) można zapisać w postaci (3.149).

$$\sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{po\;\; t:S_t^*=i,l_t^*=k}\log\mathcal{N}(\mathbf{o}_t,\hat{\boldsymbol{B}}_{ik}\boldsymbol{T},\hat{\boldsymbol{U}}_{ik}) \tag{3.149}$$

Estymator elementów macierzy $\hat{\boldsymbol{B}}_{ik}$ oraz diagonalnej macierzy kowariancji $\hat{\boldsymbol{U}}_{ik}$ można wyznaczyć rozwiązując równania powstałe przez przyrównanie pochodnych (3.149). do zera. Pochodne wyrazów z pod sumy (3.149) dla macierzy $\hat{\boldsymbol{B}}_{ik}$ przyjmują postać podaną w równaniu (3.150).

$$\frac{\partial}{\partial\hat{\boldsymbol{B}}_{ik}^{p}}\log\mathcal{N}(\mathbf{o}_t,\hat{\boldsymbol{B}}_{ik}\boldsymbol{T},\hat{\boldsymbol{U}}_{ik})=\hat{\boldsymbol{U}}_{\boldsymbol{ik}}^{-1}(\mathbf{o}_t-\sum_{l=0}^{D}\hat{\boldsymbol{B}}_{\boldsymbol{ik}}^{\boldsymbol{l}}d_t^{l})d_t^{p},\qquad p=0,1,...,D \tag{3.150}$$

Po uwzględnieniu (3.150) w (3.149) można wyznaczyć zależność (3.151).

$$\hat{\boldsymbol{U}}_{ik}^{-1}\sum_{po\;\; t:S_t^*=i,l_t^*=k}(\mathbf{o}_t-\sum_{l=0}^{D}\hat{\boldsymbol{B}}_{\boldsymbol{ik}}^{\boldsymbol{l}}d_t^{l})d_t^{p}=0,\qquad p=0,1,...,D \tag{3.151}$$

Równanie (3.151) można przekształcić do postaci (3.152).

$$\sum_{l=0}^{D}C_{ik}(l+p)\hat{\boldsymbol{B}}_{\boldsymbol{ik}}^{\boldsymbol{l}}=\boldsymbol{V}_{ik}(p),\qquad p=0,1,...,D,$$
$$C_{ik}(m)=\sum_{po\;\; t:S_t^*=i,l_t^*=k}d_t^{m}\quad m=0,1,...,2D,\quad \boldsymbol{V}_{ik}(p)=\sum_{po\;\; t:S_t^*=i,l_t^*=k}\mathbf{o}_t d_t^{p},\quad p=0,1,...,D, \tag{3.152}$$

Macierz $\hat{\boldsymbol{B}}_{ik}$ wyznacza się z takiego samego równania macierzowego jak w algorytmie Bauma-Welcha, różnica algorytmów polega na sposobie wyznaczania współczynników macierzy $\boldsymbol{C}_{\boldsymbol{ik}}$ oraz $\boldsymbol{V}_{\boldsymbol{ik}}$.

3.6.3. Wyznaczenie rozkładu „*prior*" dla niestacjonarnego rozkładu normalnego z diagonalną macierzą kowariancji

W pracy [77] został podany algorytm Bayesa adaptacji modelu niestacjonarnego HMM. W algorytmie tym wykorzystuje się rozkład „*prior*" modelu, ponieważ nie został wyjaśniony sposób wyznaczenia rozkładu „*prior*" w tym podrozdziale zostały wyznaczone pochodne tego rozkładu konieczne w celu zastosowania algorytmu opartego na kryterium (MAP). W celu wyznaczenia pochodnych została wykorzystana zaproponowana przez autora pracy metoda wyznaczenia tego rozkładu.

Należy założyć, że są dwa zbiory wektorów obserwacji, $\{\boldsymbol{x}_i\}$ o liczności τ, oraz $\{\boldsymbol{y}_i\}$ o liczności T. Dla modelu niestacjonarnego dla każdego wektora obserwacji określony jest parametr związany z czasem przebywania w stanie d_{x_i}, d_{y_j}. Podobnie jak dla rozkładu stacjonarnego można postawić pytanie, jakie powinny być parametry $\boldsymbol{B}$, $\boldsymbol{U}$ niestacjonarnego rozkładu normalnego, żeby maksymalizować iloczyn podany równaniem (3.153), oraz w jaki sposób wyznaczyć rozkład „*prior*" dla tego rozkładu. Rozpatrywany rozkład normalny jest to rozkład $\mathcal{N}$ z diagonalną macierzą kowariancji $\boldsymbol{U}$. Niestacjonarny wektor wartości średniej opisany jest przez wektor wielomianów rządu $H \le D$. Wektor ten uzyskuje się przez mnożenia macierzy współczynników wielomianów $\boldsymbol{B}$ przez wektor $\boldsymbol{T}$.

$$P(\{\boldsymbol{x}_i\},\{\boldsymbol{y}_j\}) = \prod_{i=1}^{\tau} \mathcal{N}(\boldsymbol{x}_i, \boldsymbol{B}\boldsymbol{T}_{x_i}, \boldsymbol{U}) \prod_{j=1}^{T} \mathcal{N}(\boldsymbol{y}_j, \boldsymbol{B}\boldsymbol{T}_{y_j}, \boldsymbol{U}), \tag{3.153}$$

$$\begin{aligned}\mathcal{N}(\boldsymbol{x}, \boldsymbol{B}\boldsymbol{T}_x, \boldsymbol{U}) &= K \frac{1}{|\boldsymbol{U}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{B}\boldsymbol{T}_x)^T \boldsymbol{U}^{-1}(\boldsymbol{x}-\boldsymbol{B}\boldsymbol{T}_x)\right] = \\ &= K \frac{1}{\prod_{d=1}^{D} u_d} \exp\left[-\frac{1}{2}\sum_{d=1}^{D} \frac{[x_d - (b_{d0} + b_{d1} d_x + \ldots + b_{dH} d_x^H)]^2}{u_d^2}\right],\end{aligned} \tag{3.154}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_D \end{bmatrix}, \boldsymbol{U} = \begin{bmatrix} u_1^2 & & 0 \\ & \ddots & \\ 0 & & u_D^2 \end{bmatrix}, \boldsymbol{B} = [\boldsymbol{B}(0) \cdots \boldsymbol{B}(H)] = \begin{bmatrix} b_{10} & b_{11} & b_{1H} \\ & & \\ b_{D0} & b_{D1} & b_{DH} \end{bmatrix}, \boldsymbol{T}_x^T = \begin{bmatrix} 1 \\ d_x \\ \vdots \\ d_x^H \end{bmatrix},$$

Parametry rozkładu $\widehat{\boldsymbol{B}}$ oraz $\widehat{\boldsymbol{U}}$ dla danych $\{\boldsymbol{x}_i\}$, $\{\boldsymbol{y}_i\}$ można dobrać rozwiązując równania (3.155) powstałe przez przyrównanie pochodnych logarytmu równania (3.153) do zera. Nie jest konieczne sprawdzanie znaku drugiej pochodnej, ponieważ maksymalizowana funkcja posiada jedno maksimum.

$$\begin{aligned}&\frac{\partial \log P(\{\mathbf{x}_i\},\{\mathbf{y}_j\})}{\partial \widehat{\boldsymbol{B}}(p)} = \widehat{\boldsymbol{U}}^{-1} \sum_{i=1}^{\tau} [\boldsymbol{x}_i - \sum_{l=0}^{H} \widehat{\boldsymbol{B}}(l) d_{x_i}^l] d_{x_i}^p + \widehat{\boldsymbol{U}}^{-1} \sum_{j=1}^{T} [\boldsymbol{y}_j - \sum_{l=0}^{H} \widehat{\boldsymbol{B}}(l) d_{y_j}^l] d_{y_j}^p = 0, \quad p = 0,1,...,H \\ &\frac{\partial \log P(\{\mathbf{x}_i\},\{\mathbf{y}_j\})}{\partial \widehat{\boldsymbol{U}}} = -\widehat{\boldsymbol{U}}^{-3/2} \sum_{i=1}^{\tau} [\widehat{\boldsymbol{U}} - (\boldsymbol{x}_i - \widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i})(\boldsymbol{x}_i - \widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i})^T] - \widehat{\boldsymbol{U}}^{-3/2} \sum_{j=1}^{T} [\widehat{\boldsymbol{U}} - (\boldsymbol{y}_j - \widehat{\boldsymbol{B}}\boldsymbol{T}_{y_j})(\boldsymbol{y}_j - \widehat{\boldsymbol{B}}\boldsymbol{T}_{y_j})^T] = 0\end{aligned} \tag{3.155}$$

Po przekształceniu (3.155) uzyskuje się zależności (3.156). Dla macierzy kowariancji wyznaczane są tylko elementy na diagonalnej.

$$\begin{aligned}&\boldsymbol{V}_x(p) + \boldsymbol{V}_y(p) = \sum_{l=0}^{H} [C_x(l+p) + C_y(l+p)] \widehat{\boldsymbol{B}}(l), \qquad p = 0,1,...,H \\ &\widehat{\boldsymbol{U}} = \frac{\sum_{i=1}^{\tau} (\boldsymbol{x}_i - \widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i})(\boldsymbol{x}_i - \widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i})^T + \sum_{j=1}^{T} (\boldsymbol{y}_j - \widehat{\boldsymbol{B}}\boldsymbol{T}_{y_j})(\boldsymbol{y}_j - \widehat{\boldsymbol{B}}\boldsymbol{T}_{y_j})^T}{\tau + T}\end{aligned} \tag{3.156}$$

gdzie:

$$\boldsymbol{V}_x(p) = \sum_{i=1}^{\tau} \boldsymbol{x}_i d_{x_i}^p, \quad \boldsymbol{V}_y(p) = \sum_{j=1}^{T} \mathbf{y}_j d_{y_j}^p, \quad C_x(m) = \sum_{i=1}^{\tau} d_{x_i}^m \quad C_y(m) = \sum_{j=1}^{T} d_{y_j}^m \quad m = 0,1,...,2H,$$

Należy wprowadzić dodatkowe oznaczenia:

$$\boldsymbol{C}_{x_i} = \boldsymbol{T}_{x_i}\boldsymbol{T}_{x_i}^{\boldsymbol{T}}, \quad \boldsymbol{V}_{x_i} = \boldsymbol{x}_i\boldsymbol{T}_{x_i}^{\boldsymbol{T}}, \quad \boldsymbol{C}_x = \sum_{i=1}^{\tau}\boldsymbol{C}_{x_i} = \sum_{i=1}^{\tau}\boldsymbol{T}_{x_i}\boldsymbol{T}_{x_i}^{\boldsymbol{T}}, \quad \boldsymbol{V}_x = \sum_{i=1}^{\tau}\boldsymbol{V}_{x_i} = \sum_{i=1}^{\tau}\boldsymbol{x}_i\boldsymbol{T}_{x_i}^{\boldsymbol{T}}$$

$$\boldsymbol{C}_{y_j} = \boldsymbol{T}_{y_j}\boldsymbol{T}_{y_j}^{\boldsymbol{T}}, \quad \boldsymbol{V}_{y_j} = \boldsymbol{y}_j\boldsymbol{T}_{y_j}^{\boldsymbol{T}}, \quad \boldsymbol{C}_y = \sum_{j=1}^{T}\boldsymbol{C}_{y_j} = \sum_{j=1}^{T}\boldsymbol{T}_{y_j}\boldsymbol{T}_{y_j}^{\boldsymbol{T}}, \quad \boldsymbol{V}_y = \sum_{j=1}^{T}\boldsymbol{V}_{y_j} = \sum_{j=1}^{T}\boldsymbol{y}_j\boldsymbol{T}_{y_j}^{\boldsymbol{T}}$$

Po dekompozycji pierwszego równania (3.156) można wyznaczyć parametry macierzy $\widehat{\boldsymbol{B}}^{\boldsymbol{T}}$ podane w równaniu (3.157).

$$\begin{aligned}&(\boldsymbol{C}_{\boldsymbol{x}} + \boldsymbol{C}_{\boldsymbol{y}})\widehat{\boldsymbol{B}}^{\boldsymbol{T}} = (\boldsymbol{V}_{\boldsymbol{x}} + \boldsymbol{V}_{\boldsymbol{y}})^{\boldsymbol{T}}, \quad \textit{stąd}:\\ &\widehat{\boldsymbol{B}}^{\boldsymbol{T}} = (\boldsymbol{C}_{\boldsymbol{x}} + \boldsymbol{C}_{\boldsymbol{y}})^{-1}(\boldsymbol{V}_{\boldsymbol{x}} + \boldsymbol{V}_{\boldsymbol{y}})^{\boldsymbol{T}}\end{aligned} \tag{3.157}$$

Jeżeli przyjmie się, że stopień wielomianu H=0, oraz d_{x_i}=0, d_{y_j}=0 dla każdego i, j wówczas $\boldsymbol{C}_x=\tau$, $\boldsymbol{C}_y=T$, natomiast $\boldsymbol{B}$, $\boldsymbol{V}_x$, $\boldsymbol{V}_y$ są wektorami, wtedy zależność (3.157) sprowadza się do zależności wyprowadzonych dla rozkładu stacjonarnego.

Jeżeli zostaną uwzględnione tylko obserwacje $\{\boldsymbol{x}_i\}$, estymator (ML) parametrów rozkładu przyjmuje postać podaną w równaniu (3.158).

$$\begin{cases}\boldsymbol{B}^{\boldsymbol{T}} = \boldsymbol{C}_{\boldsymbol{x}}^{-1}\boldsymbol{V}_{\boldsymbol{x}}^{\boldsymbol{T}},\\ \boldsymbol{U} = \dfrac{\sum_{t=1}^{\tau}(\boldsymbol{x}_t - \boldsymbol{B}\boldsymbol{T}_{x_i})(\boldsymbol{x}_t - \boldsymbol{B}\boldsymbol{T}_{x_i})^T}{\tau}\end{cases} \tag{3.158}$$

Po uwzględnieniu (3.158) w (3.156) oraz (3.157) zależności na $\widehat{\boldsymbol{B}}^{\boldsymbol{T}}$ oraz $\widehat{\boldsymbol{U}}$ przyjmują postać podaną w równaniu (3.159). Parametry $\boldsymbol{C}$, $\boldsymbol{B}$, $\boldsymbol{V}$ są powiązane równaniem macierzowym $\boldsymbol{C}\boldsymbol{B}^{\boldsymbol{T}}=\boldsymbol{V}^T$, dlatego w celu wyeliminowania obserwacji $\{\boldsymbol{x}_i\}$ poza macierzą $\boldsymbol{\phi} = \boldsymbol{B}^{\boldsymbol{T}}$ wystarczy dodatkowo zapamiętać macierz $\boldsymbol{M}=\boldsymbol{C}_x$.

$$\begin{aligned}&\widehat{\boldsymbol{B}}^{\boldsymbol{T}} = (\boldsymbol{C}_{\boldsymbol{x}} + \boldsymbol{C}_{\boldsymbol{y}})^{-1}(\boldsymbol{V}_{\boldsymbol{x}} + \boldsymbol{V}_{\boldsymbol{y}})^{\boldsymbol{T}} = (\boldsymbol{M} + \boldsymbol{C}_{\boldsymbol{y}})^{-1}(\boldsymbol{M}\boldsymbol{\phi} + \boldsymbol{V}_{\boldsymbol{y}}^{\boldsymbol{T}})\\ &\widehat{\boldsymbol{U}} = \frac{\sum_{i=1}^{\tau}(\boldsymbol{x}_i - \widehat{\boldsymbol{B}}\boldsymbol{T}_i)(\boldsymbol{x}_i - \widehat{\boldsymbol{B}}\boldsymbol{T}_i)^T + \sum_{j=1}^{T}(\boldsymbol{y}_j - \widehat{\boldsymbol{B}}\boldsymbol{T}_j)(\boldsymbol{y}_j - \widehat{\boldsymbol{B}}\boldsymbol{T}_j)^T}{\tau + T}\end{aligned} \tag{3.159}$$

Po przekształceniach pierwszego czynnika z mianownika drugiego równania (3.159) w celu wyeliminowania obserwacji $\{\boldsymbol{x}_i\}$ uzyskuje się zależność (3.160).

$$\begin{aligned}&\sum_{i=1}^{\tau}(\boldsymbol{x}_i - \widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i})(\boldsymbol{x}_i - \widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i})^T = \sum_{i=1}^{\tau}\boldsymbol{x}_i\boldsymbol{x}_i^T - \sum_{i=1}^{\tau}\boldsymbol{x}_i(\widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i})^T - \sum_{i=1}^{\tau}(\widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i})\boldsymbol{x}_i^T + \sum_{i=1}^{\tau}(\widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i})(\widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i})^T =\\ &= \sum_{i=1}^{\tau}\boldsymbol{x}_i\boldsymbol{x}_i^T - \sum_{i=1}^{\tau}(\boldsymbol{B}\boldsymbol{T}_{x_i})(\boldsymbol{B}\boldsymbol{T}_{x_i})^T + \sum_{i=1}^{\tau}(\boldsymbol{B}\boldsymbol{T}_{x_i})(\boldsymbol{B}\boldsymbol{T}_{x_i})^T - \sum_{i=1}^{\tau}\boldsymbol{V}_{x_i}\widehat{\boldsymbol{B}}^T - \sum_{i=1}^{\tau}\widehat{\boldsymbol{B}}\boldsymbol{V}_{x_i}^T + \sum_{i=1}^{\tau}(\widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i})(\widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i})^T =\end{aligned} \tag{3.160}$$

$$
\begin{aligned}
&= \sum_{i=1}^{\tau}[\boldsymbol{x}_i\boldsymbol{x}_i^T - (\boldsymbol{B}\boldsymbol{T}_{x_i})(\boldsymbol{B}\boldsymbol{T}_{x_i})^T] + \sum_{i=1}^{\tau}\boldsymbol{B}\boldsymbol{T}_{x_i}\boldsymbol{T}_{x_i}^T\boldsymbol{B}^T - \boldsymbol{V}_x\widehat{\boldsymbol{B}}^T - \widehat{\boldsymbol{B}}\boldsymbol{V}_x^T + \sum_{i=1}^{\tau}\widehat{\boldsymbol{B}}\boldsymbol{T}_{x_i}\boldsymbol{T}_{x_i}^T\widehat{\boldsymbol{B}}^T = \\
&= 2\boldsymbol{\beta} + \sum_{i=1}^{\tau}\boldsymbol{B}\boldsymbol{C}_{x_i}\boldsymbol{B}^T - \boldsymbol{B}\boldsymbol{C}_x^T\widehat{\boldsymbol{B}}^T - \widehat{\boldsymbol{B}}\boldsymbol{C}_x\boldsymbol{B}^T + \sum_{i=1}^{\tau}\widehat{\boldsymbol{B}}\boldsymbol{C}_{x_i}\widehat{\boldsymbol{B}}^T = \\
&= 2\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{C}_x\boldsymbol{B}^T - \boldsymbol{B}\boldsymbol{C}_x\widehat{\boldsymbol{B}}^T - \widehat{\boldsymbol{B}}\boldsymbol{C}_x\boldsymbol{B}^T + \widehat{\boldsymbol{B}}\boldsymbol{C}_x\widehat{\boldsymbol{B}}^T = 2\boldsymbol{\beta} + (\boldsymbol{B}-\widehat{\boldsymbol{B}})\,\boldsymbol{C}_x(\boldsymbol{B}-\widehat{\boldsymbol{B}})^T
\end{aligned}
\tag{3.160}
$$

Po wprowadzeniu wyników z (3.160) do (3.159) można wyznaczyć estymator (MAP) podany w równaniu (3.161).

$$
\begin{aligned}
&\widehat{\boldsymbol{B}}^{\boldsymbol{T}} = \widehat{\boldsymbol{M}}^{-1}(\boldsymbol{M}\boldsymbol{\phi} + \boldsymbol{V}_y^{\boldsymbol{T}}) \\
&\widehat{\boldsymbol{U}} = \frac{2\boldsymbol{\beta} + (\boldsymbol{\phi}^T - \widehat{\boldsymbol{B}})\,\widehat{\boldsymbol{M}}(\boldsymbol{\phi}^T - \widehat{\boldsymbol{B}})^T + \sum_{j=1}^{T}(\boldsymbol{y}_j - \widehat{\boldsymbol{B}}\boldsymbol{T}_j)(\boldsymbol{y}_j - \widehat{\boldsymbol{B}}\boldsymbol{T}_j)^T}{\tau + T}
\end{aligned}
\tag{3.161}
$$

Parametry rozkładu „*prior*” powinny być modyfikowane według zależności (3.162).

$$
\begin{aligned}
&\widehat{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{C}_y \\
&\widehat{\boldsymbol{B}}^T = \widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{M}}^{-1}(\boldsymbol{M}\boldsymbol{\phi} + \boldsymbol{V}_y^{\boldsymbol{T}}) \\
&\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta} + (\boldsymbol{\phi}^T - \widehat{\boldsymbol{B}})\,\widehat{\boldsymbol{M}}(\boldsymbol{\phi}^T - \widehat{\boldsymbol{B}})^T + \frac{1}{2}\sum_{j=1}^{T}[\boldsymbol{y}_j\boldsymbol{y}_j^T - (\widehat{\boldsymbol{B}}\boldsymbol{T}_{y_j})(\widehat{\boldsymbol{B}}\boldsymbol{T}_{y_j})^T] \\
&\widehat{\tau} = \tau + T
\end{aligned}
\tag{3.162}
$$

Pochodną rozkładu „*prior*” można wyznaczyć z zależności (3.155). Po elementarnych przekształceniach uzyskuje się zależności (3.163).

$$
\begin{aligned}
&\boldsymbol{U}^{-1}[\boldsymbol{\phi}^T\boldsymbol{M}^T(p) - \sum_{l=0}^{H}[\boldsymbol{M}(l+p)\boldsymbol{\phi}^T(l)] + \boldsymbol{U}^{-1}\sum_{j=1}^{T}[\boldsymbol{y}_j - \sum_{l=0}^{H}\boldsymbol{B}(l)d_{y_j}^l]d_{y_j}^p = 0, \quad p = 0,1,...,H \\
&-\boldsymbol{U}^{-3/2}[\tau\boldsymbol{U} - 2\boldsymbol{\beta} + (\boldsymbol{\phi}^T - \widehat{\boldsymbol{B}})\,\widehat{\boldsymbol{M}}(\boldsymbol{\phi}^T - \widehat{\boldsymbol{B}})^T] - \boldsymbol{U}^{-3/2}\sum_{j=1}^{T}[\boldsymbol{U} - (\boldsymbol{y}_j - \boldsymbol{B}\boldsymbol{T}_{y_j})(\boldsymbol{y}_j - \boldsymbol{B}\boldsymbol{T}_{y_j})^T] = 0
\end{aligned}
\tag{3.163}
$$

Korzystając z zależności (3.163) można zauważyć, że rozkład „*prior*” powinien spełniać własności (3.164).

$$
\begin{cases}
\dfrac{\partial \log g(\widehat{\boldsymbol{B}},\widehat{\boldsymbol{U}})}{\partial \boldsymbol{B}(p)} = \boldsymbol{U}^{-1}[\boldsymbol{\phi}^T\boldsymbol{M}(p) - \sum_{l=0}^{H}[\boldsymbol{M}(l+p)]\boldsymbol{\phi}^T(l)], \quad p = 0,1,...,H \\
\dfrac{\partial \log g(\widehat{\boldsymbol{B}},\widehat{\boldsymbol{U}})}{\partial \widehat{\boldsymbol{U}}} = -\boldsymbol{U}^{-3/2}[\tau\boldsymbol{U} - 2\boldsymbol{\beta} + (\boldsymbol{\phi}^T - \widehat{\boldsymbol{B}})\,\widehat{\boldsymbol{M}}(\boldsymbol{\phi}^T - \widehat{\boldsymbol{B}})^T]
\end{cases}
\tag{3.164}
$$

Z pochodnych (3.164) można wyznaczyć rozkład „*prior*”. Pomimo, że nie została wyznaczona jawna postać tego rozkładu możliwe jest zastosowanie algorytmu opartego na kryterium (MAP), wystarcza, że znane są pochodne tego rozkładu podane w zależnościach (3.164).

Wyznaczone zależności estymatora nie odbiegają w znaczny sposób od zależności podanych w pracy [77].

3.7. Skalowanie

Wartości parametrów modeli HMM: δ_i, γ_{ij} oraz $\pi_{o_t i}$ są mniejsze od jedynki, stąd dla t większych od 10 wartość parametrów: $\alpha_t(i)$ oraz $\beta_t(i)$ zmierza ekspotencjalnie do zera. Dla t większych od 100 wartość $\alpha_t(i)$ może przekroczyć precyzji komputera. Jedynym sposobem nie dopuszczenia do przekroczenia precyzji komputera jest zastosowanie skalowania.

Skalowanie przeprowadza się w celu utrzymania skalowanych wartości $\alpha_t(i)$ oraz $\beta_t(i)$ w pewnym dopuszczalnym przedziale dla czasu $1 \le t \le T$. Procedura skalowania powiela $\alpha_t(i)$ o czynnik, który jest niezależny od stanu i (*jest jedynie zależny od czasu t*). Po zastosowaniu skalowania czynnik, o który powiela się parametry redukuj się w zależnościach na estymator modelu.

W celu przeprowadzenia skalowania przez $\alpha_t(i)$ oznacza się współczynnik bez skalowania, przez $\hat{\alpha}_t(i)$ współczynnik po skalowaniu, dodatkowo przez $\bar{\hat{\alpha}}_t(i)$ został oznaczony lokalny współczynnik przed skalowaniem. Początkowo wylicza się $\alpha_1(i)$ dla t=1, według równania podstawowego oraz przyjmuje się, że $\bar{\hat{\alpha}}_1(i) = \alpha_1(i)$, oraz $\hat{\alpha}_1(i) = c_1 \alpha_1(i)$ gdzie: $c_1 = \dfrac{1}{\sum_{i=1}^{m} \alpha_1(i)}$. Następnie dla każdego t, $2 \le t \le T$ wylicza się $\bar{\hat{\alpha}}_t(i)$ według równania podstawowego korzystając z wcześniej skalowanych współczynników $\hat{\alpha}_{t-1}(i)$ według zależności (3.165).

$$\bar{\hat{\alpha}}_t(i) = \sum_{j=1}^{m} \hat{\alpha}_{t-1}(j) \gamma_{ji} \pi_{o_t i} \tag{3.165}$$

W następnej kolejności wyznacza się współczynnik skali c_t według zależności (3.166).

$$c_t = \frac{1}{\sum_{i=1}^{m} \bar{\hat{\alpha}}_t(i)} \tag{3.166}$$

Po wyznaczeniu współczynnika skaluje się wszystkie współczynniki $\bar{\hat{\alpha}}_t(i)$ dla czasu t wymnażając je przez znaleziony czynnik c_t jak w równaniu (3.167).

$$\hat{\alpha}_t(i) = c_t \bar{\hat{\alpha}}_t(i) \tag{3.167}$$

Korzystając z równań (3.165), (3.166), (3.167) skalowane $\hat{\alpha}_t(i)$ można zapisać w postaci podanej w równaniu (3.168).

$$\hat{\alpha}_t(i) = \frac{\sum_{j=1}^{m} \hat{\alpha}_{t-1}(j) \gamma_{ji} \pi_{o_t i}}{\sum_{i=1}^{m} \sum_{j=1}^{m} \hat{\alpha}_{t-1}(j) \gamma_{ji} \pi_{o_t i}} \tag{3.168}$$

Z indukcji wynika zależność (3.169).

$$\hat{\alpha}_{t-1}(j) = \left(\prod_{\tau=1}^{t-1} c_\tau \right) \alpha_{t-1}(j) \tag{3.169}$$

Korzystając z zależności (3.168), (3.169) współczynnik $\hat{\alpha}_t(i)$ można zapisać jak w równaniu (3.170).

$$\hat{\alpha}_t(j) = \frac{\sum_{j=1}^{m} \alpha_{t-1}(j) \left(\prod_{\tau=1}^{t-1} c_\tau \right) \gamma_{ji} \pi_{o_t i}}{\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_{t-1}(j) \left(\prod_{\tau=1}^{t-1} c_\tau \right) \gamma_{ji} \pi_{o_t i}} = \frac{\alpha_t(i)}{\sum_{i=1}^{m} \alpha_t(i)} \tag{3.170}$$

Z zależności (3.170) wynika, że $\alpha_t(i)$ jest efektywnie skalowane przez sumę $\alpha_t(i)$. Przy wyliczaniu β wykorzystuje się tą samą skalę c_t, stąd skalowanie $\beta_t(i)$ przybiera formę.

$$\hat{\beta}_t(i) = c_t \hat{\beta}_t(i) \tag{3.171}$$

Ponieważ wektor skalowania c_t pozostaje po skalowaniu $\alpha_t(i)$, a rozmiary $\alpha_t(i)$ oraz $\beta_t(i)$ są porównywalne, stąd użycie tego samego wektora jest efektywne w celu uzyskania rozsądnych wartości. Zależności na skalowane współczynniki mogą zostać zapisane jak równaniach (3.172).

$$\begin{aligned}
\hat{\alpha}_t(i) &= \left(\prod_{k=1}^{t} c_k \right) \alpha_1(i) = C_t \alpha_t(i) \\
\hat{\beta}_{t+1}(j) &= \left(\prod_{k=t+1}^{T} c_k \right) \beta_{t+1}(j) = D_{t+1} \beta_{t+1}(j) \\
C_T &= C_t D_{t+1} = \prod_{k=1}^{t} c_k \prod_{k=t+1}^{T} c_k = \prod_{k=1}^{T} c_k
\end{aligned} \tag{3.172}$$

Wprowadzony czynnik C_T jest niezależne od t. W równaniach na estymator (3.64) następuje skrócenie czynnika C_T. W zależności od formy estymatora parametru $\hat{\gamma}_{ij}$ modelu konieczne jest wymnożenie każdego elementu sumy licznika przez c_t co zostało pokazane w zależnościach (3.177). Procedura skalowania nie powinna być stosowana w każdej chwili t, ale powinna być wykonywana, gdy jest to konieczne.

Poza modyfikacją estymatora $\hat{\gamma}_{ij}$ kolejną zmianą spowodowaną skalowaniem modelu HMM jest procedura wyliczania $P(\mathbf{O}|M)$. W przypadku stosowania skalowania nie można obliczyć P, ale $\log(P)$, który można wyznaczyć z zależności (3.173).

$$\sum_{i=1}^{m} \hat{\alpha}_T(i) = 1 = \prod_{t=1}^{T} c_t \sum_{i=1}^{m} \alpha_T(i) = C_T \sum_{i=1}^{m} \alpha_T(i) \tag{3.173}$$

Korzystając z (3.173) uzyskuje się zależności (3.174), z których można wyznaczyć $\log(P(\mathbf{O}|M))$.

$$\begin{aligned}
&\prod_{t=1}^{T} c_t P(\mathbf{O}/M) = 1 \\
&P(\mathbf{O}/M) = \frac{1}{\prod_{t=1}^{T} c_t} \\
&\log[P(\mathbf{O}/M)] = -\sum_{t=1}^{T} \log c_t
\end{aligned} \tag{3.174}$$

Procedura (3.174) może być zastosowana tylko wtedy, jeżeli $\alpha_T(i)$ jest skalowane. W algorytmie Viterbiego, wykorzystuje się logarytmy współczynników a nie skalowania, co zostało pokazane w alternatywnym algorytmie Viterbiego zależności (3.51).

3.8. Wielokrotne sekwencje obserwacji

W celu odpowiedniego doboru parametrów modelu HMM konieczne jest zastosowanie wielu obserwacji treningowych. W przypadku zastosowania algorytmu Bayesa nie jest konieczne stosowanie wielokrotnych sekwencji obserwacji, ponieważ możliwe jest korygowanie parametrów do kolejnych sekwencji z uwzględnieniem poprzednich danych treningowych. W algorytmie Bauma-Welcha po wprowadzeniu nowych sekwencji treningowej zapominana jest informacja zawarta w modelu HMM związana z poprzednimi danymi treningowymi.

Zarówno w algorytmie Bauma-Welcha jak i w algorytmie Bayesa możliwe jest uwzględnienie kilku obserwacji (*zbioru obserwacji*) $\mathbf{O}=\{\mathbf{O}^1,\mathbf{O}^2,\ldots,\mathbf{O}^K\}$ w jednym kroku algorytmu optymalizacji, gdzie: $\mathbf{O}^k=(o_1^k,o_2^k,...,o_{T_k}^k)$ jest k tą sekwencją obserwacji. Zakłada się, że każda sekwencja obserwacji jest niezależna od innych sekwencji.

Funkcja celu (*kryterium optymalizacji*) w algorytmie Bauma-Welcha dla wielokrotnych sekwencji jest to iloczynu (3.175).

$$P(\mathbf{O}\,|\,\mathrm{M})=\prod_{k=1}^{K}P(\mathbf{O}^k\,|\,M)=\prod_{k=1}^{K}P_k \tag{3.175}$$

W celu wyprowadzenia zależności na estymator dla wielokrotnych sekwencji wprowadza się zmodyfikowaną funkcję pomocniczą Q w postaci sumy (3.176).

$$Q(M,\widehat{M})=\sum_{k=1}^{K}\frac{1}{P_k}\sum_{S_1,S_2,\ldots,S_T}P(\mathbf{O}^k,S\,|\,M)\log P(\mathbf{O}^k,S\,|\,\widehat{\mathrm{M}}) \tag{3.176}$$

Po minimalizacji funkcji (3.176) uzyskuje się zależności na estymator Bauma-Welcha dla wielokrotnych sekwencji obserwacji. Zmodyfikowane formuły dla parametrów δ_i, γ_{ij} oraz π_{ki} są podane w zależnościach (3.177).

$$\begin{aligned}
\widehat{\delta}_i &= \frac{\sum_{k=1}^{K}\frac{1}{P_k}\alpha_1^k(i)\beta_1^k(i)}{\sum_{k=1}^{K}\frac{1}{P_k}\sum_{j=1}^{m}\alpha_T^k(i)}=\frac{\sum_{k=1}^{K}\frac{1}{P_k}\alpha_1^k(i)\beta_1^k(i)}{K}=\frac{\sum_{k=1}^{K}\hat{\alpha}_1^k(i)\hat{\beta}_1^k(i)/c_t^k}{K}\\
\widehat{\gamma}_{ij} &= \frac{\sum_{k=1}^{K}\frac{1}{P_k}\sum_{t=1}^{T_k-1}\alpha_t^k(i)\gamma_{ij}\pi_{o_{t+1}^{(k)}j}\beta_{t+1}^k(i)}{\sum_{k=1}^{K}\frac{1}{P_k}\sum_{t=1}^{T_k-1}\alpha_t^k(i)\beta_t^k(i)}=\frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\hat{\alpha}_t^k(i)\gamma_{ij}\pi_{o_{t+1}^k j}\hat{\beta}_{t+1}^k(j)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\hat{\alpha}_t^k(i)\hat{\beta}_t^k(i)/c_t^k}\\
\widehat{\pi}_{ki} &= \frac{\sum_{k=1}^{K}\frac{1}{P_k}\sum_{(po\ \ t\ \ t.\dot{z}.o_t=k)}\alpha_t^k(i)\beta_t^k(i)}{\sum_{k=1}^{K}\frac{1}{P_k}\sum_{t=1}^{T_k}\alpha_t^k(i)\beta_t^k(i)}=\frac{\sum_{k=1}^{K}\sum_{(po\ \ t\ \ t.\dot{z}.o_t=k)}\hat{\alpha}_t^k(i)\hat{\beta}_t^k(i)/c_t^k}{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\hat{\alpha}_t^k(i)\hat{\beta}_t^k(i)/c_t^k}
\end{aligned} \tag{3.177}$$

W zależnościach (3.177) zostało uwzględnione skalowanie współczynników $\alpha_t(i)$ oraz $\beta_t(i)$. Po wprowadzeniu skalowania konieczne jest wprowadzenie współczynnika c_t^k w zależnościach estymatora w podanej formie.

$$\prod_{t=1}^{T_k} c_t^k = \frac{1}{P(\mathbf{O}^k \mid M)} = \frac{1}{P_k}, \quad \hat{\alpha}_t^k(i)\hat{\beta}_t^k(i) = \frac{1}{P_k} c_t^k \alpha_t^k(i)\beta_t^k(i) \tag{3.178}$$

Ponieważ każda sekwencja obserwacji ma własny wektor skalowania, dlatego każdy element sumy w (3.177) jest pomnożony przez czynnik skalujący (3.178), dlatego podczas skalowania nie jest konieczne wymnażanie przez czynnik $1/P_k$ jak to zostało błędnie pokazane w pracy [3] strona 370 zależność (6.112).

3.9. Metoda adaptacji MLLR (*Maximum Likelihood Linear Regression*)

Metoda MLLR (*Maximum Likelihood Linear Regression*) dotyczy modeli HMM z ciągłymi rozkładami prawdopodobieństw w stanach. Metoda ta jest stosowana w celu adaptacji modeli HMM do mówcy, podczas gdy nie istnieje wystarczająca ilość danych treningowych. W metodzie wyznaczane są przekształcenia liniowe służące do modyfikacji wartości średnich oraz macierzy kowariancji rozkładów w węzłach modeli HMM.

Ponieważ możliwe jest wyznaczenie jednego przekształcenia dla wszystkich rozkładów modelu, a nawet dla kilku modeli HMM, metoda ta może być z powodzeniem zastosowana nawet dla bardzo małej ilości danych treningowych, jest to istotną cechą tej metody.

W wyniku wprowadzenia adaptacji MLLR uzyskano znaczącą poprawę wyników rozpoznawania w porównaniu z modelami HMM niezależnymi do mówcy. Wprowadzenie modyfikacji liniowej jest uzasadnione i związana z różnicą długości traktów głosowych różnych mówców.

W tym rozdziale zostały wyprowadzone zależności pozwalające wyznaczać macierz przekształcenia liniowego $\boldsymbol{W}$ dla wartości średniej rozkładów, oraz macierz $\boldsymbol{H}$ przekształcenia macierzy kowariancji.

3.9.1. Modyfikacja wartości średniej

Modyfikację wartości średniej przeprowadza się według zależności (3.179).

$$\hat{\boldsymbol{\mu}} = \boldsymbol{W}\boldsymbol{\xi}, \quad \boldsymbol{\xi} = \begin{bmatrix} \omega \\ \mu_1 \\ \vdots \\ \mu_D \end{bmatrix}, \quad \omega = 1, \tag{3.179}$$

Wartość parametru ω może mieć różne wartości najczęściej przyjmuje się, że $\omega=1$. Jeżeli ω=0 oraz kolumny macierzy $\boldsymbol{W}$ od 2 do N+1 stanowią macierz jednostkową wówczas po przeprowadzeniu przekształcenia wektor wartości średniej nie zostaje zmodyfikowany.

Rozpatrywany model HMM posiada w stanach kombinację liniową rozkładów normalnych opisaną zależnością (3.180).

$$\pi_{\mathbf{o}j} = \sum_{k=1}^{D} c_{jk}\pi_{\mathbf{o}jk}$$
$$\pi_{\mathbf{o}jk} = \frac{1}{(2\pi)^{D/2} \mid \boldsymbol{U}_{jk} \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{o}-\boldsymbol{\mu}_{jk})^T \boldsymbol{U}_{jk}^{-1}(\mathbf{o}-\boldsymbol{\mu}_{jk}) \right] \tag{3.180}$$

Macierz przekształcenia wyznaczana jest przez minimalizację funkcji pomocniczej $Q(M \mid \widehat{M})$ identycznej jak w algorytmie opartym na kryterium (ML). Funkcję pomocniczą (3.181) maksymalizuje się dla danej sekwencji treningowej $\mathbf{O}=(\mathbf{o}_1,\ldots,\mathbf{o}_T)$ nowego mówcy zgodnej z modyfikowanym modelem *M*.

$$\begin{aligned} &Q(M \mid \widehat{M}) = \\ &\sum_{S_1,S_2,\ldots,S_T,l_1,l_2,\ldots,l_t} \left\{ P(\mathbf{O},S,l \mid M)\log\widehat{\delta}_{S_1} + P(\mathbf{O},S,l \mid M)\sum_{t=1}^{T-1}\log\widehat{\gamma}_{S_tS_{t+1}} + \right. \\ &\left. + P(\mathbf{O},S,l \mid M)\sum_{t=1}^{T}\log\widehat{c}_{S_tl_t} + P(\mathbf{O},S,l \mid M)\sum_{t=1}^{T}\log\widehat{\pi}_{\mathbf{o}_tS_tl_t} \right\} = \\ &= \sum_{i=1}^{m} P(\mathbf{O},S_1=i \mid M)\log\widehat{\delta}_i + \sum_{i=1}^{m}\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(\mathbf{O},S_t=i,S_{t+1}=j \mid M)\log\widehat{\gamma}_{ij} + \\ &+ \sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{t=1}^{T} P(\mathbf{O},S_t=i,l_t=k \mid M)\log\widehat{c}_{ik} + \sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{t=1}^{T} P(\mathbf{O},S_t=i,l_t=k \mid M)\log\widehat{\pi}_{\mathbf{o}_tik} \end{aligned} \tag{3.181}$$

W zmodyfikowanym modelu $\widehat{M}$ uwzględniona jest macierz przekształcenia $\boldsymbol{W}$, stąd rozkłady w nowym modelu opisane są zależnością (3.182).

$$\widehat{\pi}_{\mathbf{o}ik} = \frac{1}{(2\pi)^{D/2} \mid \boldsymbol{U}_{ik} \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{o}-\boldsymbol{W}_{ik}\boldsymbol{\xi}_{ik})^T \boldsymbol{U}_{ik}^{-1}(\mathbf{o}-\boldsymbol{W}_{ik}\boldsymbol{\xi}_{ik}) \right] \tag{3.182}$$

Po wyznaczeniu pochodnej funkcji $Q(M \mid \widehat{M})$ po elementach macierzy $\widehat{\boldsymbol{W}}_{ik}$ uzyskuje się zależności (3.183).

$$\begin{aligned} \frac{\partial}{\partial\widehat{\boldsymbol{W}}_{ik}} Q(M \mid \widehat{M}) &= \sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{t=1}^{T} P(\mathbf{O},S_t=i,l_t=k \mid M)\frac{\partial}{\partial\widehat{\boldsymbol{W}}_{ik}}\log\widehat{\pi}_{\mathbf{o}_tik} = \\ &= \sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{t=1}^{T}\gamma_t(i,k)\frac{\partial}{\partial\widehat{\boldsymbol{W}}_{ik}}\log\widehat{\pi}_{\mathbf{o}_tik} \end{aligned} \tag{3.183}$$

Parametr $\gamma_t(i,k)$ zdefiniowany jest jak w zależnościach (3.184).

$$\begin{aligned} &v_1(i) = \delta_1\beta_1(i) \\ &v_t(i) = \beta_t(i)\sum_{j=1}^{m}\gamma_{ji}\alpha_{t-1}(j), \quad t=2,\ldots,T, \\ &\gamma_t(i,k) = v_t(s)c_{ik}\pi_{\mathbf{o}_tik} \end{aligned} \tag{3.184}$$

Pochodna rozkładu normalnego zapisana jest w zależności (3.185).

$$\frac{\partial}{\partial \widehat{\boldsymbol{W}}_{ik}} \log \hat{\pi}_{\mathbf{o}_t ik} = -\frac{1}{2}\frac{\partial}{\partial \widehat{\boldsymbol{W}}_{ik}}[D\log(2\pi) + \log|\boldsymbol{U}_{ik}| + (\mathbf{o}_t - \widehat{\boldsymbol{W}}_{ik}\boldsymbol{\xi}_{ik})^T \boldsymbol{U}_{ik}^{-1}(\mathbf{o}_t - \widehat{\boldsymbol{W}}_{ik}\boldsymbol{\xi}_{ik})] = \\ = \boldsymbol{U}_{ik}^{-1}(\mathbf{o}_t - \widehat{\boldsymbol{W}}_{ik}\boldsymbol{\xi}_{ik})\boldsymbol{\xi}_{ik}^{\boldsymbol{T}} \tag{3.185}$$

Po podstawieniu (3.185) do równania (3.183) oraz przyrównaniu pochodnej Q do zera dla każdego, że stanów i oraz dla każdej składowej kombinacji k uzyskuje się równanie (3.186).

$$\sum_{t=1}^{T}\gamma_t(i,k)\boldsymbol{U}_{ik}^{-1}\mathbf{o}_t\boldsymbol{\xi}_{ik}^{\boldsymbol{T}} = \sum_{t=1}^{T}\gamma_t(i,k)\boldsymbol{U}_{ik}^{-1}\widehat{\boldsymbol{W}}_{ik}\boldsymbol{\xi}_{ik}\boldsymbol{\xi}_{ik}^{\boldsymbol{T}} \tag{3.186}$$

Po rozwiązaniu równania (3.186) uzyskuje się estymator wartości średniej (3.187), jest on identyczny jak estymator w algorytmie Bauma-Welcha opartym na kryterium (ML).

$$\hat{\boldsymbol{\mu}}_{ik} = \widehat{\boldsymbol{W}}_{ik}\boldsymbol{\xi}_{ik} = \frac{\sum_{t=1}^{T}\gamma_t(i,k)\mathbf{o}_t}{\sum_{t=1}^{T}\gamma_t(i,k)} \tag{3.187}$$

Ideą metody MLLR jest dobór macierzy $\widehat{\boldsymbol{W}}$ dla kilku stanów modelu HMM, dlatego podczas wyznaczania macierzy przekształcenia stosuje się regresje liniową polegającą na przyjęciu, że macierz jest identyczna dla kilku rozkładów normalnych, w związku z takim założeniem w zależności (3.186) uwzględnia się kilka stanów i=1,2,…,m, oraz (lub) komponentów kombinacji liniowej k=1,2,…,N, przez wprowadzenie sumy podanej w zależności (3.188).

Jeżeli przekształcenie $\widehat{\boldsymbol{W}}$ będzie wyznaczone dla każdego ze stanów modelu HMM wówczas należy przeprowadzić sumowanie po składowych kombinacji k=1,2,…,N, w wyniku tego zostanie wyznaczona oddzielna macierz dla każdego stanu i=1,2,…,m, w przypadku szukania jednej macierzy dla całego modelu HMM sumowanie należy przeprowadzić po stanach i=1,2,…,m oraz składowych kombinacji w stanie k=1,2…,N jak w zależności (3.188).

$$\sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{k=1}^{N}\gamma_t(i,k)\boldsymbol{U}_{ik}^{-1}\mathbf{o}_t\boldsymbol{\xi}_{ik}^{\boldsymbol{T}} = \sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{k=1}^{N}\gamma_t(i,k)\boldsymbol{U}_{ik}^{-1}\widehat{\boldsymbol{W}}\boldsymbol{\xi}_{ik}\boldsymbol{\xi}_{ik}^{\boldsymbol{T}} \tag{3.188}$$

W przypadku wyznaczania macierzy dla kilku modeli HMM, konieczne jest zastosowanie dodatkowych sekwencji obserwacji odpowiadających pozostałym modelom HMM.

W celu wyznaczenia macierzy $\widehat{\boldsymbol{W}}$ wprowadza się oznaczenia (3.189).

$$\boldsymbol{Z} = \sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{k=1}^{N}\gamma_t(i,k)\boldsymbol{U}_{ik}^{-1}\mathbf{o}_t\boldsymbol{\xi}_{ik}^{\boldsymbol{T}}, \quad \boldsymbol{V}^{ik} = \sum_{t=1}^{T}\gamma_t(i,k)\boldsymbol{U}_{ik}^{-1}, \quad \boldsymbol{D}^{ik} = \boldsymbol{\xi}_{ik}\boldsymbol{\xi}_{ik}^{\boldsymbol{T}} \tag{3.189}$$

Po wprowadzeniu oznaczeń zależność (3.188) można zapisać w postaci (3.190).

$$\boldsymbol{Z} = \sum_{i=1}^{m}\sum_{k=1}^{N}\boldsymbol{V}^{ik}\widehat{\boldsymbol{W}}\boldsymbol{D}^{ik} \tag{3.190}$$

Elementy macierzy $\boldsymbol{Z}^{ik}$, $\boldsymbol{V}^{ik}$ oraz $\boldsymbol{D}^{ik}$, które można wyznaczyć z zależności (3.189) są oznaczane odpowiednio przez z_{ab}^{ik}, v_{ab}^{ik}, d_{ab}^{ik}, natomiast elementy szukanej macierzy $\widehat{\boldsymbol{W}}$ są oznaczone przez $\hat{w}_{ab}$. Jeżeli macierz kowariancji jest diagonalna, wówczas macierz $\boldsymbol{Z}$ można zapisać w postaci (3.191).

$$\boldsymbol{Z}=\sum_{i=1}^{m}\sum_{k=1}^{N}\begin{bmatrix} v_{11}^{ik} & & 0 \\ & \ddots & \\ 0 & & v_{DD}^{ik} \end{bmatrix}\begin{bmatrix} \widehat{w}_{11} & \cdots & \widehat{w}_{1D} & \widehat{w}_{1D+1} \\ \vdots & & \vdots & \vdots \\ \widehat{w}_{D1} & \cdots & \widehat{w}_{DD} & \widehat{w}_{DD+1} \end{bmatrix}\begin{bmatrix} d_{11}^{ik} & \cdots & d_{1D}^{ik} & d_{1D+1}^{ik} \\ \vdots & & \vdots & \vdots \\ d_{D1}^{ik} & \cdots & d_{DD}^{ik} & d_{DD+1}^{ik} \\ d_{D+11}^{ik} & \cdots & d_{D+1D}^{ik} & d_{D+1D+1}^{ik} \end{bmatrix}=$$

$$\sum_{i=1}^{m}\sum_{k=1}^{N}\begin{bmatrix} v_{11}^{ik}\widehat{w}_{11} & \cdots & v_{11}^{ik}\widehat{w}_{1D} & v_{11}^{ik}\widehat{w}_{1D+1} \\ \vdots & & \vdots & \vdots \\ v_{DD}^{ik}\widehat{w}_{D1} & \cdots & v_{DD}^{ik}\widehat{w}_{DD} & v_{DD}^{ik}\widehat{w}_{DD+1} \end{bmatrix}\begin{bmatrix} d_{11}^{ik} & \cdots & d_{1D}^{ik} & d_{1D+1}^{ik} \\ \vdots & & \vdots & \vdots \\ d_{D1}^{ik} & \cdots & d_{DD}^{ik} & d_{DD+1}^{ik} \\ d_{D+11}^{ik} & \cdots & d_{D+1D}^{ik} & d_{D+1D+1}^{ik} \end{bmatrix}=\begin{bmatrix} \boldsymbol{z}_1^T \\ \vdots \\ \boldsymbol{z}_D^T \end{bmatrix} \tag{3.191}$$

Po przeniesieniu elementów v_{aa}^{ik} wiersza macierzy $\boldsymbol{V}^{ik}\widehat{\boldsymbol{W}}$ do macierzy $\boldsymbol{D}^{ik}$ a-ty wiersz $\boldsymbol{z}_a^T$ macierzy $\boldsymbol{Z}$ można zapisać jak w zależności (3.192).

$$\boldsymbol{z}_a^T=\sum_{i=1}^{m}\sum_{k=1}^{N}\begin{bmatrix} w_{aa} & \cdots & w_{aD} & w_{aD+1} \end{bmatrix}\begin{bmatrix} v_{aa}^{ik}d_{11}^{ik} & \cdots & v_{aa}^{ik}d_{1D}^{ik} & v_{aa}^{ik}d_{1D+1}^{ik} \\ \vdots & & \vdots & \vdots \\ v_{aa}^{ik}d_{D1}^{ik} & \cdots & v_{aa}^{ik}d_{DD}^{ik} & v_{aa}^{ik}d_{DD+1}^{ik} \\ v_{aa}^{ik}d_{D+11}^{ik} & \cdots & v_{aa}^{ik}d_{D+1D}^{ik} & v_{aa}^{ik}d_{D+1D+1}^{ik} \end{bmatrix}= \tag{3.192}$$

$$=\sum_{i=1}^{m}\sum_{k=1}^{N}\widehat{\boldsymbol{w}}_a^T\boldsymbol{G}_{\boldsymbol{a}}^{ik}=\widehat{\boldsymbol{w}}_a^T\sum_{i=1}^{m}\sum_{k=1}^{N}\boldsymbol{G}_{\boldsymbol{a}}^{ik}=\widehat{\boldsymbol{w}}_a^T\boldsymbol{G}_{\boldsymbol{a}}$$

Po lewostronnym pomnożeniu zależności (3.192) przez $\boldsymbol{G}^{-1}$ uzyskuje się zależność na a-ty wiersz $\widehat{\boldsymbol{w}}_a^T$ szukanej macierzy $\widehat{\boldsymbol{W}}$ (3.193).

$$\widehat{\boldsymbol{w}}_a^T=\boldsymbol{z}_a^T\boldsymbol{G}_{\boldsymbol{a}}^{-1},$$

$$\widehat{\boldsymbol{W}}=\begin{bmatrix} \widehat{\boldsymbol{w}}_1^T \\ \vdots \\ \widehat{\boldsymbol{w}}_D^T \end{bmatrix},\quad \boldsymbol{G}_{\boldsymbol{i}}=\begin{bmatrix} g_{11}^a & \cdots & g_{1D+1}^a \\ \vdots & & \vdots \\ g_{D+11}^a & \cdots & g_{D+1D+1}^a \end{bmatrix},\quad g_{pq}^a=\sum_{i=1}^{m}\sum_{k=1}^{N}v_{aa}^{ik}d_{pq}^{ik}, \tag{3.193}$$

3.9.2. Modyfikacja macierzy kowariancji

Modyfikację macierzy kowariancji $\boldsymbol{U}$ przeprowadza się według zależności (3.194).

$$\widehat{\boldsymbol{U}}=\boldsymbol{B}^{\boldsymbol{T}}\widehat{\boldsymbol{H}}\boldsymbol{B},\quad \boldsymbol{B}=\boldsymbol{C}^{-1},\quad \boldsymbol{U}^{-1}=\boldsymbol{C}\boldsymbol{C}^{\boldsymbol{T}} \tag{3.194}$$

W celu wyznaczenia macierzy przekształcenia $\widehat{\boldsymbol{H}}$ korzysta się z takiej samej pomocniczej funkcji $Q(M\mid\widehat{M})$ jak podczas modyfikacji wartości średniej, po opuszczeniu części niezależnych od $\widehat{\boldsymbol{H}}$ uzyskuje się zależność (3.195).

$$\frac{\partial}{\partial\widehat{\boldsymbol{H}}_{\boldsymbol{ik}}}Q(M\mid\widehat{M})=\sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{t=1}^{T}\gamma_t(i,k)\frac{\partial}{\partial\widehat{\boldsymbol{H}}_{\boldsymbol{ik}}}\log\widehat{\pi}_{\mathbf{o}_t ik} \tag{3.195}$$

W rozkładach nowego modelu $\widehat{M}$ wprowadza się zależność (3.194) na zmodyfikowaną macierz kowariancji $\widehat{\boldsymbol{U}}_{ik}$ oraz wyznaczoną wcześniej zmodyfikowana wartość średnią $\widehat{\boldsymbol{\mu}}_{ik}$, stąd logarytm rozkładu normalnego można zapisać jak w zależności (3.196). W zależności zostały

wykorzystane dwie własności rachunku macierzowego: $|\boldsymbol{B}^T\boldsymbol{H}\boldsymbol{B}|=|\boldsymbol{B}|^2|\boldsymbol{H}|$ oraz $[\boldsymbol{B}^T\boldsymbol{H}\boldsymbol{B}]^{-1}=\boldsymbol{C}\boldsymbol{H}^{-1}\boldsymbol{C}^T$, gdzie: $\boldsymbol{B}=\boldsymbol{C}^{-1}$ oraz $\boldsymbol{B}$, $\boldsymbol{C}$ - macierze diagonalne.

$$\begin{aligned}\log\hat{\pi}_{\mathbf{o}_t ik} &= -\frac{1}{2}[D\log(2\pi)+\log|\boldsymbol{B}_{ik}^{\boldsymbol{T}}\hat{\boldsymbol{H}}_{ik}\boldsymbol{B}_{ik}|+(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})^T\left[\boldsymbol{B}_{ik}^{\boldsymbol{T}}\hat{\boldsymbol{H}}_{ik}\boldsymbol{B}_{ik}\right]^{-1}(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})]= \\ &= -\frac{1}{2}[D\log(2\pi)+2\log|\boldsymbol{B}_{ik}|+\log|\hat{\boldsymbol{H}}_{ik}|+(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})^T\boldsymbol{C}_{ik}\hat{\boldsymbol{H}}_{ik}^{-1}\boldsymbol{C}_{ik}^{\boldsymbol{T}}(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})]= \\ &= -\frac{1}{2}[D\log(2\pi)+2\log|\boldsymbol{B}_{ik}|+\log|\hat{\boldsymbol{H}}_{ik}|+\{[\boldsymbol{C}_{ik}^{\boldsymbol{T}}(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})]^T\hat{\boldsymbol{H}}_{ik}^{-1}[\boldsymbol{C}_{ik}^{\boldsymbol{T}}(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})]\}\end{aligned} \tag{3.196}$$

Korzystając z zależność (3.196) pochodną rozkładu normalnego można zapisać jak w równaniu (3.197). W wyznaczeniu pochodnej (3.197) zostały wykorzystane własności: $\partial(\log|\boldsymbol{H}|)/\partial\boldsymbol{H}=(\boldsymbol{H}^{-1})^T$ oraz $(\partial/\partial\boldsymbol{H})\boldsymbol{w}^T\boldsymbol{H}^{-1}\boldsymbol{w}=-\boldsymbol{H}^{-1}\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{H}^{-1}$.

$$\frac{\partial}{\partial\hat{\boldsymbol{H}}_{ik}}\log\hat{\pi}_{\mathbf{o}_t ik}=\frac{1}{2}[(\hat{\boldsymbol{H}}_{ik}^{-1})^T-\hat{\boldsymbol{H}}_{ik}^{-1}[\boldsymbol{C}_{ik}^{\boldsymbol{T}}(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})][\boldsymbol{C}_{ik}^{\boldsymbol{T}}(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})]^T\hat{\boldsymbol{H}}_{ik}^{-1}] \tag{3.197}$$

Po podstawieniu pochodnej (3.197) do równania (3.195) oraz przyrównaniu pochodnej Q do zera dla każdego, że stanów i oraz dla każdej składowej kombinacji k można wyznaczyć równanie (3.198).

$$(\hat{\boldsymbol{H}}_{ik}^{-1})^T\sum_{t=1}^{T}\gamma_t(i,k)=(\hat{\boldsymbol{H}}_{ik}^{-1})\sum_{t=1}^{T}\gamma_t(i,k)[\boldsymbol{C}_{ik}^{\boldsymbol{T}}(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})][\boldsymbol{C}_{ik}^{\boldsymbol{T}}(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})]^T(\hat{\boldsymbol{H}}_{ik}^{-1}) \tag{3.198}$$

Po prawostronnym oraz lewostronnym pomnożeniu równania (3.198) przez $\hat{\boldsymbol{H}}$ oraz uwzględnieniu tego, że $(\boldsymbol{H}^{-1})^T=(\boldsymbol{H}^{-1})$ uzyskuje się równanie (3.199).

$$\hat{\boldsymbol{H}}_{ik}=\frac{\sum_{t=1}^{T}\gamma_t(i,k)[\boldsymbol{C}_{ik}^{\boldsymbol{T}}(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})][\boldsymbol{C}_{ik}^{\boldsymbol{T}}(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})]^T}{\sum_{t=1}^{T}\gamma_t(i,k)}=\boldsymbol{C}_{ik}^{\boldsymbol{T}}\frac{\sum_{t=1}^{T}\gamma_t(i,k)(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})^T}{\sum_{t=1}^{T}\gamma_t(i,k)}\boldsymbol{C}_{ik} \tag{3.199}$$

Nowa macierz kowariancji wyraża się zależnością (3.200). Wyznaczona zależność jest identyczna jak estymator w algorytmie Bauma-Welcha.

$$\hat{\boldsymbol{U}}_{ik}=\boldsymbol{B}_{ik}^{\boldsymbol{T}}\hat{\boldsymbol{H}}\boldsymbol{B}_{ik}=\frac{\sum_{t=1}^{T}\gamma_t(i,k)(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})^T}{\sum_{t=1}^{T}\gamma_t(i,k)} \tag{3.200}$$

Po uwzględnieniu wszystkich stanów oraz składowych kombinacji w stanie uzyskuje się zależność (3.201).

$$\hat{\boldsymbol{U}}_{ik}=\boldsymbol{B}_{ik}^{\boldsymbol{T}}\hat{\boldsymbol{H}}\boldsymbol{B}_{ik}=\boldsymbol{B}_{ik}^{\boldsymbol{T}}\frac{\sum_{i=1}^{m}\sum_{k=1}^{N}\left\{\boldsymbol{C}_{ik}^{\boldsymbol{T}}\left[\sum_{t=1}^{T}\gamma_t(i,k)(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})(\mathbf{o}_t-\hat{\boldsymbol{\mu}}_{ik})^T\right]\boldsymbol{C}_{ik}\right\}}{\sum_{i=1}^{m}\sum_{k=1}^{N}\sum_{t=1}^{T}\gamma_t(i,k)}\boldsymbol{B}_{ik} \tag{3.201}$$

Nowa macierz kowariancji jest diagonalna ze względu na wymnażanie przez diagonalną macierz $\boldsymbol{B}_{ik}$, dlatego wystarczające jest wyznaczenie tylko wartości na diagonalnej dla macierzy przekształcenia $\widehat{\boldsymbol{H}}$.

3.10. Redukcja wektora cech metodami LDA, HLDA

W systemach rozpoznawania stosuje się wektory cech reprezentujące dźwięk w oknie, wymiar wektorów jest najczęściej równy 39. Ponieważ modelowanie wektorów o tak dużej ilości składowych za pomocą modeli HMM jest złożone obliczeniowo, dlatego w aplikacjach można zastosować redukcje ilości składowych. W celu redukcji ilości składowych stosuje się dwie metody LDA (*Linear Discriminant Analysis*) oraz HLDA (*Heteroscadestic LDA*).

3.10.1. Metoda LDA

W celu redukcji wektora cech wyznacza się przekształcenie liniowe $\boldsymbol{\theta}\ \mathfrak{R}^n \rightarrow \mathfrak{R}^p$ dla $p<n$ podane w zależności (3.202). Przekształcenie to pozwala na wyznaczenie zredukowanego wektora $\boldsymbol{y}$ z zachowaniem cech sygnału dotyczących dyskryminacji (*zróżnicowania pomiędzy np. fonemami*).

$$\boldsymbol{y} = \boldsymbol{\theta}^T \boldsymbol{x} \tag{3.202}$$

Do wyznaczenia przekształcenia wykorzystuje się zbiór N wektorów treningowych $\{\boldsymbol{x}_{ij}\}$, gdzie indeks j opisuje przynależność do jedną z J klas, do której należy dana, natomiast i=1,2,…,N_j jest to indeks elementu klasy j. W zależnościach (3.203) podane są wartości średnie oraz macierze kowariancji dla poszczególnych klas.

$$\bar{\boldsymbol{x}}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} \boldsymbol{x}_{ij}, \qquad \overline{\boldsymbol{W}}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)^T \tag{3.203}$$

W zależnościach (3.204) podana jest wartość średnia oraz macierz kowariancji dla wszystkich klas.

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{j=1}^{J}\sum_{\boldsymbol{i=1}}^{N_j} \boldsymbol{x}_{ij}, \qquad \overline{\boldsymbol{T}} = \frac{1}{N}\sum_{j=1}^{J}\sum_{\boldsymbol{i=1}}^{N_j} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}})^T \tag{3.204}$$

Wartość N_j oznacza liczność wektorów w klasie j, stąd zachodzi (3.205).

$$\sum_{j=1}^{J} N_j = N \tag{3.205}$$

Przekształcenie $\boldsymbol{\theta}$ wyznacza się maksymalizując kryterium podane w zależności (3.206).

$$\widehat{\boldsymbol{\theta}}_p = \arg\max_{\boldsymbol{\theta}_p} \frac{\left|\boldsymbol{\theta}_p^T \overline{\boldsymbol{T}} \boldsymbol{\theta}_p\right|}{\left|\boldsymbol{\theta}_p^T \overline{\boldsymbol{W}} \boldsymbol{\theta}_p\right|}, \qquad gdzie: \quad \overline{\boldsymbol{W}} = \frac{1}{N}\sum_{j=1}^{J} N_j \overline{\boldsymbol{W}}_j \tag{3.206}$$

Ponieważ przekształcenie $\boldsymbol{\theta}$ jest to rzutowanie $R^n \rightarrow R^p$ dla $p<n$, dlatego wymaga się, żeby po rzutowaniu wartości średnie poszczególnych klas były maksymalnie oddalone oraz rozproszenie wektorów dla poszczególnych klas po rzutowaniu było jak najmniejsze. W kryterium (3.206), licznik odpowiada sumie kwadratów odległości pomiędzy poszczególnymi

klasami, natomiast mianownik odpowiada sumie czynników odpowiadających rozproszeniu zmiennych klas po przekształceniu $\boldsymbol{\theta}$.

Wiadomo, że rozwiązaniem zadania maksymalizacji (3.206) jest macierz $\widehat{\boldsymbol{\theta}}_p$ o wymiarach $n \times p$, której wiersze są to wektory własne odpowiadające największym wartością własnym macierzy $\boldsymbol{W}^{-1}\boldsymbol{U}$.

3.10.2. Metoda HLDA

Metoda HLDA jest rozszerzeniem metody LDA. W metodzie tej również szuka się przekształcenia liniowego $\widehat{\boldsymbol{\theta}}_p$ pozwalającego zredukować ilość parametrów. Metoda LDA nie działa prawidłowo dla pewnej klasy problemów w porównaniu z metodą HLDA.

W metodzie HLDA macierz transformacji $\boldsymbol{\theta}$ zostaje rozszerzona do wymiaru $n \times n$ jak w równaniu (3.207).

$$\boldsymbol{y} = \boldsymbol{\theta}^T \boldsymbol{x} = \begin{bmatrix} \boldsymbol{\theta}_p \\ \boldsymbol{\theta}_{(n-p)} \end{bmatrix} \boldsymbol{x} \tag{3.207}$$

W celu wyznaczenia transformacji najpierw wprowadza się całą macierz przekształcenia $\boldsymbol{\theta}$ a następnie maksymalizuje się wprowadzone kryterium (3.210) związane z wektorami cech po przekształceniu $\boldsymbol{\theta}$.

Zakłada się, ze tylko pierwsze p składowych wektorów po przekształceniu zachowują dyskryminację pomiędzy klasami, natomiast pozostałe (n-p) składowe nie przenoszą żadnych informacji różnicujących klasy.

Zakłada się, że dane każdej klasy po przekształceniu $\boldsymbol{\theta}$ mogą być modelowanie przez n wymiarowy rozkład Gaussa. Dla każdej klasy wprowadza się parametry rozkładu podane w równaniach (3.208). Ponieważ ostatnie (n-p) składowe wektora po przekształceniu nie przenoszą informacji różnicujących klasy, dlatego można przyjąć, że parametry rozkładów $\bar{\boldsymbol{y}}_{0,(n-p)}$ oraz $\overline{\boldsymbol{U}}_0^{(n-p)}$ są takie same dla każdej klasy.

$$\begin{aligned} \bar{\boldsymbol{y}}_j &= \begin{bmatrix} y_{j,1} & y_{j,2} & \cdots & y_{j,p} & y_{0,p+1} & y_{0,p+2} & \cdots & y_{0,n} \end{bmatrix}^T = \begin{bmatrix} \bar{\boldsymbol{y}}_{j,p} \\ \bar{\boldsymbol{y}}_{0,(n-p)} \end{bmatrix} \\ \overline{\boldsymbol{U}}_j &= \begin{bmatrix} \overline{\boldsymbol{U}}_j^p & 0 \\ 0 & \overline{\boldsymbol{U}}_0^{(n-p)} \end{bmatrix} \end{aligned} \tag{3.208}$$

Rozkład prawdopodobieństwa dla wektora i klasy j po przekształceniu przyjmuje postać (3.209).

$$P(\boldsymbol{y}_{ij}) = \frac{|\boldsymbol{\theta}|}{\sqrt{(2\pi)^n |\overline{\boldsymbol{U}}_j|}} \exp[-\frac{1}{2}(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_j)^T \overline{\boldsymbol{U}}_j^{-1} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_j)], \quad i = 1,2,...N_j, \quad j = 1,2,...,J \tag{3.209}$$

W celu znalezienia przekształcenia $\boldsymbol{\theta}$ definiuje się kryterium maksymalizacji (3.210). Przekształcenie $\boldsymbol{\theta}$ znajduje się rozwiązując równia powstałe przez przyrównanie pochodnych kryterium do zera. Pochodne kryterium liczone są po parametrach $\boldsymbol{\theta}$.

$$\log P(\{\boldsymbol{y}_{ij}\}) = \sum_{j=1}^{J} \sum_{i=1}^{N_j} [\log|\boldsymbol{\theta}| - \frac{1}{2}\log((2\pi)^n |\overline{\boldsymbol{U}}_j|) - \frac{1}{2}(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_j)^T \overline{\boldsymbol{U}}_j^{-1} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_j)] \tag{3.210}$$

Ponieważ parametry $\bar{\boldsymbol{y}}_j$ oraz $\overline{\boldsymbol{U}}_j$ zależą od przekształcenia $\boldsymbol{\theta}$, dlatego możliwe jest ich wcześniejsze wyznaczenie w postaci funkcji parametrów szukanego przekształcenia $\boldsymbol{\theta}$. Wartość wektora średniej $\bar{\boldsymbol{y}}_j$ dla danej klasy wyznaczona jest w równaniu (3.211), natomiast macierz kowariancji $\overline{\boldsymbol{U}}_j$ w równaniu (3.212).

$$\begin{aligned}
&\frac{\partial \log P}{\partial \bar{\boldsymbol{y}}_j} = -\frac{1}{2}\sum_{i=1}^{N_j}\frac{\partial}{\partial \bar{\boldsymbol{y}}_j}(\boldsymbol{\theta}^T\boldsymbol{x}_{ij} - \bar{\boldsymbol{y}}_j)^T\overline{\boldsymbol{U}}_j^{-1}(\boldsymbol{\theta}^T\boldsymbol{x}_{ij} - \bar{\boldsymbol{y}}_j)\\
&= \frac{1}{2}\sum_{i=1}^{N_j}[\overline{\boldsymbol{U}}_j^{-1} + (\overline{\boldsymbol{U}}_j^{-1})^T](\boldsymbol{\theta}^T\boldsymbol{x}_{ij} - \bar{\boldsymbol{y}}_j) = 0\\
&\Rightarrow N_j[\overline{\boldsymbol{U}}_j^{-1} + (\overline{\boldsymbol{U}}_j^{-1})^T]\bar{\boldsymbol{y}}_j = \sum_{i=1}^{N_j}[\overline{\boldsymbol{U}}_j^{-1} + (\overline{\boldsymbol{U}}_j^{-1})^T]\boldsymbol{\theta}^T\boldsymbol{x}_{ij}\\
&\Rightarrow \bar{\boldsymbol{y}}_j = \frac{1}{N_j}\sum_{i=1}^{N_j}\boldsymbol{\theta}^T\boldsymbol{x}_{ij} = \boldsymbol{\theta}^T\frac{1}{N_j}\sum_{i=1}^{N_j}\boldsymbol{x}_{ij} = \boldsymbol{\theta}^T\bar{\boldsymbol{x}}_j
\end{aligned} \tag{3.211}$$

$$\begin{aligned}
&\frac{\partial \log P}{\partial \overline{\boldsymbol{U}}_j} = -\frac{1}{2}\sum_{i=1}^{N_j}\frac{\partial}{\partial \overline{\boldsymbol{U}}_j}[(\boldsymbol{\theta}^T\boldsymbol{x}_{ij} - \boldsymbol{\mu}_j)^T\overline{\boldsymbol{U}}_j^{-1}(\boldsymbol{\theta}^T\boldsymbol{x}_{ij} - \boldsymbol{\mu}_j) + \log\left|\overline{\boldsymbol{U}}_j\right|] =\\
&= \frac{1}{2}\sum_{i=1}^{N_j}[\overline{\boldsymbol{U}}_j^{-1}(\boldsymbol{\theta}^T\boldsymbol{x}_{ij} - \boldsymbol{\mu}_j)(\boldsymbol{\theta}^T\boldsymbol{x}_{ij} - \boldsymbol{\mu}_j)^T\overline{\boldsymbol{U}}_j^{-1} - (\overline{\boldsymbol{U}}_j^T)^{-1}] =\\
&= -\frac{1}{2}N_j\overline{\boldsymbol{U}}_j^{-1} + \overline{\boldsymbol{U}}_j^{-1}\frac{1}{2}\sum_{i=1}^{N_j}(\boldsymbol{\theta}^T\boldsymbol{x}_{ij} - \boldsymbol{\mu}_j)(\boldsymbol{\theta}^T\boldsymbol{x}_{ij} - \boldsymbol{\mu}_j)^T\overline{\boldsymbol{U}}_j^{-1} = 0\\
&\Rightarrow \overline{\boldsymbol{U}}_j = \frac{1}{N_j}\sum_{i=1}^{N_j}\boldsymbol{\theta}^T(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)[\boldsymbol{\theta}^T(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)]^T\\
&\Rightarrow \overline{\boldsymbol{U}}_j = \boldsymbol{\theta}^T\frac{1}{N_j}\sum_{i=1}^{N_j}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)^T\boldsymbol{\theta} = \boldsymbol{\theta}^T\overline{\boldsymbol{W}}_j\boldsymbol{\theta}
\end{aligned} \tag{3.212}$$

Po wyznaczeniu parametrów $\bar{\boldsymbol{y}}_j$ oraz $\overline{\boldsymbol{U}}_j$ kryterium (3.210) dla jednej klasy *j* przyjmuje postać (3.213). W zależności (3.213) został wykorzystane założenie, że $\bar{\boldsymbol{y}}_{0,(n-p)}$ oraz $\overline{\boldsymbol{U}}_0^{(n-p)}$ są takie same dla każdej klasy.

$$\begin{aligned}
&\log P(\boldsymbol{y}_{1j}, \boldsymbol{y}_{2j}, ..., \boldsymbol{y}_{N_j j}) = \sum_{i=1}^{N_j}[-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{\theta}^T\overline{\boldsymbol{W}}_j\boldsymbol{\theta}\right| + \log|\boldsymbol{\theta}| -\\
&-\frac{1}{2}(\boldsymbol{y}_{ij} - \boldsymbol{\theta}^T\bar{\boldsymbol{x}}_j)^T(\boldsymbol{\theta}^T\overline{\boldsymbol{W}}_j\boldsymbol{\theta})^{-1}(\boldsymbol{y}_{ij} - \boldsymbol{\theta}^T\bar{\boldsymbol{x}}_j)] =\\
&= \frac{-N_j n}{2}\log(2\pi) - \frac{N_j}{2}\log\left|\boldsymbol{\theta}_p^T\overline{\boldsymbol{W}}_j\boldsymbol{\theta}_p\right| - \frac{N_j}{2}\log\left|\boldsymbol{\theta}_{(n-p)}^T\overline{\boldsymbol{T}}\boldsymbol{\theta}_{(n-p)}\right| + N_j\log|\boldsymbol{\theta}| -\\
&-\frac{1}{2}\sum_{i=1}^{N_j}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)^T\boldsymbol{\theta}(\boldsymbol{\theta}^T\overline{\boldsymbol{W}}_j\boldsymbol{\theta})^{-1}\boldsymbol{\theta}^T(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)
\end{aligned} \tag{3.213}$$

Ponieważ ostatni element sumy w wyrażeniu (3.213) nie zależy od $\boldsymbol{\theta}$ co zostało pokazane w zależności (3.214), dlatego może zostać opuszczony podczas maksymalizacji.

$$\frac{1}{2}\sum_{i=1}^{N_j}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_j)^T\boldsymbol{\theta}(\boldsymbol{\theta}^T\overline{\boldsymbol{W}}_j\boldsymbol{\theta})^{-1}\boldsymbol{\theta}^T(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_j)=\boldsymbol{\theta}(\boldsymbol{\theta}^T\overline{\boldsymbol{W}}_j\boldsymbol{\theta})^{-1}\boldsymbol{\theta}^T\frac{1}{2}\sum_{i=1}^{N_j}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_j)^T(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_j)=$$
$$=\frac{1}{2}\boldsymbol{\theta}(\boldsymbol{\theta}^T\overline{\boldsymbol{W}}_j\boldsymbol{\theta})^{-1}\boldsymbol{\theta}^T\overline{\boldsymbol{W}}_j=\frac{n}{2} \tag{3.214}$$

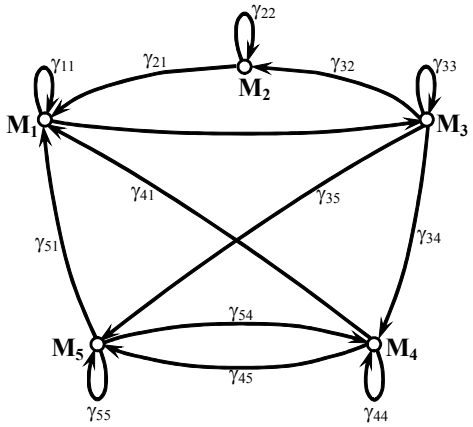
Ostatecznie kryterium maksymalizacji w metodzie HLDA przyjmuje postać (3.215).

$$\widehat{\boldsymbol{\theta}}=\arg\max_{\boldsymbol{\theta}}\sum_{j=1}^{J}\log P(\boldsymbol{y}_{1j},\boldsymbol{y}_{2j},...,\boldsymbol{y}_{N_jj}),\qquad gdzie:$$
$$\sum_{j=1}^{J}\log P(\boldsymbol{y}_{1j},\boldsymbol{y}_{2j},...,\boldsymbol{y}_{N_jj})=-\frac{N}{2}\log\left|\boldsymbol{\theta}_{(n-p)}^T\overline{\boldsymbol{T}}\boldsymbol{\theta}_{(n-p)}\right|-\sum_{j=1}^{J}\frac{N_j}{2}\log\left|\boldsymbol{\theta}_p^T\overline{\boldsymbol{W}}_j\boldsymbol{\theta}_p\right|+N\log|\boldsymbol{\theta}| \tag{3.215}$$

Nie istnieje zamknięta metoda rozwiązania zadania (3.215). Zadanie to rozwiązuje się metodami numerycznymi.

4. Model Sieciowy

W tym rozdziale jest omówiony model sieciowy HMM. Model sieciowy skonstruowany jest z modeli HMM z dyskretnymi rozkładami w stanach połączonych w sieć jak na rysunku (4.1). W sieci określone są prawdopodobieństwa początkowe, prawdopodobieństwa przejść oraz prawdopodobieństwa pozostawania w poszczególnych węzłach sieci.



Rysunek 4.1. Przykład modelu sieciowego z pięcioma pod modelami.

Poniżej zostały zdefiniowane parametry modelu sieciowego HMM dla dyskretnych rozkładów w stanach:

1. Z - liczba węzłów w modelu. Węzły zostały ponumerowane od 1 do z.
2. Macierz rozkładu prawdopodobieństw przejść pomiędzy węzłami oznaczona jest przez $\Gamma=\{\gamma_{ij}\}$, $\gamma_{ij}=P(C_{k+1}=j|C_k=i)$, $1\le i,j\le z$. Wartość γ_{ij} jest prawdopodobieństwem tego, że model znajdzie się w węźle j, przy założeniu, że wcześniej znajdował się w węźle i. Jeżeli założy się ze każdy węzeł może być osiągalny z każdego innego węzła w jednym kroku wówczas $\gamma_{ij}>0$ dla wszystkich i,j.
3. Wektor rozkładu początkowego prawdopodobieństwa dla sieci oznaczony jest przez $\delta=\{\delta_i\}$, gdzie $\delta_i=P(C_1=i)$, $1\le i\le z$. Wartość δ_i określa prawdopodobieństwo rozpoczęcia pracy w węźle i.
4. Rozkład prawdopodobieństwo pozostawania w węźle i przez okres d opisany jest przez rozkład dyskretny prawdopodobieństwa opisany zależnością (4.1).

$$p_i(d), \quad 1 \le d \le D$$
$$\sum_{k=1}^{D} p_i(k) = 1 \tag{4.1}$$

   Parametr D jest to maksymalny czas, przez jaki model może przebywać w poszczególnych węzłach, jest to wartość stała dla wszystkich węzłów.

W poszczególnych węzłach modelu sieciowego umieszczone są standardowe modele HMM z dyskretnymi rozkładami w stanach. Model sieciowy jest złożony z z pod modeli HMM $M^1,M^2,...,M^z$. Parametry pod modeli oznaczone są jak poniżej:

5. *m*-oznacza liczbę stanów w poszczególnych pod modelach. Stany modeli są ponumerowane od 1 do *m*. Stan w czasie *t* jest oznaczony jako S_t^z. Modele HMM w każdym węźle modelu sieciowego mają taką samą liczbę stanów.
6. *n* - oznacza liczbę różnych obserwacji w stanie. Obserwacje *o* przyjmują wartości od 1 do *n*. Przez o_t oznaczana jest obserwacja w chwili *t*.
7. Macierz rozkładu prawdopodobieństw przejść pomiędzy stanami dla poszczególnych pod modeli oznaczona jest przez $\Gamma^z=\{\gamma_{ij}^z\}$, $\gamma_{ij}^z=P(S_{t+1}^z=j|S_t^z=i)$, $1\le i,j\le m$. Wartość γ_{ij}^z jest prawdopodobieństwem tego, że pod model *z* znajdzie się w stanie *j*, przy założeniu, że wcześniej znajdował się w stanie *i*. Jeżeli założy się, że każdy stan może być osiągalny z każdego innego stanu w jednym kroku wówczas $\gamma_{ij}^z>0$ dla wszystkich *i,j*.
8. Macierz rozkładów prawdopodobieństwa obserwacji dla poszczególnych pod modeli oznaczona jest przez $\pi^z=\{\pi_{ki}{}^z\}$, $\pi_{ki}{}^z=P(o_t=k|S_t^z=i)$, $1\le k\le n$, $1\le i\le m$, $1\le z\le Z$. Wartość $\pi_{ki}{}^z$ jest prawdopodobieństwem tego, że pod model *z* wygeneruje obserwacje *k* w stanie *i*.
9. Wektor rozkładu prawdopodobieństwa początkowego w pod modelach oznaczony jest przez $\delta^z=\{\delta_i^z\}$, $\delta_i^z=P(S_1{}^z=i)$, $1\le i\le m$. Wartość δ_i^z jest prawdopodobieństwem, z jakim pod model *z* rozpocznie pracę w stanie *i*.

W powyższych punktach został w pełni zdefiniowany model sieciowy HMM. Tak zdefiniowany model będzie oznaczany skrótowo przez $M^z=(\Gamma,p,\delta,(\Gamma^z,\pi^z,\delta^z))$.

Należy zaznaczyć, ze model będąc w węźle $C_k=i$ generuje obserwacje $o_{d_1+d_2+\ldots+d_{l-1}+1},o_{d_1+d_2+\ldots+d_{l-1}+2},\ldots,o_{d_l}$ zgodnie z prawdopodobieństwem, oznaczonym przez $P(o_{d_1+d_2+\ldots+d_{l-1}+1},o_{d_1+d_2+\ldots+d_{l-1}+2},..,o_{d_l}\,|\,M^i)$. Prawdopodobieństwo to można policzyć metodami. opisanymi w rozdziale 3 (*Rozwiązanie problemu* 1).

4.1. Obliczenie prawdopodobieństwa obserwacji

W celu wyznaczenia $P(\mathbf{O}|M)$ dla modelu sieciowego muszą zostać wprowadzone zmiany do procedur dla modelu z czasem przebywania. Należy wprowadzić parametr $\alpha_t(i)$ opisany zależnością (4.2).

$$\alpha_t(i)=P(C_1,C_2,\ldots,C_t,\ przebywanie\ w\ \ węźle\ \ i\ \ kończy\ się\ w\ chwili\ \ t) \tag{4.2}$$

Należy założyć, że model był w *r* węzłach podczas pierwszych *t* chwil czasowych, węzły, w których przebywał są oznaczone przez $C_1,C_2,..,C_r$. Jeżeli czas pozostawania w każdym węźle zostanie oznaczony przez $d_1,d_2,\ldots,d_r$ wówczas zachodzi warunek (4.3).

$$\begin{aligned} &C_r=i \\ &\sum_{k=1}^{r} d_k=t \end{aligned} \tag{4.3}$$

Parametr $\alpha_t(i)$ może zostać zapisany jak w zależności (4.4).

$$\begin{aligned} \alpha_t(i)=\sum_C\sum_d &\delta_{C_1}\cdot p_{C_1}(d_1)\cdot P(o_1,o_2,\ldots,o_{d_1}\,|\,C_1) \\ &\cdot\gamma_{C_1C_2}\,p_{C_2}(d_2)\cdot P(o_{d_1+1},o_{d_1+2},\ldots,o_{d_1+d_2}\,|\,C_2) \\ &\cdot\gamma_{C_2C_3}\,p_{C_3}(d_3)\cdot P(o_{d_1+d_2+1},o_{d_1+d_2+2},\ldots,o_{d_1+d_2+d_3}\,|\,C_3)\cdot\ \ \ldots \\ &\cdot\gamma_{C_{r-1}C_r}\,p_{C_r}(d_r)\cdot P(o_{d_1+d_2+\ldots+d_{r-1}+1},o_{d_1+d_2+\ldots+d_{r-1}+2},\ldots,o_t\,|\,C_r) \end{aligned} \tag{4.4}$$

Sumowanie w zależności (4.4) przeprowadzane jest po wszystkich możliwych kombinacjach węzłów C oraz czasów pozostawania d. Poprzez indukcje parametr $\alpha_t(i)$ można zapisać jak w zależnościach (4.5).

$$\alpha_t(i) = \delta_i p_i(t) P(o_1, o_2, ..., o_t \mid M^i) + \sum_{j=1}^{m} \sum_{d=1}^{t-1} \alpha_{t-d}(j) \gamma_{ji} p_i(d) P(o_{t-d+1}, ..., o_t \mid M^i),$$

$$dla \quad t \le D \tag{4.5}$$

$$\alpha_t(i) = \sum_{j=1}^{m} \sum_{d=1}^{D} \alpha_{t-d}(j) \gamma_{ji} p_i(d) P(o_{t-d+1}, ..., o_t \mid M^i), \qquad dla \quad t > D$$

Parametr D reprezentuje maksymalnym okres pozostawania w danym węźle. W równaniach (4.6) podane są zależności na pierwsze trzy współczynniki $\alpha_t(i)$.

$$\alpha_1(i) = \delta_i p_i(1) P(o_1 \mid M^i)$$

$$\alpha_2(i) = \delta_i p_i(2) P(o_1, o_2 \mid M^i) + \sum_{j=1}^{m} \alpha_1(j) \gamma_{ji} p_i(1) P(o_2 \mid M^i)$$

$$\alpha_3(i) = \delta_i p_i(3) P(o_1, o_2, o_3 \mid M^i) + \sum_{j=1}^{m} \alpha_1(j) \gamma_{ji} p_i(2) P(o_2, o_3 \mid M^j) + \sum_{j=1}^{m} \alpha_2(j) \gamma_{ji} p_i(1) P(o_3 \mid M^i)$$

$$= \delta_i p_i(3) P(o_1, o_2, o_3 \mid M^i) + \sum_{j=1}^{m} \sum_{d=1}^{2} \alpha_{3-d}(j) \gamma_{ji} p_i(d) P(o_{3-d+1}, ..., o_3 \mid M^i) \tag{4.6}$$

Równanie (4.5) może być wykorzystane dla $t>D$ z pominięciem pierwszego członu związanego z prawdopodobieństwem początkowym. Prawdopodobieństwo $P(\mathbf{O}|M)$ można obliczyć podobnie jak dla modelu podstawowego korzystając z równania (4.7).

$$P(\mathbf{O} \mid M) = \sum_{i=1}^{m} \alpha_T(i) \tag{4.7}$$

4.2. Estymacja parametrów modelu sieciowego

W rozdziale tym został przedstawiony wyznaczony zdroworozsądkowo estymator parametrów modelu sieciowego oparty na kryterium (ML). Aby wyznaczyć estymator poza parametrem $\alpha_t(i)$ dodatkowo należy zdefiniować parametry podane w równaniach (4.8).

$$\alpha_t^*(i) = P(C_1, C_2, ..., C_{t-1}, następuje \quad przejście \quad do \quad węzla \quad i \quad w \quad chwili \quad t)$$

$$\beta_t(i) = P(C_{t+1}, C_{t+2}, ..., C_T \mid przebywanie \quad w \quad węzle \quad i \quad kończy \quad się \quad w \quad chwili \quad t) \tag{4.8}$$

$$\beta_t^*(i) = P(C_t, C_{t+1}, ..., C_T \mid następuje \quad przejście \quad do \quad węzla \quad i \quad w \quad chwili \quad t)$$

W zależnościach (4.9) podane są pierwsze trzy współczynniki $\beta_t^*(i)$:

$$\beta_T^*(i) = p_i(1)P(o_T \mid M^i)$$

$$\beta_{T-1}^*(i) = p_i(1)P(o_{T-1} \mid M^i)\sum_{j=1}^{m}\gamma_{ij}\beta_T^*(j) + p_i(2)P(o_{T-1}, o_T \mid M^i)$$

$$\beta_{T-2}^*(i) = p_i(1)P(o_{T-2} \mid M^i)\sum_{j=1}^{m}\gamma_{ij}\beta_{T-1}^*(j) + p_i(2)P(o_{T-2}, o_{T-1} \mid M^i)\sum_{j=1}^{m}\gamma_{ij}\beta_T^*(j) + \qquad (4.9)$$

$$p_i(3)P(o_{T-2}, o_{T-1}, o_T \mid M^i) = p_i(d)P(o_{T-2}, ..., o_{T-2+d-1} \mid M^i)\sum_{j=1}^{m}\sum_{d=1}^{2}\gamma_{ij}\beta_{T-2+d}^*(j) +$$

$$p_i(3)P(o_{T-2}, o_{T-1}, o_T \mid M^i)$$

Korzystając z zależności (4.9) rekurencyjny wzór na $\beta_t^*(i)$ można zapisać jak w równaniach (4.10).

$$\beta_t^*(i) = p_i(d)P(o_t, ..., o_{t+d-1} \mid M^i)\sum_{j=1}^{m}\sum_{d=1}^{T-1}\gamma_{ij}\beta_{t+d}^*(j) +$$

$$+ p_i(T-t+1)P(o_t, o_{t+1}, ..., o_T \mid M^i), \quad dla \quad t > T-D \qquad (4.10)$$

$$\beta_t^*(i) = p_i(d)P(o_t, ..., o_{t+d-1} \mid M^i)\sum_{j=1}^{m}\sum_{d=1}^{D}\gamma_{ij}\beta_{t+d}^*(j), \quad dla \quad t \le T-D$$

Powiązania pomiędzy parametrami $\alpha_t(i)$, $\alpha_t^*(i)$, $\beta_t(i)$ oraz $\beta_t^*(i)$ podane są w równaniach (4.11).

$$\alpha_1^*(j) = \delta_j, \quad \alpha_t^*(j) = \sum_{i=1}^{m}\alpha_{t-1}(i)\gamma_{ij}, \quad dla\ t > 1$$

$$\alpha_t(i) = \sum_{d=1}^{D}\alpha_{t-d+1}^*(i)p_i(d)P(o_{t-d+1}, ..., o_t \mid M^i)$$

$$\beta_t(j) = \sum_{i=1}^{m}\gamma_{ji}\beta_{t+1}^*(i), \quad dla\ t < T, \qquad \beta_T(j) = 1 \qquad (4.11)$$

$$\beta_t^*(i) = \sum_{d=1}^{D}\beta_{t+d-1}(i)p_i(d)P(o_t, ..., o_{t+d-1} \mid M^i)$$

Po wprowadzeniu parametru $\beta_t^*(i)$ prawdopodobieństwo obserwacji $P(\mathbf{O}|M)$ może być wyznaczone drugim sposobem podanym w równaniu (4.12).

$$P(\mathbf{O} \mid M) = \sum_{i=1}^{m}\delta_i\beta_1^*(i) \qquad (4.12)$$

Możliwość wyliczenia prawdopodobieństwa na dwa sposoby pozwala kontrolować prawidłowość zaimplementowanego algorytmu.

W oparciu o zdefiniowane współczynniki α, β estymator parametrów $\hat{\delta}_i$, $\hat{\gamma}_{ij}$, $\hat{p}_i(d)$ jest podany w zależnościach (4.13).

$$\widehat{\delta}_i = \frac{\delta_i \beta_1^*(i)}{P(\mathbf{O}\,|\,M)} = \frac{\delta_i \beta_1^*(i)}{\sum_{i=1}^{m} \alpha_T(i)}$$

$$\widehat{\gamma}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i)\gamma_{ij}\beta_{t+1}^*(j)}{\sum_{j=1}^{m}\sum_{t=1}^{T-1} \alpha_t(i)\gamma_{ij}\beta_{t+1}^*(j)} = \frac{\sum_{t=1}^{T-1} \alpha_t(i)\gamma_{ij}\beta_{t+1}^*(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} \tag{4.13}$$

$$\widehat{p}_i(d) = \frac{\sum_{t=1}^{T-d+1} \alpha_t^*(i)\, p_i(d)\, \beta_{t+d-1}(i) P(o_t,...,o_{t+d-1}\,|\,M^i)}{\sum_{d=1}^{D}\sum_{t=1}^{T-d+1} \alpha_t^*(i)\, p_i(d)\, \beta_{t+d-1}(i) P(o_t,...,o_{t+d-1}\,|\,M^i)}$$

Zależności na $\widehat{\delta}_i$ oraz na $\widehat{\gamma}_{ij}$ są identyczne jak dla modelu z jawną gęstością prawdopodobieństwa pozostawania w danym stanie. W zależności na $\widehat{p}_i(d)$ zamiast iloczynu prawdopodobieństw obserwacji w stanie pojawia się $P(o_t,...,o_{t+d-1}|M)$.

W modelu występują nowe parametry dotyczące pod modeli: prawdopodobieństwa początkowe δ_i^z, przejść γ_{ij}^z oraz wylosowania kolorów π_{ki}^z Jedną z możliwości estymacji tych parametrów jest uśrednianie wartości parametrów pod modeli wyliczonych dla wszystkich możliwych pod sekwencji o długości od 1 do D, takich pod sekwencji w modelu występuje $(T-D)D+((D+1)/2)D$. Parametry pod modeli można wyznaczyć stosując estymator (4.14).

$$\widehat{\delta}_i^z = \frac{\sum_{t=1}^{T}\sum_{d=1}^{D'} \left\{\alpha_t^*(z)\, p_z(d)\, \beta_{t+d-1}(z)\left[\alpha_1^{z(t,d)}(i)\beta_1^{z(t,d)}(i)\right]\right\}}{\sum_{t=1}^{T}\sum_{d=1}^{D'} \left\{\alpha_t^*(z)\, p_z(d)\, \beta_{t+d-1}(z)\left[\sum_{i=1}^{m} \alpha_1^{z(t,d)}(i)\beta_1^{z(t,d)}(i)\right]\right\}}$$

$$\widehat{\gamma}_{ij}^z = \frac{\sum_{t=1}^{T}\sum_{d=1}^{D'} \left\{\alpha_t^*(z)\, p_z(d)\, \beta_{t+d-1}(z)\left[\sum_{t'=1}^{d-1} \alpha_{t'}^{z(t,d)}(i)\gamma_{ij}\pi_{o_{t+1}j}\beta_{t'+1}^{z(t,d)}(j)\right]\right\}}{\sum_{t=1}^{T}\sum_{d=1}^{D'} \left\{\alpha_t^*(z)\, p_z(d)\, \beta_{t+d-1}(z)\left[\sum_{t'=1}^{d-1} \alpha_{t'}^{z(t,d)}(i)\beta_{t'}^{z(t,d)}(i)\right]\right\}}$$

$$\widehat{\pi}_{ki}^z = \frac{\sum_{t=1}^{T}\sum_{d=1}^{D'} \left\{\alpha_t^*(z)\, p_z(d)\, \beta_{t+d-1}(z)\left[\sum_{(po\ \ t'\ \ t.\dot{z}.\, o_{t'}=k)}^{d} \alpha_{t'}^{z(\tau,d)}(i)\beta_{t'}^{z(\tau,d)}(i)\right]\right\}}{\sum_{t=1}^{T}\sum_{d=1}^{D'} \left\{\alpha_t^*(z)\, p_z(d)\, \beta_{t+d-1}(z)\left[\sum_{t'=1}^{d} \alpha_{t'}^{z(t,d)}(i)\beta_{t'}^{z(t,d)}(i)\right]\right\}}$$

gdzie:

$$D' = D \ dla\ \ T - t \geq D$$
$$D' = T - t + 1 \ dla\ \ T - t < D \tag{4.14}$$

Wielkości $\alpha_{t'}^{z(t,d)}(i), \beta_{t'}^{z(t,d)}(i)$, są wyznaczane dla pod modeli według procedur dla podstawowych modeli HMM oraz dla sekwencji obserwacji $o_t,o_{t+1},...,o_{t+d-1}$.

Podane metody estymacji zostały opracowane intuicyjnie dzięki wykorzystaniu podobieństw pomiędzy modelem z jawną gęstością prawdopodobieństwa pozostawania w danym stanie opisanym w rozdziale (3) oraz zdefiniowanym w tym rozdziale modelem sieciowym.

Model sieciowy sprowadza się do modelu z jawnym prawdopodobieństwem pozostawania w danym stanie z podrozdziału (3.4), jeśli w każdym węźle modelu są pod modele składające się z pojedynczego stanu, oraz sprowadza się do modelu podstawowego z podrozdziału (3.3) w dwóch przypadkach. W pierwszym przypadku modele są równoważne, jeżeli w modelu sieciowym jest tylko jeden węzeł, oraz jeżeli prawdopodobieństwo pozostawania w tym węźle przyjmie wartość 1 dla czasu pozostawania równego T tj. $p_1(T)=1$ oraz $p_1(d)=0$ dla $d\neq T$. Wartość T jest to długość sekwencji obserwacji, do której optymalizowany jest model albo sekwencji, dla której wyznaczane jest prawdopodobieństwo obserwacji. Model sieciowy jest również równoważny modelowi podstawowemu HMM, jeśli w każdym węźle jest model jedno-stanowy HMM oraz prawdopodobieństwo pozostawania w węźle jest równe jeden dla czasu równego jeden tj. $p_1(1)=1$ oraz $p_1(t)=0$ dla $t\neq 1$.

4.3. Sprowadzenie estymatora modelu sieciowego do estymatora model z jawnym rozkładem prawdopodobieństwa pozostawania w tym samym stanie

Model sieciowy jest równoważny modelowi z jawnym rozkładem prawdopodobieństwa pozostawania w tym samym stanie, jeżeli w każdym węźle modelu są pod modele składające się z pojedynczego stanu. Wówczas podane metody obliczania prawdopodobieństwa oraz estymacji parametrów powinny sprowadzać się do zależności stosowanych do modelu z czasem przebywania. Równoważność zależności na obliczanie prawdopodobieństwa obserwacji jest oczywista. Poniżej zostanie pokazane, w jaki sposób redukują się zależności dla parametrów pod modeli.

Dla poszczególnych pod modeli z jednym stanem prawdopodobieństwo początkowo wynosi 1, prawdopodobieństwo przejść wynosi 1, dlatego pozostaje wykazanie równoważności estymatora prawdopodobieństw kolorów podanego w równaniach (4.15).

$$\widehat{\pi}_{ki}^{z} = \frac{\sum_{t=1}^{T}\sum_{d=1}^{D'}\left\{\alpha_t^*(z)p_z(d)\beta_{t+d-1}(z)\left[\sum_{(po\ \ t'\ \ t.\dot{z}.o_{t'}=k)}^{d}\alpha_{t'}^{z(\tau,d)}(i)\beta_{t'}^{z(\tau,d)}(i)\right]\right\}}{\sum_{t=1}^{T}\sum_{d=1}^{D'}\left\{\alpha_t^*(z)p_z(d)\beta_{t+d-1}(z)\left[\sum_{t'=1}^{d}\alpha_{t'}^{z(t,d)}(i)\beta_{t'}^{z(t,d)}(i)\right]\right\}} \qquad (4.15)$$

$gdzie$:

$$D' = D\ dla\ \ T-t \geq D$$

$$D' = T-t+1\ dla\ \ T-t < D$$

Jeżeli w pod modelach jest tylko jeden stan wówczas współczynniki $\alpha_{t'}^{z(\tau,d)}(1), \beta_{t'}^{z(\tau,d)}(1)$ przyjmują postać podaną w zależnościach (4.16).

$$\alpha_{t'}^{z(\tau,d)}(1) = \prod_{w=\tau}^{\tau+t'-1}\pi_{o_w 1}^{z}, \qquad \beta_{t'}^{z(\tau,d)}(1) = \prod_{w=\tau+t'}^{\tau+d-1}\pi_{o_w 1}^{z} \qquad (4.16)$$

Korzystając z zależności (4.16) iloczyn $\alpha_{t'}^{z(\tau,d)}(1)\beta_{t'}^{z(\tau,d)}(1)$ można zapisać jak w równaniu (4.17).

$$\alpha_{t'}^{z(\tau,d)}(1)\beta_{t'}^{z(\tau,d)}(1) = \prod_{w=\tau}^{\tau+d-1}\pi_{o_w 1}^{z} \tag{4.17}$$

Po uwzględnieniu zależności (4.16) w zależności (4.15) estymator π_{ki} można zapisać jak w zależności (4.18).

$$\widehat{\pi}_{ki}^{z} = \frac{\sum_{t=1}^{T}\sum_{d=1}^{D'}\left\{\alpha_t^*(z)p_z(d)\beta_{t+d-1}(z)\left[(Liczba\ wystapień\ k\ w\ o_\tau,...,o_{\tau+d-1})\prod_{t'=t}^{t+d-1}\pi_{o_{t'}}^{z}\right]\right\}}{\sum_{t=1}^{T}\sum_{d=1}^{D'}\left\{\alpha_t^*(z)p_z(d)\beta_{t+d-1}(z)\left[d\prod_{t'=t}^{t+d-1}\pi_{o_{t'}}^{z}\right]\right\}} \tag{4.18}$$

gdzie:

$$D' = D\ dla\ \ T-t \geq D$$
$$D' = T-t+1\ dla\ \ T-t < D$$

Estymator (4.18) można zmodyfikować korzystając z przykładu pokazanego w tabeli (4.1) oraz (4.2). Dla rozpatrywanego na przykładzie modelu długość sekwencji wynosi T=4 natomiast maksymalny czas pozostawania natomiast tym samym stanie wynosi D=2. Przykład dotyczy jednego stanu z modelu HMM. Na przykładzie za pomocą nawiasów uwzględnione są przedziały obserwacji, które są generowane w rozpatrywanym stanie z. Zależności opisujące prawdopodobieństwa wygenerowania obserwacji w przedziałach pomiędzy nawiasami w rozpatrywanym stanie z podane są w zależnościach w lewej części tabel. Prawdopodobieństwa wyrażone są przez współczynniki $\alpha_t(i)$, $\alpha_t^*(i)$, $\beta_t(i)$ oraz $\beta_t^*(i)$.

	sekwencja				liczba
elementy sumy	o_1	o_2	o_3	o_4	składowych
$\alpha_t^*(z)\mathrm{p}(1)\beta_t(z)\,\pi_{o_t i}$	( )	( )	( )	( )	**4**
$\alpha_t^*(z)\mathrm{p}(\mathrm{d})\beta_{t+d-1}(z)$ $\pi_{o_t i}\pi_{o_{t+1} i}$, d=2, t=1,2,3	(	)			**3**
		(	)		
			(	)	
liczba składowych	**2**	**3**	**3**	**2**	**=10**

Tabela 4.1. Współczynniki estymatora rozkładu kolorów dla modelu sieciowego z jednostanowymi pod modelami, T=4, D=2.

	sekwencja				liczba
elementy sumy	o_1	o_2	o_3	o_4	składowych
$\alpha_1^*(z)\beta_1^*(z)$, t=1	( )				**2**
	(	)			
$\alpha_1^*(z)\beta_1^*(z)$, t=2	( )				**4-1=3**
	(	)			
$\alpha_2^*(z)\beta_2^*(z)$, t=2		( )			
		(	)		
$-\alpha_1(z)\beta_1(z)$	( )				
$\alpha_2^*(z)\beta_2^*(z)$, t=3		( )			**4-2=2** 4-1=3
		(	)		
$\alpha_3^*(z)\beta_3^*(z)$, t=3			( )		
			(	)	
$-\alpha_2(z)\beta_2(z)$		( )			
		~~( )~~			
$\alpha_3^*(z)\beta_3^*(z)$, t=4			( )		**3-2=1** 3-1=2
			(	)	
$\alpha_4^*(z)\beta_4^*(z)$, t=4				( )	
$-\alpha_3(z)\beta_3(z)$			( )		
			~~( )~~		
liczba składowych	**2**	**3**	**3**-1	**2**-1	**=10**-2=**8**

Tabela 4.2. Współczynniki zmodyfikowanego estymatora rozkładu kolorów dla modelu sieciowego z jednostanowymi pod modelami, T=4, D=2.

Korzystając z przykładu w tabelach (4.1), (4.2) zależność estymatora (4.18) można zapisać w postaci podanej w zależności (4.19).

$$\widehat{\pi}_{ki}^{z}=\frac{\sum\limits_{(po\quad t\quad t.\dot{z}.o_t=k)}\left[\sum\limits_{\substack{\tau=t-D+1\\ \tau\geq 1}}^{t}\alpha_\tau^*(z)\beta_\tau^*(z)-\sum\limits_{\substack{\tau=t-D+1\\ \tau\geq 1}}^{t-1}\alpha_\tau(z)\beta_\tau(z)+\sum\limits_{\substack{\tau=t-D+1\\ \tau\geq 1}}^{t-1}\sum\limits_{\substack{d=\tau-t+D+1\\ d\leq\tau}}^{D}\alpha_{\tau-d+1}^*(z)p_z(d)\beta_\tau(z)\prod\limits_{t'=\tau-d+1}^{\tau}\pi_{o_{t'},1}^{z}\right]}{\sum\limits_{k=1}^{n}\sum\limits_{(po\quad t\quad t.\dot{z}.o_t=k)}\left[\sum\limits_{\substack{\tau=t-D+1\\ \tau\geq 1}}^{t}\alpha_\tau^*(z)\beta_\tau^*(z)-\sum\limits_{\substack{\tau=t-D+1\\ \tau\geq 1}}^{t-1}\alpha_\tau(z)\beta_\tau(z)+\sum\limits_{\substack{\tau=t-D+1\\ \tau\geq 1}}^{t-1}\sum\limits_{\substack{d=\tau-t+D+1\\ d\leq\tau}}^{D}\alpha_{\tau-d+1}^*(z)p_z(d)\beta_\tau(z)\prod\limits_{t'=\tau-d+1}^{\tau}\pi_{o_{t'},1}^{z}\right]} \tag{4.19}$$

Zależność estymatora (4.19) nie jest równoważna zależności estymatora (6.80) podanej w pracy [3] na stronie (361) dla modelu z czasem przebywania ponieważ nie występuje w niej dodatkowy czynnik występujący w zależności (4.19). Dodatkowy czynnik zaznaczony jest na przykładzie w tabelce (4.2) przez skreślenie. Dodatkowy czynnik z zależności (4.19) nie może być wyrażony przez iloczyny $\alpha_t^*(i)\beta_t^*(i)$ oraz $\alpha_t(i)\beta_t(i)$.

W podrozdziale tym zostało wykazane, że estymator dla modelu sieciowego nie sprowadza się do estymatora modelu z czasem przebywania z powodu odmiennej definicji zależności na estymator rozkładów generowania kolorów. W zależności estymatora podanej w pracy [3] w estymatorze nie zostały uwzględnione pewne czynniki w sumie po czasie *t* z powodu mniejszej złożoności obliczeniowej zaproponowanego estymatora.

4.4. Sprowadzenie estymatora modelu sieciowego do estymatora modelu podstawowego

Model sieciowy sprowadza się do modelu podstawowego HMM w dwóch przypadkach. W pierwszym przypadku, jeżeli w modelu sieciowym jest tylko jeden węzeł, oraz jeżeli prawdopodobieństwo pozostawania w tym węźle przyjmie wartość 1 dla czasu pozostawania równego *T*. Model sieciowy jest również równoważny modelowi podstawowemu HMM, jeśli w każdym węźle jest jedno-stanowy model HMM oraz prawdopodobieństwo pozostawania w węźle jest równe jeden dla czasu równego jeden.

W podrozdziale tym zostanie rozpatrzony pierwszy przypadek, gdy w modelu sieciowym jest tylko jeden węzeł. Przy założeniu, że $p_1(T)=1$ oraz $p_1(d)=0$ dla $d\neq T$ prawdopodobieństwo przejść w sieci zewnętrznej modelu sieciowego wynosi 1. W modelu podczas estymacji parametrów nie będzie podlegał zmianom rozkład prawdopodobieństwa przebywania.

Wyrażenia na estymator parametrów jednego pod modelu redukują się do zależności na estymator dla modelu podstawowego jak w zależnościach (4.20).

$$\hat{\delta}_i^1 = \frac{\sum_{t=1}^{T}\sum_{d=1}^{T}\left\{\alpha_t^*(1)p_1(d)\beta_{t+d-1}(1)\left[\alpha_1^{1(t,d)}(i)\beta_1^{1(t,d)}(i)\right]\right\}}{\sum_{t=1}^{T}\sum_{d=1}^{T}\left\{\alpha_t^*(1)p_1(d)\beta_{t+d-1}(1)\left[\sum_{i=1}^{m}\alpha_1^{1(t,d)}(i)\beta_1^{1(t,d)}(i)\right]\right\}} = \frac{\alpha_1^{1(1,T)}(i)\beta_1^{1(1,T)}(i)}{\sum_{i=1}^{m}\alpha_1^{1(1,T)}(i)\beta_1^{1(1,T)}(i)}$$

$$\hat{\gamma}_{ij}^1 = \frac{\sum_{t=1}^{T-1}\sum_{d=1}^{T}\left\{\alpha_t^*(1)p_1(d)\beta_{t+d-1}(1)\left[\sum_{t'=t}^{t+d-1}\alpha_{t'}^{1(t,d)}(i)\gamma_{ij}\pi_{o_{t+1}j}\beta_{t'+1}^{1(t,d)}(j)\right]\right\}}{\sum_{t=1}^{T-1}\sum_{d=1}^{T}\left\{\alpha_t^*(1)p_1(d)\beta_{t+d-1}(1)\left[\sum_{t'=t}^{t+d-1}\alpha_{t'}^{1(t,d)}(i)\beta_{t'}^{1(t,d)}(i)\right]\right\}} = \frac{\sum_{t'=1}^{T-1}\alpha_{t'}^{1(1,T)}(i)\gamma_{ij}\pi_{o_{t+1}j}\beta_{t'+1}^{1(1,T)}(j)}{\sum_{t'=1}^{T-1}\alpha_{t'}^{1}(i)\beta_{t'}^{1}(i)}$$

$$\hat{\pi}_{ki}^1 = \frac{\sum_{t=1}^{T}\sum_{d=1}^{T}\left\{\alpha_t^*(1)p_1(d)\beta_{t+d-1}(1)\left[\sum_{(po\ \ t'\ \ t.\dot{z}.o_{t'}=k)}^{t+d-1}\alpha_{t'}^{1(t,d)}(i)\beta_{t'}^{1(t,d)}(i)\right]\right\}}{\sum_{t=1}^{T}\sum_{d=1}^{T}\left\{\alpha_t^*(1)p_1(d)\beta_{t+d-1}(1)\left[\sum_{t'=t}^{t+d-1}\alpha_{t'}^{1(t,d)}(i)\beta_{t'}^{1(t,d)}(i)\right]\right\}} = \frac{\sum_{(po\ \ t'\ \ t.\dot{z}.o_{t'}=k)}^{T}\alpha_{t'}^{1(1,T)}(i)\beta_{t'}^{1(1,T)}(i)}{\sum_{t'=1}^{T}\alpha_{t'}^{1(1,T)}(i)\beta_{t'}^{1(1,T)}(i)} \qquad (4.20)$$

Redukcja podana w zależnościach (4.20) zachodzi, ponieważ prawdopodobieństwo pozostawania wynosi 0 dla $d \neq T$. Suma po parametrach t i d redukuje się do jednego elementu dla $t=1$ oraz $d=T$, stąd $\alpha_1(1)=1$ $\beta_T(1)=1$ oraz $p_1(T)=1$.

4.5. Skalowanie w modelu sieciowym

Podobnie jak w modelu podstawowym w modelu sieciowym można zastosować skalowanie. Współczynniki $\alpha_t(i)$, $\alpha^*_t(i)$, $\beta_t(i)$, $\beta^*_t(i)$ składają się z sumy dużej liczby wyrażeń, gdzie każde jest w postaci iloczynu prawdopodobieństw.

Współczynniki skalowania c_t wyznaczane są przy obliczaniu tablicy $\alpha_t(i)$. Wyznaczone współczynniki następnie wykorzystywane są przy wyliczaniu tablicy $\beta^*_t(i)$ podobnie jak w modelu podstawowym. Przy wyliczaniu tablic $\alpha^*_t(i)$, $\beta_t(i)$ skalowanie nie jest konieczne, ponieważ wyznaczanie tych tablic bazuje na już skalowanych tablicach $\alpha_t(i)$, $\beta^*_t(i)$.

Należy oznaczyć przez $\alpha_t(i)$ współczynnik przed skalowaniem oraz przez $\hat{\alpha}_t(i)$ współczynnik po skalowaniu dodatkowo należy oznaczyć przez $\hat{\hat{\alpha}}_t(i)$ lokalny współczynnik przed skalowaniem. Najpierw należy wyliczyć $\alpha_1(i)$ dla $t=1$, według równania podstawowego

oraz przyjąć, że $\hat{\bar{\alpha}}_1(i) = \alpha_1(i)$, oraz $\hat{\alpha}_1(i) = c_1\alpha_1(i)$ gdzie: $c_1 = \frac{1}{\sum_{i=1}^{m}\alpha_1(i)}$. Ponieważ w modelu sieciowym przy wyliczaniu współczynników $\hat{\bar{\alpha}}_t(i)$ korzysta się z współczynników od $\hat{\alpha}_{t-d+1}(j)$ do $\hat{\alpha}_{t-1}(i)$ stąd podczas wyznaczania współczynników przed skalowaniem należy zastosować zależność (4.21), w której przeprowadza się wymnażanie przez dodatkowe iloczyny współczynników skali.

$$\begin{aligned}\hat{\bar{\alpha}}_t(j) &= \left(\prod_{t'=1}^{t-1} c_{t'}\right)\delta_j p_j(t)P(o_1,o_2,...,o_t \mid M^i) \\ &+ \sum_{i=1}^{m}\sum_{d=1}^{t-1}\left(\prod_{t'=t-d+1}^{t-1} c_{t'}\right)\hat{\alpha}_{t-d}(i)\gamma_{ij}p_j(d)P(o_{t-d+1},...,o_t \mid M^i), \qquad dla \quad t \le D \qquad (4.21)\\ \hat{\bar{\alpha}}_t(j) &= \sum_{i=1}^{m}\sum_{d=1}^{D}\left(\prod_{t'=t-d+1}^{t-1} c_{t'}\right)\hat{\alpha}_{t-d}(i)\gamma_{ij}p_j(d)P(o_{t-d+1},...,o_t \mid M^i), \qquad dla \quad t > D\end{aligned}$$

Następnie należy wyznaczyć współczynnik skali c_t, tak samo jak dla modelu podstawowym korzystając z zależności (4.22).

$$c_t = \frac{1}{\sum_{i=1}^{N}\hat{\bar{\alpha}}_t(i)} \qquad (4.22)$$

Skalowanie $\alpha_t(i)$ przeprowadza się według zależności (4.23).

$$\hat{\alpha}_t(i) = c_t\hat{\bar{\alpha}}_t(i) \qquad (4.23)$$

Podczas skalowania $\beta^*_t(i)$ należy użyć tej samej skali, która była wyznaczona przy wyliczaniu współczynnika $\alpha_t(i)$, stąd skalowanie przybiera postać (4.24).

$$\hat{\beta}^*_t(i) = c_t\hat{\bar{\beta}}^*_t(i) \qquad (4.24)$$

Współczynniki $\alpha_t(i)$ po skalowaniu można zapisać jak w równaniu (4.25).

$$\hat{\alpha}_t(i) = \left(\prod_{k=1}^{t} c_k\right)\alpha_1(i) = C_t\alpha_t(i) \qquad (4.25)$$

Współczynniki $\beta^*_t(i)$ po skalowaniu można zapisać jak w równaniu (4.26).

$$\hat{\beta}^*_{t+1}(j) = \left(\prod_{k=t+1}^{T} c_k\right)\beta^*_{t+1}(j) = D_{t+1}\beta^*_{t+1}(j) \qquad (4.26)$$

W przypadku użycia procedury skalowania do modelu sieciowego konieczna jest również modyfikacja zależności na estymatory parametrów modelu. Estymator prawdopodobieństwa początkowych, prawdopodobieństw przejść nie wymaga modyfikacji, natomiast ulegają zmianie zależności na estymatory prawdopodobieństw pozostawania oraz estymatory parametrów pod modeli HMM w węzłach. Estymator parametrów przejść w sieci przyjmuje postać (4.27).

$$\bar{\gamma}_{ij} = \frac{\sum_{t=1}^{T-1}\hat{\alpha}_t(i)\gamma_{ij}\hat{\beta}^*_{t+1}(j)}{\sum_{j=1}^{m}\sum_{t=1}^{T-1}\hat{\alpha}_t(i)\gamma_{ij}\hat{\beta}^*_{t+1}(j)} = \frac{\sum_{t=1}^{T-1}C_t\alpha_t(i)\gamma_{ij}D_{t+1}\beta^*_{t+1}(j)}{\sum_{j=1}^{m}\sum_{t=1}^{T-1}C_t\alpha_t(i)\gamma_{ij}D_{t+1}\beta^*_{t+1}(j)} = \frac{\sum_{t=1}^{T-1}\alpha_t(i)\gamma_{ij}\beta^*_{t+1}(j)}{\sum_{j=1}^{m}\sum_{t=1}^{T-1}\alpha_t(i)\gamma_{ij}\beta^*_{t+1}(j)} \qquad (4.27)$$

Powyższa procedura nie wymaga modyfikacji, ponieważ następuje skrócenie się czynnika $C_t D_{t+1}$. Estymator prawdopodobieństw pozostawania przyjmuje postać (4.28).

$$\hat{p}_i(d) = \frac{\sum_{t=1}^{T-d+1}\left(\prod_{\tau=t}^{t+d-1} c_\tau\right)\hat{\alpha}_t^*(i)p_i(d)\hat{\beta}_{t+d-1}(i)P(o_t,..,o_{t+d-1}\mid M^i)}{\sum_{d=1}^{D}\sum_{t=1}^{T-d+1}\left(\prod_{\tau=t}^{t+d-1} c_\tau\right)\hat{\alpha}_t^*(i)p_i(d)\hat{\beta}_{t+d-1}(i)P(o_t,..,o_{t+d-1}\mid M^i)} =$$

$$= \frac{\sum_{t=1}^{T-d+1}\left(\prod_{\tau=t}^{t+d-1} c_\tau\right)C_t\alpha_t^*(i)p_i(d)D_{t+d-1}\beta_{t+d-1}(i)P(o_t,..,o_{t+d-1}\mid M^i)}{\sum_{d=1}^{D}\sum_{t=1}^{T-d+1}\left(\prod_{\tau=t}^{t+d-1} c_\tau\right)C_t\alpha_t^*(i)p_i(d)D_{t+d-1}\beta_{t+d-1}(i)P(o_t,..,o_{t+d-1}\mid M^i)} \quad (4.28)$$

$$= \frac{\sum_{t=1}^{T-d+1}\alpha_t^*(i)p_i(d)\beta_{t+d-1}(i)P(o_t,..,o_{t+d-1}\mid M^i)}{\sum_{d=1}^{D}\sum_{t=1}^{T-d+1}\alpha_t^*(i)p_i(d)\beta_{t+d-1}(i)P(o_t,..,o_{t+d-1}\mid M^i)}$$

Dlatego w procedurach na prawdopodobieństwa pozostawania oraz parametry pod modeli w sieci konieczne jest wprowadzenie dodatkowego iloczynu $\left(\prod_{\tau=t}^{t+d-1} c_\tau\right)$. Iloczyn należy wprowadzić w celu zredukowania współczynników skalowania.

Podobnie jak dla modelu podstawowego w przypadku skalowania zmienia się sposób wyznaczania prawdopodobieństwa $P(\mathbf{O}|M)$ nie można wyznaczyć $P(\mathbf{O}|\mathrm{M})$, ale $\log(P(\mathbf{O}|\mathrm{M}))$.

4.6. Uczenie modelu sieciowego do wielokrotnych sekwencji obserwacji

W tym rozdziale zostaną podane zależności pozwalające optymalizować sieciowy model HMM od wielokrotnych sekwencji obserwacji. Metody estymacji modelu podstawowego HMM do wielokrotnych sekwencji obserwacji zostały opisane w rozdziale 3. Podobnie jak dla modelu podstawowego przez K została oznaczona liczba sekwencji obserwacji. $\mathbf{O} = \{\mathbf{O}^{(1)}, \mathbf{O}^{(2)}, ..., \mathbf{O}^{(K)}\}$

gdzie: $\mathbf{O}^{(k)} = (o_1^{(k)}, o_2^{(k)}, ..., o_{T_k}^{(k)})$ jest k tą sekwencją obserwacji o długości T_k. Należy założyć, że każda sekwencja obserwacji jest niezależna od innych sekwencji obserwacji. Jako funkcja celu podczas optymalizacji zostało przyjęte wyrażenie (4.29).

$$P(\mathbf{O}\mid \mathrm{M}) = \prod_{k=1}^{K} P(\mathbf{O}^{(k)}\mid \mathrm{M}) = \prod_{k=1}^{K} P_k \quad (4.29)$$

Ponieważ formuła estymacji bazuje na częstości występowania różnych obserwacji, formuła dla wielu sekwencji obserwacji jest zmodyfikowana poprzez dodanie razem poszczególnych częstości wystąpienia różnych sekwencji. Zmodyfikowane formuły dla parametrów sieci δ_i, γ_{ij}, $p_i(d)$ są podane w zależnościach (4.30).

$$\hat{\delta}_i = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \delta_i \beta_1^*(i)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{i=1}^{m} \alpha_T(i)} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \delta_i \beta_1^*(i)}{K}$$

$$\hat{\gamma}_{ij} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t(i) \gamma_{ij} \beta_{t+1}^*(j)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t(i) \beta_t(i)} \tag{4.30}$$

$$\hat{p}_i(d) = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-d+1} \alpha_t^*(i) p_i(d) \beta_{t+d-1}(i) P(o_t,...,o_{t+d-1} \mid M^i)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{d=1}^{D} \sum_{t=1}^{T_k-d+1} \alpha_t^*(i) p_i(d) \beta_{t+d-1}(i) P(o_t,...,o_{t+d-1} \mid M^i)}$$

Procedury estymacji dla parametrów pod modeli δ_i^z, γ_{ij}^z, $\pi_{S_t i}^z$ przyjmują postać podaną w zależnościach (4.31).

$$\hat{\delta}_i^z = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k} \sum_{d=1}^{D'} \left\{ \alpha_t^*(z) p_z(d) \beta_{t+d-1}(z) \left[\alpha_1^{z(t,d)}(i) \beta_1^{z(t,d)}(i) \right] \right\}}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k} \sum_{d=1}^{D'} \left\{ \alpha_t^*(z) p_z(d) \beta_{t+d-1}(z) \left[\sum_{i=1}^{m} \alpha_1^{z(t,d)}(i) \beta_1^{z(t,d)}(i) \right] \right\}}$$

$$\hat{\gamma}_{ij}^z = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k} \sum_{d=1}^{D'} \left\{ \alpha_t^*(z) p_z(d) \beta_{t+d-1}(z) \left[\sum_{t'=1}^{d-1} \alpha_{t'}^{z(t,d)}(i) \gamma_{ij} \pi_{s_{t+1} j} \beta_{t'+1}^{z(t,d)}(j) \right] \right\}}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k} \sum_{d=1}^{D'} \left\{ \alpha_t^*(z) p_z(d) \beta_{t+d-1}(z) \left[\sum_{t'=1}^{d-1} \alpha_{t'}^{z(t,d)}(i) \beta_{t'}^{z(t,d)}(i) \right] \right\}}$$

$$\hat{\pi}_{ki}^z = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k} \sum_{d=1}^{D'} \left\{ \alpha_t^*(z) p_z(d) \beta_{t+d-1}(z) \left[\sum_{t'=1}^{d} \alpha_{t'}^{z(\tau,d)}(i) \beta_{t'}^{z(\tau,d)}(i) \right] \right\}}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k} \sum_{d=1}^{D'} \left\{ \alpha_t^*(z) p_z(d) \beta_{t+d-1}(z) \left[\sum_{t'=1}^{d} \alpha_{t'}^{z(t,d)}(i) \beta_{t'}^{z(t,d)}(i) \right] \right\}}$$

$$\left(po \quad t' \quad t.\dot{z}.\, s_{t'} = k \right)$$

gdzie:

$$D' = D \; dla \; T_k - t \geq D$$
$$D' = T_k - t + 1 \; dla \; T_k - t < D \tag{4.31}$$

W przypadku skalowania konieczne jest wymnażanie przez czynnik $\left(\prod_{\tau=t}^{t+d-1} c_\tau \right)$ zależności w estymatorach, w których występuje prawdopodobieństwo $P_z(d)$. W przypadku skalowania w estymatorach dla wielokrotnych sekwencji obserwacji należy opuścić czynnik $1/P_k$ podobnie jak dla modelu podstawowego.

5. Szacowanie złożoności obliczeniowej wyznaczania prawdopodobieństwa obserwacji oraz estymacji parametrów modelu HMM

W tym rozdziale zostały umieszczone wyniki oszacowania złożoność obliczeniowej oraz ilość wymaganej pamięci operacyjnej dla zaimplementowanych w ramach pracy algorytmów opartych na kryterium (ML). Złożoność została wyznaczona dla modelu podstawowego HMM, modelu HMM z jawną gęstością pozostawania w danym stanie oraz modelu sieciowego HMM.

5.1. Szacowanie złożoności dla modelu podstawowego HMM

Model podstawowy posiada m stanów, n kolorów możliwych do wylosowania w każdym stanie oraz długość sekwencji wynosi T.

Obliczanie Prawdopodobieństwa obserwacji:

Wymagana pamięć operacyjna:
Tablice $\alpha_t(i)$, $\beta_t(i)$, wektor skalowania $c(t)$: $2Tm+T$ jednostek pamięci

Ilość działań:

Wypełnianie tablic $\alpha_t(i)$, $\beta_t(i)$:	$(T\text{-}1)(3m+1)+m$ iloczyn, $2(T\text{-}1)(m\text{-}1)$ suma
Skalowanie $c(t)$:	$T(m\text{-}1)$ suma, Tm iloraz
Prawdopodobieństwa $P(\mathbf{O}\|M)$:	m-1 suma

Złożoność obliczeniowa wyznaczenia prawdopodobieństwa wynosi Tm.
Złożoność pamięci wynosi Tm.

Estymacja parametrów modelu:

Dodatkowa wymagana pamięć operacyjna:

Mianowniki estymatorów γ_{ij}, oraz π_{ki}:	$2m$ jednostek pamięci

Dodatkowa ilość działań:

Prawdopodobieństwa początkowe δ_i:	m iloczyn, m iloraz
Prawdopodobieństwa przejść γ_{ij}:	$3(T\text{-}1)m^2$ iloczyn, $(T\text{-}1)m^2$ suma, m^2 iloraz
Prawdopodobieństwa kolorów π_{ki}:	Tm iloczyn, $(T\text{-}n)m$ suma, mn iloraz
Mianowniki estymatorów γ_{ij}, oraz π_{ki}:	$m(m\text{-}1)$ suma, $m(n\text{-}1)$ suma

Złożoność obliczeniowa procesu estymacji parametrów wynosi Tm^2.
Złożoność pamięci wynosi m.

5.2. Szacowanie złożoności dla modelu z jawnym rozkładem prawdopodobieństwa pozostawania w tym samym stanie

Model podstawowy posiada m stanów, n kolorów możliwych do wylosowania w każdym. Model może przebywać w jednym stanie przez maksymalnie D chwil czasowych. Długość sekwencji obserwacji wynosi T.

Obliczanie Prawdopodobieństwa obserwacji:

Wymagana pamięć operacyjna:

Tablice $\alpha_t(i)$, $\beta_t(i)$, wektor skalowania $c(t)$:	$4Tm+T$ jednostek pamięci
Współczynniki wagowe (iloczyny prawd. kolorów w stanie):	$\{TD-[(D-1)/2]D\}m$ jednostek pamięci

Ilość działań:

Współczynniki wagowe:	$\{[(D-1)/2]DT-(1/3D^3+1/2D^2+1/6D)\}m$ iloczyn
Wypełnianie tablic $\alpha_t(i)$ $\beta_t^*(i)$:	$2\{3[D^2-(D/2)(D+1)]m+3(T-D)D(2m-2)\}m+3mD$ iloczyn $2\{[D^2-(D/2)(D+1)](m-2)+(T-D)D(m-2)\}m+2mD$ suma
Skalowanie $c(t)$:	$T(m-1)$ suma, $2Tm$ iloraz
Wypełnienie tablic $\alpha_t^*(i)$ $\beta_t(i)$:	$2m^2T$ iloczyn, $2m^2T$ suma
Policzenie prawdopodobieństwa $P(\mathbf{O}\|M)$:	m suma

Złożoność obliczeniowa wyznaczenia prawdopodobieństwa wynosi Tm^2D, $m>D$, $T>D$
Złożoność pamięci wynosi TmD, $T>D$.

Estymacja parametrów modelu:

Dodatkowa wymagana pamięć operacyjna

Mianowniki estymatorów γ_{ij}, π_{ki} oraz $p_i(d)$:	$3m$ jednostek pamięci

Dodatkowa ilość działań:

Prawdopodobieństwa początkowe δ_i:	m iloczyn, m iloraz
Prawdopodobieństwa przejść γ_{ij}:	$2Tm^2$ iloczyn, $(T-1)m^2$ suma, m^2 iloraz
Prawdopodobieństwa kolorów π_{ki}:	$mn(T-n)2D$ iloczyn, $mn(T-n)2D$ suma, $mn(T-n)$ różnica, mn iloraz
Prawdopodobieństwa pozostawania $p_i(d)$:	$3TmD$ iloczyn, $(T-1)mD$, suma, mD dzieleń
Mianowniki estymatorów γ_{ij}, π_{ki} oraz $p_i(d)$:	$m(m-1)$ suma, $m(n-1)$ suma, $m(D-1)$ suma

Złożoność obliczeniowa procesu estymacji parametrów wynosi $TmnD$, $T>n$, $Dn>m$.
Złożoność pamięci wynosi m.

5.3. Szacowanie złożoności dla modelu sieciowego

Model sieciowy posiada z węzłów oraz pod modeli, każdy pod model posiada m stanów oraz istnieje n kolorów możliwych do wylosowania w każdym stanie pod modelu. Długość sekwencji wynosi T.

Obliczanie Prawdopodobieństwa obserwacji:

Wymagana pamięć operacyjna:
Tablice $\alpha_t(i)$, $\beta_t(i)$ wektor skalowania $c(t)$: $4Tz+T$ jednostek pamięci
Współczynniki pod modeli: $(2m+1)\{TD(D+1)/2-[1/3(D-1)^3+(D-1)^2+2/3(D-1)]\}z$

Ilość działań:
Współczynniki pod modeli: $m\{TD(D+1)/2-[1/3(D-1)^3+(D-1)^2+2/3(D-1)]\}z$
Wypełnianie tablic $\alpha_t(i)$ $\beta_t^*(i)$: $2\{3[D^2-(D/2)(D+1)]z+3(T-D)D(2z-2)\}z+3zD$ iloczyn
$2\{[D^2-(D/2)(D+1)](z-2)+(T-D)D(z-2)\}z+2zD$ suma
Skalowanie $c(t)$: $T(z-1)$ suma, $2Tz$ iloraz
Wypełnienie tablic $\alpha_t^*(i)$ $\beta_t(i)$: $2z^2T$ iloczyn, $2z^2T$ suma
Policzenie prawdopodobieństwa $P(\mathbf{O}|M)$: z suma

Stąd złożoność obliczeniowa wyznaczenia prawdopodobieństwa wynosi TmD^2z.
Złożoność pamięci wynosi TmD^2z.

Estymacja parametrów modelu:

Dodatkowa wymagana pamięć operacyjna:
Współczynniki pod modeli δ_i^z, γ_{ij}^z, π_{ki}^z: $(3m+m^2+mn)[(TD+D(D-1)/2)]z$ jednostek pam.
Prawdopodobieństwa δ_i^z, γ_{ij}^z, π_{ki}^z: $zm(3+m+n)$ jednostek pamięci
Mianowniki estymatorów γ_{ij}, oraz $p_i(d)$: $2z$ jednostek pamięci
Mianowniki estymatorów δ_i^z , γ_{ij}^z oraz π_{ki}^z: $z+2zm$ jednostek pamięci

Dodatkowa ilość działań:
Współczynniki pod modeli δ_i^z, γ_{ij}^z, π_{ki}^z: $m^2\{TD-[1/3(D-1)^3+(D-1)^2+2/3(D-1)]\}z$
Prawdopodobieństwa początkowe δ_i: z iloczyn z iloraz
Prawdopodobieństwa przejść γ_{ij}: $2(T-1)z^2$ iloczyn, $(T-2)z^2$ suma, z^2 iloraz
Prawdopodobieństwa pozostawania $p_i(d)$: $3z[(TD+D(D-1)/2)]$ iloczyn,
$z[(TD+(D-1)(D-2)/2)]$ suma, Dz iloraz
Mianowniki estymatorów γ_{ij}, oraz $p_i(d)$: $z(z-1)$ suma, $z(D-1)$ suma
Prawdopodobieństwa początkowe δ_i^z: $3TDzm$ iloczyn, $TDzm$ suma, zm iloraz
Prawdopodobieństwa przejść γ_{ij}^z: $3TDzm^2$ iloczyn, $TDzm^2$ suma, zm^2 iloraz
Prawdopodobieństwa pozostawania π_{ki}^z: $3TDzmn$ iloczyn $TDzmn$ suma, zmn iloraz
Mianowniki estymatorów δ_i^z, γ_{ij}^z oraz π_{ki}^z: $z(m-1)$ suma, $zm(m-1)$ suma, $zn(m-1)$ suma

Złożoność obliczeniowa procesu estymacji parametrów wynosi Tm^2Dz.
Złożoność pamięci wynosi Tm^2Dz.

W tabeli (5.1) przedstawione jest zestawienie złożoności oraz wymaganej pamięci dla omawianych w pracy modeli HMM. W tabeli przyjęto oznaczenia:

T– długość sekwencji obserwacji,

m– liczna stanów dla modeli standardowego oraz modelu z jawną gęstością prawdopodobieństwa, liczba węzłów dla modelu sieciowego,

n– ilość obserwacji (symboli) w stanie,

D– maksymalny czas pozostawania w jednym stanie dla modelu z jawna gęstością prawdopodobieństwa oraz maksymalny czas pozostawania w jednym węźle dla modelu sieciowego,

z– ilość stanów w pod modelach modelu sieciowego.

	Złożoność przy obliczaniu prawdopodobieństwa obserwacji	Wymagana pamięć operacyjna przy obliczaniu prawdopodobieństwa obserwacji	Złożoność przy estymacji parametrów modelu	Wymagana pamięć operacyjna przy estymacji parametrów modelu
Model podstawowy HMM	Tm	Tm	Tm^2	m
Model z czasem przebywania	Tm^2D	TmD	$TmnD$	m
Model sieciowy	TmD^2z	TmD^2z	Tm^2Dz	Tm^2Dz

Tabela 5.1. Złożoność obliczeniowa oraz wymagana pamięć operacyjna przy obliczaniu prawdopodobieństwa obserwacji oraz estymacji parametrów modeli HMM.

W celu estymacji parametrów należy najpierw wyznaczyć współczynniki α, β związane z wyznaczaniem prawdopodobieństwa sekwencji obserwacji. Dlatego w tabeli (5.1) podana jest dodatkowa wymagana pamięć operacyjna oraz złożoność obliczeniowa związana z estymacją parametrów po wcześniejszym obliczeniu prawdopodobieństwa sekwencji obserwacji oraz związanych z nim współczynników α, β.

6. Algorytmy sztucznej inteligencji

W rozdziale tym zostały przedstawione zaprojektowane przez autora pracy algorytmy sztucznej inteligencji wykorzystane do optymalizacji modeli HMM.

W praktycznych zastosowaniach do optymalizacji modeli HMM wykorzystywane są algorytmy polegające na poprawianiu już istniejącego rozwiązania. Algorytmy te zostały omówione w poprzednich rozdziałach. Przedstawione algorytmy zatrzymują się w krytycznym punkcie. Wartość znalezionego rozwiązania zależy od doboru parametrów początkowych modelu. Dlatego uzasadnione jest użycie algorytmów sztucznej inteligencji takich jak Algorytmy ewolucyjne oraz algorytmów przybliżonych takich jak „*Tabu Search*" do optymalizacji modeli HMM oraz porównanie wyników uzyskanych dzięki zastosowaniu tych algorytmów oraz omówionych w pracy algorytmów poprawiających.

6.1. Algorytm genetyczny

W opracowanym algorytmie tak jak w większości algorytmów genetycznych wykorzystane zostały podstawowe procedury tj.: reprodukcja, krzyżowanie oraz mutacja. Zaproponowany algorytm działa krokowo. W pierwszym kroku tworzona jest populacja początkowa, w której każdy osobnik jest reprezentowany przez model HMM zakodowany w postaci chromosomu, w każdym następnym kroku algorytmu w procedurze reprodukcji z krzyżowaniem tworzone są osobniki potomne, które są następnie poddawane procedurze mutacji. Z populacji początkowej oraz osobników potomnych w procedurze selekcji tworzona jest nowa populacja. W procedurach reprodukcji z krzyżowaniem oraz selekcji dokonywany jest wybór osobników ze względu na wartość ich funkcji przystosowania.

W podrozdziałach (6.1.1)÷(7.1.7) opisane jest tworzenie nowej populacji, algorytm konstrukcyjny, kodowanie chromosomu, funkcja oceny, procedury reprodukcji i krzyżowania oraz procedura mutacji.

6.1.1. Tworzenie nowej populacji

Nowe modele HMM populacji generowane są w sposób losowy, ale mogą być też dodawane modele wcześniej poddane optymalizacji przy pomocy algorytmu Bauma-Welcha opisanego w rozdziale (3). W algorytmie ustalona jest liczba rozwiązań w populacji, jest ona oznaczona przez x. Rozwiązania losowe tworzone są z wykorzystaniem algorytmu konstrukcyjnego.

6.1.2. Algorytm konstrukcyjny

Losowe rozwiązania początkowe zarówno w algorytmie Bauma-Welcha oraz w algorytmach przybliżonych są tworzone w następujący sposób. W programie ustalona jest wartość całkowita L_{max}. Parametrom modelu przypisywane są liczby całkowite z przedziału

<0,L_{max}>. Po przypisaniu wartości całkowitych parametrom modelu HMM przeprowadzana jest procedura normalizacji jak w równaniach (6.1), (6.2), (6.3).

$$\delta_i = \frac{\overline{\delta}_i}{\sum_{l=1}^{m} \overline{\delta}_l}, \qquad i = 1,2,...,m \tag{6.1}$$

$$\gamma_{ij} = \frac{\overline{\gamma}_{ij}}{\sum_{l=1}^{m} \overline{\gamma}_{il}}, \qquad i,j = 1,2,...,m \tag{6.2}$$

$$\hat{\pi}_{kj} = \frac{\overline{\pi}_{s_k j}}{\sum_{l=1}^{n} \overline{\pi}_{lj}}, \qquad j = 1,2,...,m, \quad k = 1,2,...,n \tag{6.3}$$

W wyniku powyższej metody uzyskuje się model HMM, którego wagi mogą przyjmować wartość prawdopodobieństwa z przedziału od 0 do 1 z krokiem zależnym od wartości parametru L_{max}.

6.1.3. Kodowanie chromosomów

Kodowanie chromosomów osobnika M przeprowadzone jest w sposób pokazany w zależności (6.4).

$$M = \left(\delta_i \quad \gamma_{1j} \mid \gamma_{2j} \quad \ldots \quad \gamma_{mj} \quad \pi_{k1} \mid \pi_{k2} \quad \ldots \quad \pi_{km}\right) \tag{6.4}$$

i,j=1,2,..,m k=1,2,..,n

Chromosom reprezentowany jest przez wektor prawdopodobieństw początkowych $\{\delta_i\}$, wektory prawdopodobieństw przejść z węzła pierwszego $\{\gamma_{1j}\}$, węzła drugiego $\{\gamma_{2j}\}$ aż do węzła m-tego $\{\gamma_{mj}\}$ oraz wektory prawdopodobieństw wylosowania kolorów w węźle pierwszym $\{\pi_{k1}\}$, węźle drugim $\{\pi_{k2}\}$ aż do węzła m-tego $\{\pi_{km}\}$, gdzie: i,j=1,2,...,m, k=1,2,..n.

6.1.4. Funkcja oceny

W każdym kroku algorytmu ewolucyjnego przeprowadzana jest ocena osobników populacji oraz osobników potomków. Wartość funkcji oceny została oznaczona przez $Q(M)$. Zastosowane zostały trzy metody oceny.

W pierwszej metodzie wybierana jest losowo sekwencja **O** reprezentująca sylabę, następnie liczona jest wartość funkcji oceny jako prawdopodobieństwo wygenerowania wylosowanej sekwencji jak w zależności (6.5).

$$Q(M) = P(rand(\mathbf{O}_i) \mid M), \qquad i = 1,2,...,z, \tag{6.5}$$

W drugiej metodzie wartość funkcji oceny przyjmuje minimalną wartość prawdopodobieństwa jak w zależności (6.6). Prawdopodobieństwa są wyznaczane dla wszystkich sekwencji uczących 1,2,…,z.

$$Q(M) = \min P(\mathbf{O}_i \mid M), \qquad i = 1,2,...,z \tag{6.6}$$

W trzeciej metodzie oceny wartość funkcji wyliczana dla osobnika jest iloczynem logarytmów prawdopodobieństw wygenerowania poszczególnych sekwencji ze zbioru sekwencji uczących jak w zależności (6.7).

$$Q(M) = \sum_{i=1}^{z} \log P(\mathbf{O}_i \mid M) \tag{6.7}$$

6.1.5. Reprodukcja z krzyżowaniem

Modele HMM reprezentujące rodziców wybierane są w sposób częściowo losowy, ale z uwzględnieniem wartości funkcji oceny. Z populacji wybierane są dwa rozwiązani w sposób losowy. Pierwsze o lepszej wartości funkcji przystosowania brane jest jako pierwszy rodzic. Proces ten jest powtarzany w celu wyodrębnienia drugiego rodzica.

Rozwiązania potomków są tworzone z wykorzystaniem rozwiązań rodziców według standardowej procedury krzyżowania w wyniku, której uzyskuje się *y* rozwiązań potomnych. W algorytmie została zastosowana dwupunktowa operacja krzyżowania, w której dwa punkty krzyżowania wybierane są w sposób losowy. Pierwszy potomek tworzony jest z genów rodzica pierwszego na prawo od pierwszego punktu krzyżowania oraz na lewo od drugiego punktu krzyżowania oraz z genów drugiego z rodziców, które są pomiędzy dwoma punktami krzyżowania. Drugi potomek tworzony jest w ten sam sposób tylko rodzice brani są w odwrotnej kolejności. W zależnościach (6.8), (6.9) pokazany jest przykład dwóch rodziców wykorzystanych do operacji krzyżowana. Punkty krzyżowania są wyznaczone w sposób losowy na przykładzie są one zaznaczone przez trzy punkty w pionie.

$$M = \left(\delta_i \quad \gamma_{1j} \vdots \gamma_{2j} \quad \ldots \quad \gamma_{mj} \quad \pi_{o_k 1} \vdots \pi_{o_k 2} \quad \ldots \quad \pi_{o_k m}\right) \tag{6.8}$$

$$\widetilde{M} = \left(\widetilde{\delta}_i \quad \widetilde{\gamma}_{1j} \vdots \widetilde{\gamma}_{2j} \quad \ldots \quad \widetilde{\gamma}_{mj} \quad \widetilde{\pi}_{o_k 1} \vdots \widetilde{\pi}_{o_k 2} \quad \ldots \quad \widetilde{\pi}_{o_k m}\right) \tag{6.9}$$

Pierwszy punkt krzyżowania umieszczony jest pomiędzy wektorem prawdopodobieństw przejść z pierwszego węzła a wektorem prawdopodobieństw przejść z drugiego węzła, drugi pomiędzy wektorem wylosowania kolorów w pierwszym węźle a wektorem wylosowania kolorów w drugim węźle.

W wyniku krzyżowani otrzymuje się dwa osobniki potomków M' oraz $\widetilde{M}'$. Przykład dwóch potomków uzyskanych w wyniku operacji krzyżowania wykonanej na rodzicach z zależności (6.8), (6.9) pokazany jest w zależnościach (6.10), (6.11).

$$M' = \left(\delta_i \quad \gamma_{1j} \vdots \widetilde{\gamma}_{2j} \quad \ldots \quad \widetilde{\gamma}_{mj} \quad \widetilde{\pi}_{o_k 1} \vdots \pi_{o_k 2} \quad \ldots \quad \pi_{o_k m}\right) \tag{6.10}$$

$$\widetilde{M}' = \left(\widetilde{\delta}_i \quad \widetilde{\gamma}_{1j} \vdots \gamma_{2j} \quad \ldots \quad \gamma_{mj} \quad \pi_{o_k 1} \vdots \widetilde{\pi}_{o_k 2} \quad \ldots \quad \widetilde{\pi}_{o_k m}\right) \tag{6.11}$$

W wyniku operacji krzyżowania otrzymuje się parzystą liczbę potomków.

6.1.6. Mutacja

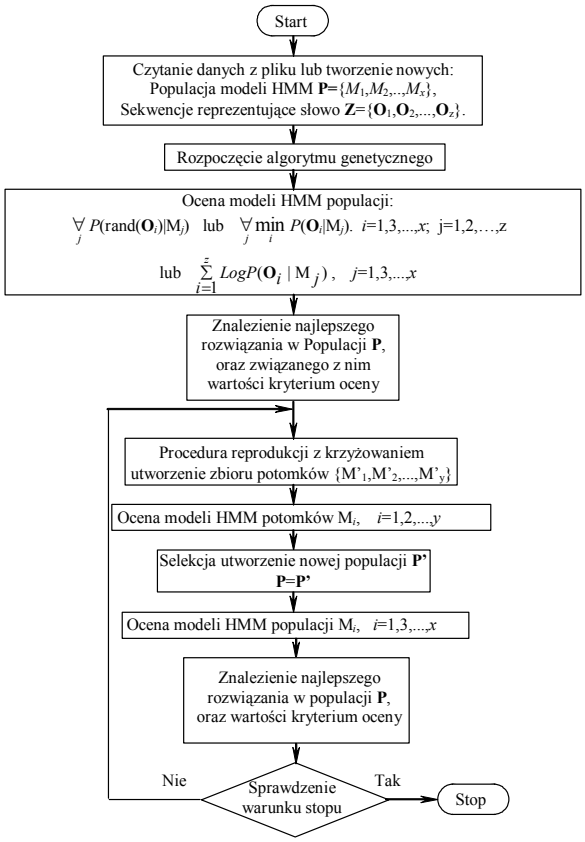
Operacja mutacji przeprowadzana jest w każdym kroku algorytmu. Mutacji poddawane są rozwiązana potomków. W algorytmie przyjęto parametr mutacji z zakresu 0.001 do 0.5 określający prawdopodobieństwo, z jakim poddawana jest modyfikacji każda składowa genu odpowiadająca wadze modelu HMM. Według przyjętego prawdopodobieństwa parametry

modeli tj. prawdopodobieństwa początkowe δ_i, przejść γ_{ij}, oraz kolorów π_{oi} poddawane są modyfikacji. Losowana jest liczba z przedziału 0 do L_{max} następnie parametr poddawany modyfikacji przyjmuje wylosowaną wartość. Jeżeli nastąpiła zmiana parametrów modelu stosowana jest normalizacji jak w zależnościach (6.1)÷(6.3).

6.1.7. Selekcja

W procedurze selekcji tworzona jest nowa populacja. Osobniki nowej populacji mogą pochodzić zarówno ze zbioru potomków jak i populacji początkowej lub tylko ze zbioru potomków. Grupę osobników, z której tworzona jest nowa populacja nazwiemy grupa bazową. W procesie selekcji wykorzystywane są dwie metody wyboru osobników do nowej populacji z grupy bazowej. Z grupy bazowej mogą być wybierane najlepsze osobniki według zastosowanej wartości funkcji oceny bądź stosowana jest metoda grupowania osobników według wartości funkcji przystosowania. Dzięki zastosowaniu pierwszej metody zostają zabezpieczone dobre wartości rozwiązań, ale może ona powodować lokalną zbieżność algorytmu. W drugiej metodzie wyboru stosowane jest grupowanie osobników. Osobniki z grupy bazowej dzielone są na trzy podgrupy, przynależność do każdej z grup zależy od wartości funkcji przystosowania, na przykład najlepiej przystosowane 40% mniej przystosowane 40% oraz najgorzej przystosowane 20%. Po grupowaniu osobniki do nowej populacji są wybierane z każdej grupy w odpowiednich z góry określonych proporcjach np. 40% z pierwszej grupy 40% z drugiej grupy, 20% z trzeciej grupy. Celem zastosowania drugiej metody jest utrzymanie różnorodności jak i dobrej jakości rozwiązań.

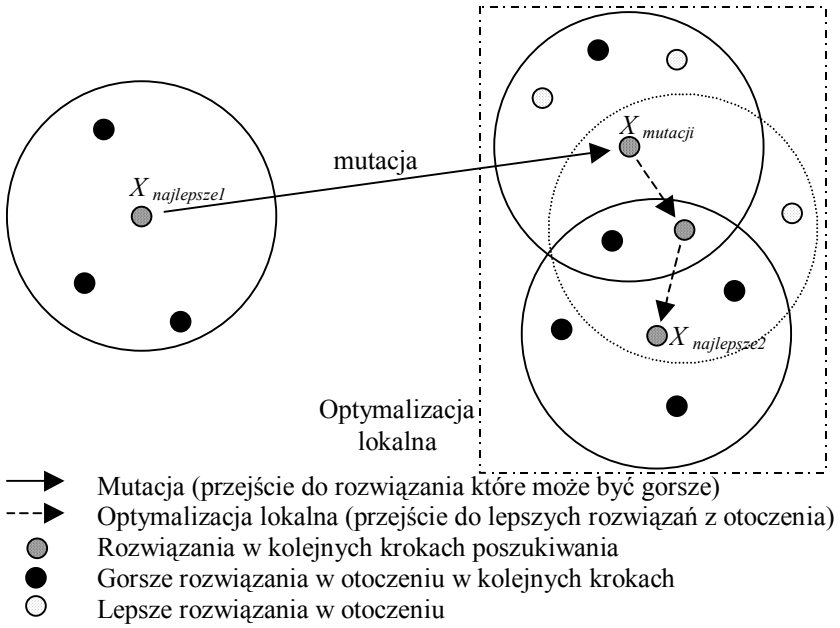
Na rysunku (6.1) przedstawiony jest schemat blokowy zastosowanego algorytmu genetycznego.



Rysunek 6.1. Schemat blokowy algorytmu genetycznego zastosowanego do optymalizacji modeli HMM.

6.2. Algorytm iteracyjny

Do optymalizacji modeli sieciowych został zaimplementowany przybliżony algorytm iteracyjny. W algorytmie ILS jest wykorzystana zasada poprawiania rozwiązania startowego wyznaczonego za pomocą algorytmu konstrukcyjnego. Proces poprawy przebiega do chwili zadziałania warunku stopu. Rozwiązanie startowe najpierw jest poprawiane lokalnie w wyniku optymalizacji lokalnej uzyskuje się rozwiązanie lokalnie optymalne $x_{najlepsze1}$. Rozwiązanie lokalnie optymalne jest mutowane i ponownie optymalizowane lokalnie. Z dwóch rozwiązań lokalnie optymalnych $x_{najlepsze1}$ oraz $x_{najlepsze2}$. jest wybierane lepsze i proces optymalizacji jest powtarzany, co przedstawia rysunek (6.2).



Rysunek 6.2. Zasada działania algorytmu ILS

Algorytm optymalizacji lokalnej szuka pierwszego lepszego rozwiązania w otoczeniu rozwiązania po mutacji i przechodzi do tego rozwiązania. Proces poszukiwania w otoczeniu otrzymanego rozwiązania jest powtarzany do momentu, gdy w otoczeniu nie ma już rozwiązania lepszego. Oznacza to ze zostało znalezione optimum lokalne.

Rozwiązania początkowe są generowane z wykorzystaniem takiej samej metody jak w algorytmie genetycznym albo jako rozwiązania początkowe wykorzystywane są rozwiązania uzyskane wcześniej innymi metodami np. przy pomocy algorytmu Bauma-Welcha.

Procedura mutacji dokonuje zmian wartości części składowych w rozwiązaniu $x_{najlepsze1}$, rozwiązania optymalnego lokalnie. W mechanizmie generującym zmiany został zastosowany taki sam algorytm mutacji jak w algorytmie genetycznym.

Algorytm iteracyjny dzięki zastosowaniu mutacji rozwiązania może wychodzić z optimów lokalnych. Do tworzenia rozwiązań początkowych został wykorzystany algorytm konstrukcyjny działający na takiej samej zasadzie jak algorytm konstrukcyjny w algorytmie genetycznym.

6.2.1. Otoczenie rozwiązania

Poprzez zmiany rozwiązania aktualnego w granicach rozwiązań dopuszczalnych uzyskuje się nowe rozwiązania z otoczenia. Zmiany rozwiązania dokonuje się według przyjętej reguły.

Podzbiór rozwiązań dopuszczalnych związanych ze zmianami rozwiązania aktualnego jest nazywany sąsiedztwem (*otoczeniem*) $N([x_{ijk}])$.

Jeśli rozwiązanie znajduje się w optimum lokalnym w otoczeniu nie występują rozwiązania lepsze ze względu na rozpatrywaną funkcje celu. Wyjście z optimów lokalnych następuje poprzez przejście przez gorsze rozwiązania niż rozwiązanie optimum lokalnego.

Zmiany, które są przeprowadzane w rozwiązaniu aktualnym, polegają na zwiększeniu lub zmniejszeniu jednej składowej prawdopodobieństwa modelu o wartość zależną od przyjętego w programie parametru określającego procent zmiany składowej. Dla modelu z trzema węzłami prawdopodobieństwo przejść miedzy węzłami opisane jest przez macierz o dziewięciu elementach. Przy przeszukiwaniu otoczenia każdy z elementów macierzy będzie zwiększany oraz zmniejszany o pewną wartość jak w zależności (6.12).

$$\begin{matrix} \widehat{\gamma}_{ij} = \gamma_{ij} + (1-\gamma_{ij})\Delta\% \\ \widehat{\gamma}_{ij} = \gamma_{ij} - (\gamma_{ij})\Delta\% \end{matrix} \quad , \qquad 1 \le i, j \le n \tag{6.12}$$

Parametr $\Delta\%$ określa procent zmiany wartości składowej.

Po dokonaniu zmiana cały wektor prawdopodobieństw w wierszu macierzy przejść, w którym została dokonana zmiana składowej jest normalizowany. W podobny sposób uzyskiwane są elementy otoczenia przez zmiany składowych wektora prawdopodobieństw początkowych, prawdopodobieństw pozostawania w danym stanie oraz prawdopodobieństw kolorów.

7. Zależności wykorzystywane do wyznaczenia algorytmu Bauma-Welcha

W rozdziale tym zostały umieszczone własności wykorzystywane podczas wyznaczaniu algorytmu Bauma-Welcha omówionego w rozdziale (3). Wyprowadzenia zależności można znaleźć w pracy [6].

Wszystkie cztery własności są prawdziwe dla łańcuch Markowa opisanego przez zbiór stanów $\{S_t: t\in\mathrm{N}\}$ oraz dla danego zbioru obserwacji $\{o_t: t\in\mathrm{N}\}$. Oba te zbiory są skończoną przestrzenią stanu procesu losowego. Model Markowa $M=\{S_t: t\in\mathrm{N}\}$ oraz obserwacje $\mathbf{O}=(o_1,o_2,...,o_T)$ mają tą własność, że przy określonym ciągu stanów $S^{(T)}=S_1,S_2, \ldots ,S_T$, losowe wartości $o_1,o_2,...,o_T$ są wzajemnie niezależne tzn. wartość prawdopodobieństwa $P(o_t|S_t)$ wystąpienia obserwacji o_t w stanie S_t zależy tylko od stanu S_t natomiast nie zależy od innych stanów S_k, $k\neq t$.

Własność (2.1) Dla t=1,2,...,T:

$$P(S_1,\ldots,S_T \mid S_t) = P(S_1,\ldots,S_t \mid S_t)P(S_{t+1},\ldots,S_T \mid S_t)$$

Własność (2.2) Dla t=1,2,...,T-1:

$$P(S_1,\ldots,S_T \mid S_t,S_{t+1}) = P(S_1,\ldots,S_t \mid S_t)P(S_{t+1},\ldots,S_T \mid S_{t+1})$$

Własność (2.4) Dla t = 1,2,...,T:

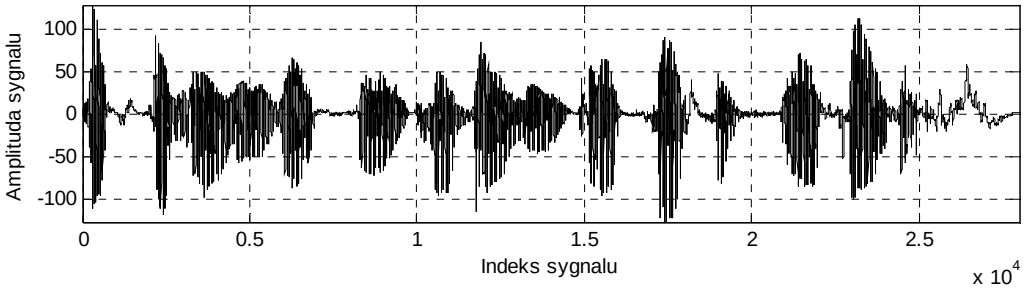
$$P(o_t,\ldots,o_T \mid S_t) = P(o_t \mid S_t)P(o_{t+1},\ldots,o_T \mid S_t)$$

Własność (2.3) Dla wszystkich l oraz t takich, że $1\leq t\leq l\leq T$:

$$P(o_l,\ldots,o_T \mid S_t,\ldots,S_l) = P(o_l,\ldots,o_T \mid S_l)$$

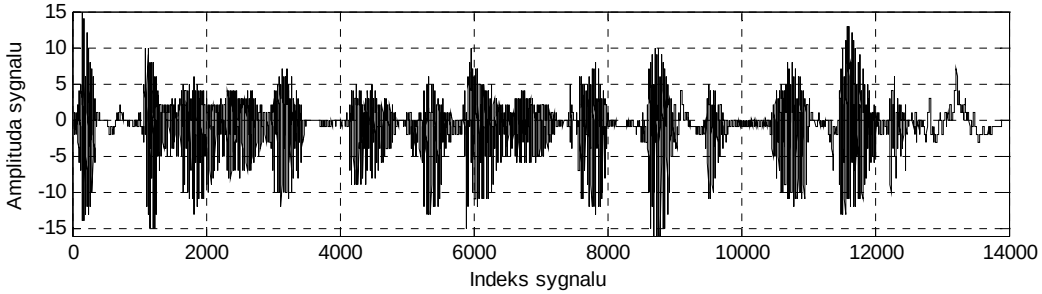
8. Tworzenie sekwencji uczących

Badania przedstawionych algorytmów zostały przeprowadzone na sekwencjach uczących utworzonych bezpośrednio z próbek dźwięku oraz z wartości uzyskanych z zastosowaniem metod LPC oraz MFCC oraz kwantyzacji wektorowej VQ. Badania zostały przeprowadzone na sylabach pochodzących z nagrań zdania „*Optymalizacja modeli procesów losowych*", oraz liczbach od 0 do 9. Na rysunku (8.1) przedstawione jest jedno z nagrań.



Rysunek 8.1. Zdanie „*Optymalizacja modeli procesów losowych*". Częstotliwość 8000 Htz Rozdzielczość 8 bitów na próbkę.

Dla sekwencji reprezentowanych bezpośrednio przez próbki dźwięk został nagrany w formacie (**.wav*) z częstotliwością 48000Htz i rozdzielczością 8 oraz bitów na próbkę. Z nagranych zdań wyodrębniono sylaby, w wyniku tego otrzymano 15 sekwencji uczących dla każdego nagrania. Częstotliwość próbkowania w sekwencjach uczących może przyjmować wartość 1000Htz, 2000Htz, 3000Htz 4000Htz, 6000Htz oraz 8000Htz natomiast rozdzielczość od 2 do 8 bitów na próbkę. Jeżeli przyjmie się rozdzielczość sekwencji uczącej 5 bitów na próbkę wówczas uzyskuje się sekwencję, w której liczba możliwych obserwacji wynosi n=32. Nowe wartości próbek przy zmianie częstotliwości zostały wyliczane poprzez linearyzację. Dodatkowo w sekwencjach uczących został podniesiony poziom dźwięku w taki sposób, żeby największe wartości próbek w nagranym dźwięku miały maksymalne wartości w sekwencji uczącej. Rysunek (8.2) przedstawia dźwięk po wprowadzonych modyfikacjach dla parametrów 4000Htz oraz 5 bitach na próbkę.



Rysunek 8.2. Sygnał dźwięku częstotliwość próbkowania 4000 Htz, rozdzielczość 5 bitów na próbkę.

W dźwięku zostały dokonane zmiany częstotliwości oraz rozdzielczości, ponieważ w przypadku częstotliwości większych niż 8000Htz oraz rozdzielczości 8 bitów prawdopodobieństwo wygenerowania sekwencji o długości 1000 próbek przez model HMM jest tak małe ze przekracza precyzje komputera. Przy formacie dźwięku np. 4000Htz oraz 5 bitów zachowana jest treść wypowiedzi oraz wartości uzyskiwanych prawdopodobieństw są zadawalające.

Sekwencje uczące dla rozdzielczości 5 bitów zawierają najwięcej wartości na poziomie środkowym o wartości 16 w wyniku tego po dobraniu modelu HMM do jakiejś sekwencji treningowej największe prawdopodobieństwo można uzyskać dla sekwencji mających wszystkie próbki o wartościach na poziomie 16.

Uzyskane sekwencje treningowe mają zróżnicowane długości. W celu porównania sekwencji o różnej długości nie można porównywać prawdopodobieństw wygenerowania tych sekwencji, dlatego zastosowana jest metoda opisana w rozdziale (9) pozwalające obliczyć prawdopodobieństwo wygenerowania danej sekwencji na *K* próbek niezależnie od długości danej sekwencji.

Poza bezpośrednią reprezentacją dźwięku w postaci próbek zostały przeprowadzane badania dla dźwięku po przetworzeniu metodą LPC oraz metodą MFCC. Metoda predykcji została opisana w rozdziale (8.2), natomiast metoda MFCC w rozdziale (8.3). W wyniku zastosowania powyższych metod uzyskuje się parametry dźwięku w postaci wektorów (*cepstrum*) charakteryzujących widmo sygnału w oknie. W metodach ekstrakcji parametrów wykorzystuje się transformację Fouriera opisaną w rozdziale (8.1).

W ramach pracy został zaimplementowany model sieciowy HMM w wersji dyskretnej, dlatego do parametrów dźwięku została zastosowana kwantyzacja wektorowa VQ. W wyniku kwantyzacji wektory reprezentujące widmo sygnału w oknie są indeksowane. W pracy do badania modeli HMM została zastosowana metoda kwantyzacji opisana w podrozdziale (8.4).

8.1. Krótko-czasowa transformacja Fouriera (STFT)

W tym podrozdziale zostanie opisana transformacji Fouriera. Po wyznaczaniu transformaty Fouriera dla sygnału dźwięku $x(n)$ w oknie o długości N próbek uzyskuje się zespolony wektor $X(k)$ o długości D nazywany widmem. Zależności (8.1) przedstawiają dyskretną transformatę Fouriera oraz transformatę odwrotną.

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-ik\frac{2\pi}{N}\frac{N}{D}n},\quad 0\le k\le D-1\qquad x(n)=\frac{1}{D}\sum_{k=0}^{D-1}X(k)e^{i\frac{2\pi}{N}nk\frac{N}{D}},\quad 0\le n\le N-1 \tag{8.1}$$

Współczynniki szeregu Fouriera a oraz b można wyznaczyć z zależności (8.2).

$$a_0=X(0),\quad a_k=re(X(k)),\quad b_k=-im(X(k)),\quad 1\le k\le D-1 \tag{8.2}$$

Zależności pomiędzy funkcją transformowaną $x(n)$ oraz współczynnikami szeregu Fouriera wyrażone są zależnością (8.3).

$$x(n)=\frac{a_0}{D}+\frac{1}{D}\sum_{k=1}^{D-1}a_k\cos(k\frac{2\pi}{D}n)+b_k\sin(k\frac{2\pi}{D}n),\quad 0\le n\le N-1,\quad 0\le n\le D-1 \tag{8.3}$$

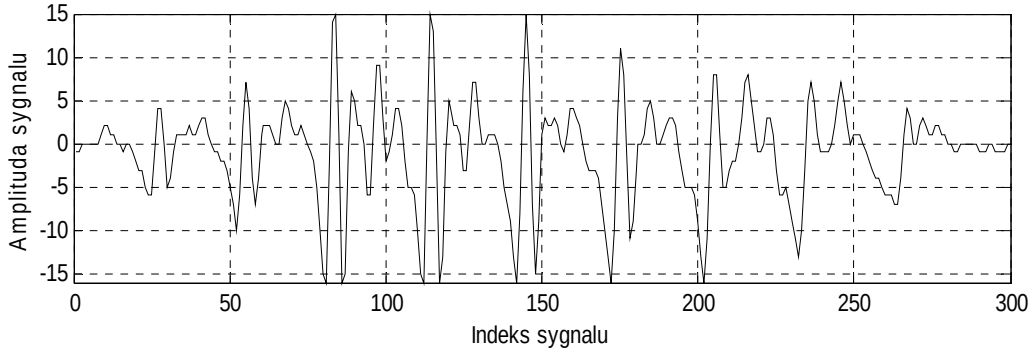
W wyniku zastosowania Transformaty uzyskuje się składową stałą a_0 oraz współczynniki poziomu harmonicznych. Moduł widma sygnału wyznacza się według zależności (8.4).

$$\left|X(k)\right|=\sqrt{(reX(k))^2+(imX(k))^2},\quad 0\le k\le D-1 \tag{8.4}$$

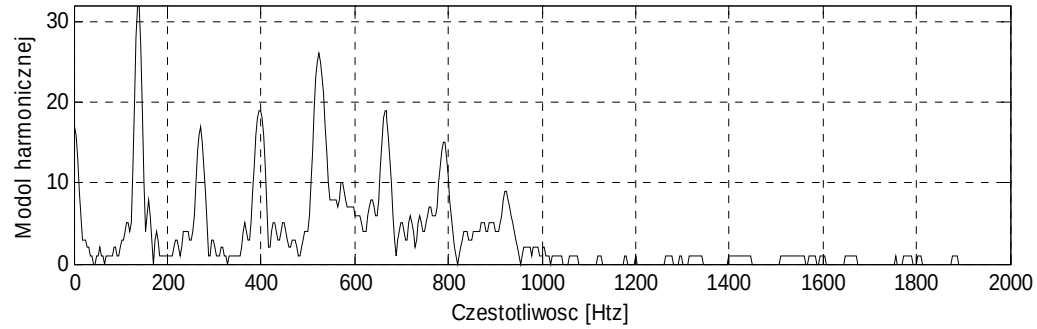
Widmo przedstawia udziały poszczególnych harmonicznych. Częstotliwości danej harmonicznej można wyznaczyć z zależności (8.5). W zależności (8.5) k oznacza wartość z dziedziny widma.

$$f_{Htz}=(f_p/(D-1))*k \tag{8.5}$$

Poniżej na rysunku (8.3) pokazana jest postać czasowa sygnału odpowiadającego sylabie „*op*”.



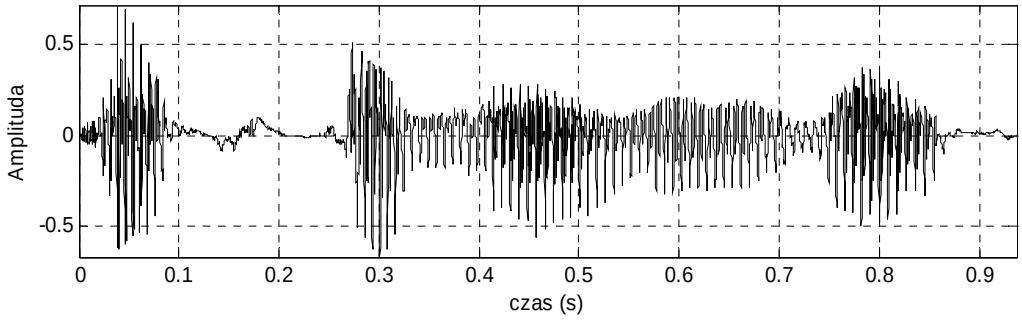
Rysunek 8.3. Sekwencja wartości reprezentujący sylabę „*op*”. Częstotliwość próbkowania 4000 Htz, rozdzielczość 5 bitów na próbkę czas 0.075s.



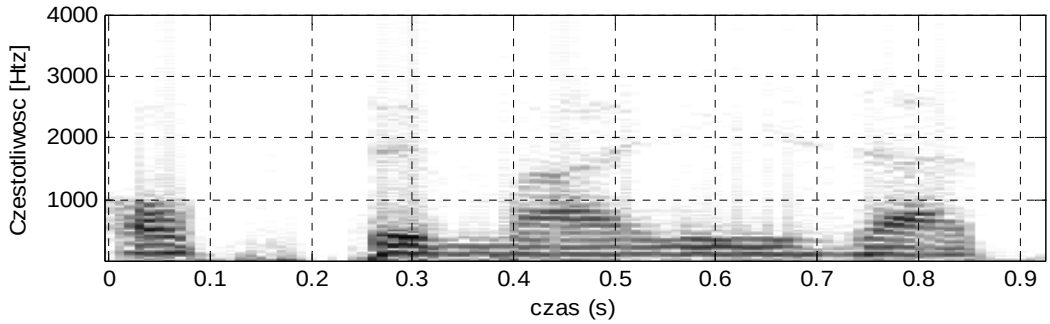
Rysunek 8.4. Moduł widma sylaby „*op*”, utworzona z sekwencji o częstotliwości próbkowania 4000 Htz, rozdzielczość 5 bitów na próbkę z rysunku 8.3.

Na rysunku (8.4) pokazany jest moduł widma odpowiadające sylabie „*op*” wyznaczone z sygnału z rysunku (8.3) dla parametru D=1200. Na rysunku przedstawione są składowe współczynniki harmonicznych od 0Htz do 2000Htz z krokiem 3.3(3)Htz.

Przedstawiając dziesiętne logarytmy modułu widma w oknach w skali szarości uzyskuje się spektrogram. Spektrogram sygnału liczby „*cztery*” pokazany jest na rysunku (8.6). Został on wyznaczony dla sygnału z rysunku (8.5). Na rysunku (8.7) pokazany jest moduł widma sygnału w skali logarytmicznej dla sygnału z rysunku (8.5).

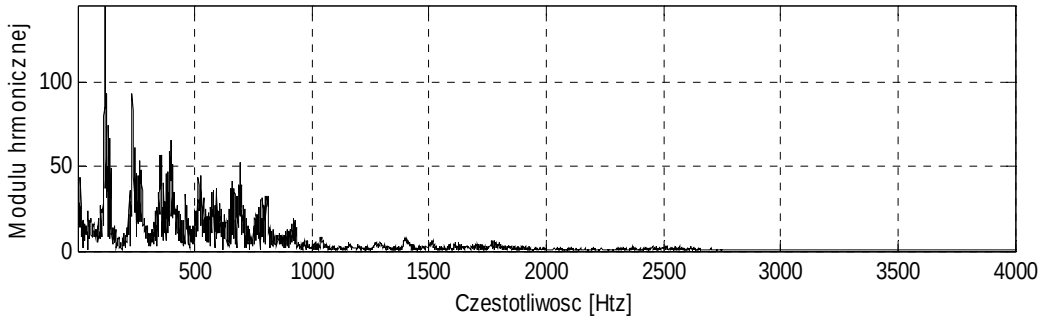


Rysunek 8.5. Sygnał dźwięku liczby „*optymalizacja*” 8000 Htz, rozdzielczość 16 bitów na próbkę.



Rysunek 8.6. Spektrogram sygnału z rysunku 8.5. Szerokość okna N=128, przesuniecie M=40, zakres transformaty *D*=512.

Spektrogram wyznaczony jest dla okien o szerokości *N*=128 przesuwanych o *N*=40 próbek, w każdym oknie została wyznaczona transformata Fouriera dla parametru *D*=512.



Rysunek 8.7. Moduł widma sygnału z rysunku 8.5. Szerokość okna transformaty z całego zakresu sygnału.

Z powyższych rysunków widać, że w celu dobrej reprezentacji dźwięku konieczna jest jego reprezentacja w oknach w celu uwzględnienia zmienności składowych harmonicznych w czasie.

8.2. Metoda predykcji liniowej (LPC)

W tym podrozdziale zostanie opisana metoda predykcji liniowej (*Linear Predictive Coding*) LPC stosowana w systemach rozpoznawania mowy.

Jest to jedna z najbardziej efektywnych metod reprezentacji dźwięku. Metoda ta jest również stosowana np. do kompresji w formacie mp3. W systemach rozpoznawania mowy sygnał reprezentowany jest przez wektory (*cepstrum*) wyznaczane z współczynników predykcji.

Podstawowa ideą w metodzie predykcji liniowej jest to, że sygnał $x(n)$ może być aproksymowany w oknie przez liniową kombinację p wcześniejszych próbek jak w równaniu (8.6).

$$x(n) \approx a_1 x(n-1) + a_2 x(n-2) + \cdots + a_p x(n-p) \tag{8.6}$$

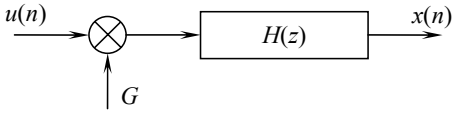
Przyjmuje się, że współczynniki predykcji $a_1,a_2,\ldots a_p$ są stałe dla analizowanego okna. Powyższe równanie można zapisać w postaci (8.7).

$$x(n) = \sum_{i=1}^{p} a_i x(n-i) + Gu(n) \tag{8.7}$$

Po zastosowaniu transformacji *z* do równania (8.7) uzyskuje się równania (8.8).

$$H(z)=\frac{1}{A(z)}=\frac{X(z)}{GU(z)}=\frac{1}{1-\sum_{i=1}^{p}a_i z^i},\quad GH(z)U(z)=X(z) \tag{8.8}$$

Zależność na H możne być traktowana jako transmitancja jak na schemacie blokowym rysunek (8.8), wartość $u(n)$ może być traktowana jako unormowane sterowanie natomiast czynnik G jako współczynnik poziomu sygnału.



Rysunek 8.8. Model blokowy predykcji liniowej.

Jeżeli przyjmie się, że liniowa kombinacja $\tilde{x}(n)$ ostatnich p próbek sygnału jest przybliżeniem sygnału $x(n)$ wówczas można obliczyć błąd predykcji $e(n)$ jak w zależności (8.9).

$$e(n)=Gu(n)=x(n)-\tilde{x}(n)=x(n)-\sum_{i=1}^{p}a_i x(n-i) \tag{8.9}$$

Współczynniki predykcji liniowej dobiera się w taki sposób, żeby zminimalizować sumaryczny błąd E_n opisany równaniem (8.10). Minimalizacja przeprowadzana jest dla sygnału $x(n)$ w oknie o szerokości N dla próbek od n do $n+N$, poza tym przedziałem przyjmuje się, że sygnał $x(n)$ jest równy zero. Błąd predykcji wyznacza się dla większego okna o szerokości $N-1+p$ dla próbek od n do $n+N-1+p$. Dodatkowo sygnał $x(n)$ w oknie wymnażany jest przez funkcje $w(m)$.

$$E_n=\sum_{m=0}^{N-1+p}e_n^2(m) \tag{8.10}$$

$$E_n=\sum_{m=0}^{N-1+p}\left[x_n(m)-\sum_{k=1}^{p}a_k x_n(m-k)\right]^2 \tag{8.11}$$

gdzie:

$$x_n(m)=\begin{cases}x(n+m)w(m), & dla \quad 0\le m<N\\ 0 & dla \quad m<0,\quad m\ge N\end{cases}$$

$$e_n(m)=e(n+m)$$

$$w(m)=0.54-0.46\cos\left(\frac{2\pi m}{N-1}\right)$$

W celu znalezienie optymalnych współczynników predykcji $\hat{a}_1,\hat{a}_2,...,\hat{a}_p$, dla których E_n przyjmuje wartość minimalną różniczkuje się błąd E_n po a_k i przyrównuje pochodne do zera, w wyniku uzyskuje się zależności (8.12).

$$\frac{\partial E_n}{\partial a_k}=0,\quad k=1,2,...,p$$

$$\sum_{m=0}^{N-1+p}x_n(m-i)x_n(m)=\sum_{k=1}^{p}\hat{a}_k\sum_{m=0}^{N-1+p}x_n(m-i)x_n(m-k),\qquad i=1,2,...,p \tag{8.12}$$

Wprowadzając funkcję $\phi_n(i,k)$ zdefiniowaną w równaniu (8.13) równania (8.12) można zapisać w postaci (8.14).

$$\phi_n(i,k) = \sum_{m=0}^{N-1+p} x_n(m-i)x_n(m-k), \qquad k = 1,2,...,p; \quad i = 1,2,...,p \tag{8.13}$$

$$\phi_n(i,0) = \sum_{k=1}^{p} \hat{a}_k \phi_n(i,k), \qquad i = 1,2,...,p \tag{8.14}$$

Wprowadzając zamianę zmiennych $l=m-i$ do (8.13) oraz funkcję autokorelacji $r(n)$ uzyskuje się zależność (8.15).

$$\phi_n(i,k) = r_n(i-k) = \sum_{l=-i}^{N-1-(i-p)} x_n(l)x_n(l+i-k) = \sum_{l=0}^{N-1-(i-k)} x_n(l)x_n(l+i-k) \tag{8.15}$$

Funkcja $\phi_n(i,k)$ zależne tylko od $(i-k)$ ponieważ wprowadzona funkcja autokorelacji r_n jest symetryczna $r(-k)=r(k)$ stąd równania (8.14) można zapisać jak w równaniu (8.16).

$$r_n(i) = \sum_{k=1}^{p} \hat{a}_k r_n(|i-k|), \qquad k = 1,2,...,p; \quad i = 1,2,...,p \tag{8.16}$$

Równania (8.16) w zapisie macierzowym przyjmują postać (8.17).

$$\begin{bmatrix} r_n(0) & r_n(1) & \cdots & r_n(p-1) \\ r_n(1) & r_n(1) & \cdots & r_n(p-2) \\ \vdots & \vdots & & \vdots \\ r_n(p-1) & r_n(p-2) & \cdots & r_n(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ \vdots \\ r_n(p) \end{bmatrix}, \quad \mathsf{R}_p \mathsf{a}_p = \mathsf{r}_p \tag{8.17}$$

Istnieją dwa sposoby wyznaczenia współczynników $\hat{a}_1, \hat{a}_2, ..., \hat{a}_p$, z równań (8.14) oraz (8.17). Podczas rozwiązywania układu równań korzysta się z iteracyjnej metody (*Durbina-Levisona*). Błąd predykcji E_n można zapisać jak w równaniu (8.18).

$$\begin{aligned} E_n &= \sum_{m=0}^{N-1+p} e_n^2(m) = \sum_{m=0}^{N-1+p} \left[s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right]^2 = \\ &\sum_{m=0}^{N-1+p} \left[s_n^2(m) - 2s_n(m) \sum_{k=1}^{p} a_k s_n(m-k) + \left(\sum_{k=1}^{p} a_k s_n(m-k) \right)^2 \right] \end{aligned} \tag{8.18}$$

Wprowadzając dodatkowe oznaczenia na wektor $\mathsf{s}_p^T(m)=[s_n(m\text{-}1)\ s_n(m\text{-}1)\ \ldots\ s_n(m\text{-}p)]$ poza oznaczeniami na: R_p, a_p, r_p błąd predykcji można zapisać jak w równaniu (8.19).

$$\begin{aligned} E_n &= \sum_{m=0}^{N-1+p} s_n^2(m) - 2 \sum_{m=0}^{N-1+p} s_n(m) \sum_{k=1}^{p} a_k s_n(m-k) + \sum_{m=0}^{N-1+p} \left(\sum_{k=1}^{p} a_k s_n(m-k) \right)^2 = \\ &= r(0) - 2\sum_{k=1}^{p} a_k \sum_{m=0}^{N-1+p} s_n(m)s_n(m-k) + \sum_{m=0}^{N-1+p} \mathsf{a}_p^{\mathsf{T}} \mathsf{s}(m) \mathsf{s}_p^{\mathsf{T}}(m) \mathsf{a}_p = \\ &= r(0) - 2\sum_{k=1}^{p} a_k r(k) + \sum_{m=0}^{N-1+p} \mathsf{a}_p^{\mathsf{T}} \mathsf{s}_p(m) \mathsf{s}_p^{\mathsf{T}}(m) \mathsf{a}_p = \\ &= r(0) - 2\sum_{k=1}^{p} a_k r(k) + \mathsf{a}_p^{\mathsf{T}} \left[\sum_{m=0}^{N-1+p} \mathsf{s}_p(m) \mathsf{s}_p^{\mathsf{T}}(m) \right] \mathsf{a}_p = \end{aligned} \tag{8.19}$$

$$= r(0) - 2\mathbf{a}_p^{\mathsf{T}}\mathbf{r}_p + \mathbf{a}_p^{\mathsf{T}}\mathbf{R}_p\mathbf{a}_p = [1 \quad \mathbf{a}_p^{\mathsf{T}}]\mathbf{R}_{p+1}[1 \quad \mathbf{a}_p]$$

Dla optymalnych wartości parametrów predykcji można wyznaczyć minimalny błąd $E_{n\,min}$, korzystając z równania (8.20), jest on często oznaczany w literaturze przez σ^2 oraz wykorzystywany w modelach auto-regresywnych HMM.

$$E_{n\min} = \sigma^2 = r(0) - 2\mathbf{a}_p^{\mathsf{T}}\mathbf{r}_p + \mathbf{a}_p^{\mathsf{T}}\mathbf{r}_p = r(0) - \mathbf{a}_p^{\mathsf{T}}\mathbf{r}_p = |\mathbf{R}_p| / |\mathbf{R}_{p\text{-}1}| = r(0) - \sum_{k=1}^{p} \hat{a}_k r(k) \qquad (8.20)$$

W metodzie predykcji zakłada się, że sygnał sterowania $u(n)$ jest unormowany, co można zapisać jak w równaniu (8.21).

$$\sum_{n=-\infty}^{\infty} u^2(n) = 1$$
$$e(n) = Gu(n) \qquad (8.21)$$

stąd:

$$\sigma^2 = G^2$$

Szczegóły dotyczące iteracyjnej metody (*Durbina-Levisona*) rozwiązywania układu równań liniowych podczas wyznaczania współczynników predykcji można znaleźć w pracy [12]. W zależnościach (8.22) przedstawiony jest iteracyjny algorytm podany w pracy [3].

$$E(0) = r(0)$$
$$k_i = \{r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(|i-j|)\} / E^{(i-1)}, \qquad 1 \le i \le p$$
$$\alpha_i^{(i)} = k_i \qquad (8.22)$$
$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}, \qquad 1 \le j < i$$
$$E^{(i)} = (1 - k_i^2) E^{(i-1)},$$

Równania (8.22) rozpoczynające się od akapitu oblicza się iteracyjnie dla i=1,2,..,p po wyznaczaniu powyższych zależności współczynniki predykcji wyznaczane są według zależności (8.23).

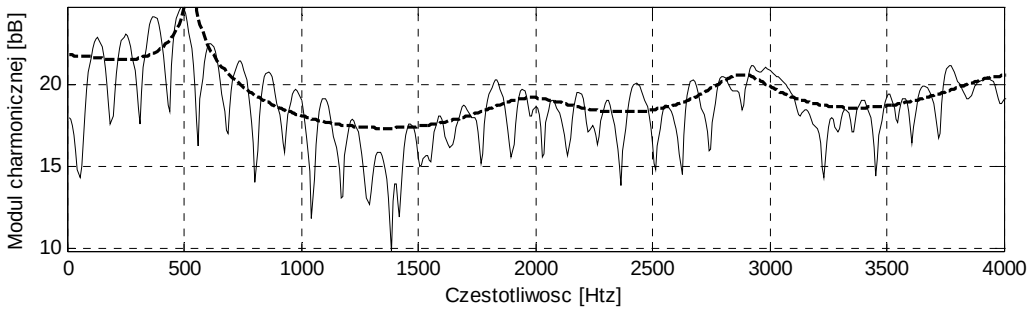
$$\hat{a}_i = \alpha_i^{(p)}, \qquad 1 \le i \le p \qquad (8.23)$$

Dzięki zastosowaniu metody LPC uzyskuje się współczynniki predykcji $\hat{a}_1, \hat{a}_2, ..., \hat{a}_p$. Modułu widma transmitancji $GH(z)$ jest podobna do modułu widma sygnału $x(n)$, co można zapisać jak w zależności (8.24).

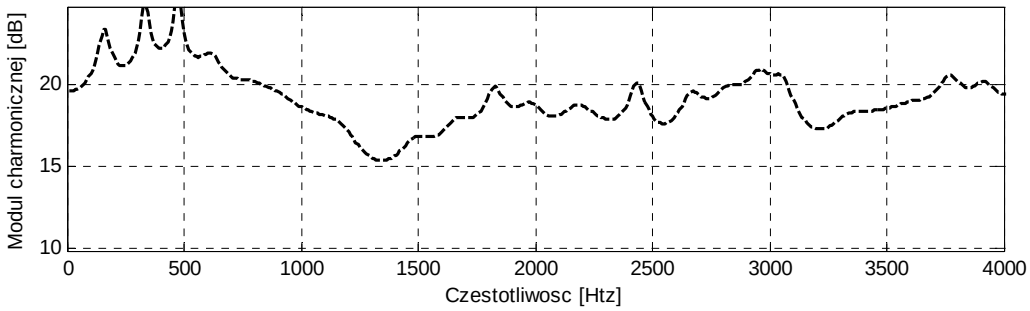
Podobieństwo modułów zależy od wartości parametru p dla większych wartości parametru uzyskuje się większe podobieństwo z powodu mniejszego błędu predykcji. Opuszczając czynnik G otrzymuje się charakterystyki częstotliwościowe transmitancji modelu predykcyjnego niezależne od poziomu sygnału.

$$G|H(e^{j\omega})| = \frac{G}{|A(e^{j\omega})|} \qquad \sim |X(e^{j\omega})|$$
$$G|H(e^{j\omega})| = \frac{|X(e^{j\omega})|}{|U(e^{j\omega})|} = G\frac{|X(e^{j\omega})|}{|E(e^{j\omega})|} \qquad \sim |X(e^{j\omega})| \qquad (8.24)$$

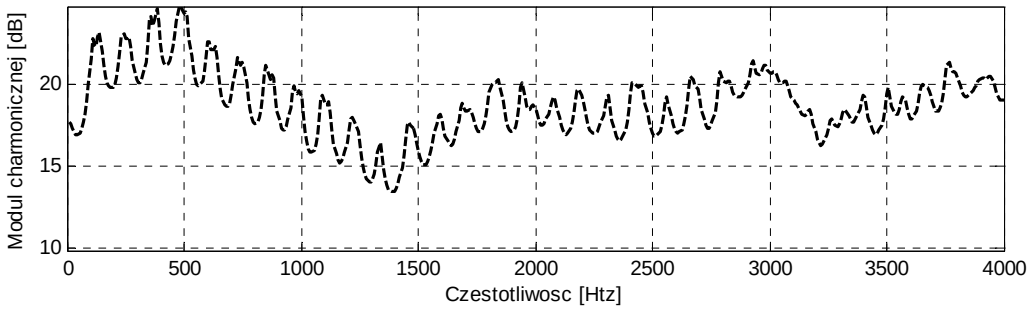
Na rysunku (8.9) pokazany jest kwadrat moduł widma sylaby '*op*' linia ciągła oraz kwadrat moduł widma transmitancji w skali logarytmicznej dla parametru p=12. Na rysunkach (8.10)-(8.11) pokazane są kwadraty moduły widma w skali logarytmicznej transmitancji dla parametrów p=50 oraz p=150.



Rysunek 8.9. Kwadrat modułu widma sygnału w skali logarytmicznej $\log(|(X(k)|^2)$ oraz kwadrat moduł widma transmitancji w skali logarytmicznej $\log_{10}(\delta^2|H(k)|^2)$ modelu predykcyjnego dla p=12.



Rysunek 8.10. Kwadrat moduł widma transmitancji w skali logarytmicznej $\log(\delta^2|H(k)|^2)$ modelu predykcyjnego dla p=50.



Rysunek 8.11. Kwadrat moduł widma transmitancji w skali logarytmicznej $\log(\delta^2|H(k)|^2)$ modelu predykcyjnego dla p=150.

Przed wyznaczaniem parametrów predykcji LPC sygnał poddawany jest premfrazie jak w zależność (8.25).

$$\hat{x}(n) = x(n) - 15/16x(n-1) \tag{8.25}$$

8.3. Metoda (MFCC)

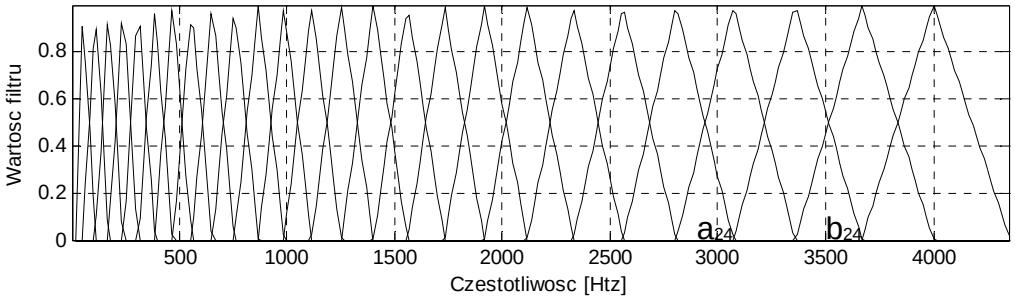
W tym podrozdziale została opisana metoda przetwarzania dźwięku (MFCC) (*Mel-frequency cepstral coefficients*). Pierwsze publikacje na temat tej metody zostały opracowane w roku 1980 (praca [124]). W metodzie MFCC stosuję się skalowania charakterystyki częstotliwościowej, ponieważ dźwięk o niskiej częstotliwości jest słyszany z większą

rozdzielczością niż o wysokiej. Podczas skalowania stosuje się wyznaczane eksperymentalnie funkcje skalujące, najczęściej funkcje (8.26).

$$Mel(f) = 1127.01048 \log(1 + f/700) \tag{8.26}$$

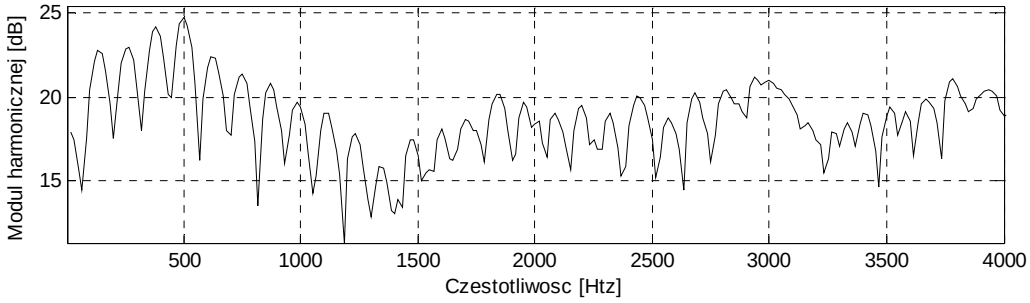
Najpierw z sygnału w oknie wyznacza się kwadrat widma $S(k)$, następnie z $S(k)$ wyznacza się współczynnik Y_i i=1,2,...,M według równania (8.27). Współczynniki Y_i reprezentują widmo w skali „*Mel*". Podczas wyznaczania Y_i korzysta się z zespołu filtrów $H_i(k)$ $a_i \le k \le b_i$ i=1,2,…P o trójkątnej charakterystyce. Na rysunku (8.12) pokazany jest zespół 26 filtrów dla sygnału próbkowanego z częstotliwością *fp* =8000KHtz.

$$Y_i = \log\left[\sum_{k=a_i}^{b_i} |S(k)|\, H_i(k)\right] \tag{8.27}$$

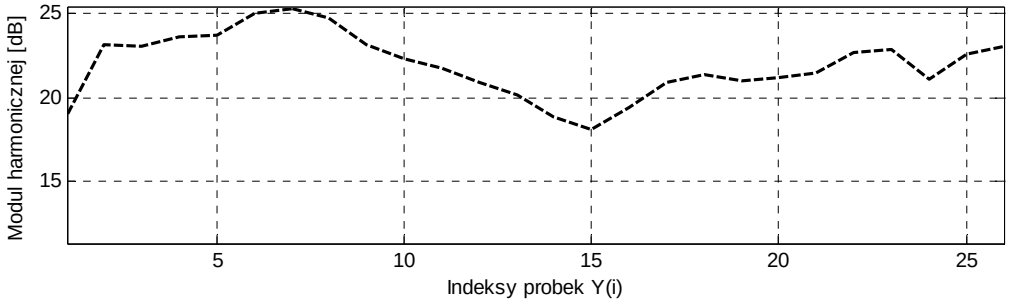


Rysunek 8.12. Zespół 26 filtrów 240 elementowych dla częstotliwości próbkowania sygnału f_p=8000 Htz.

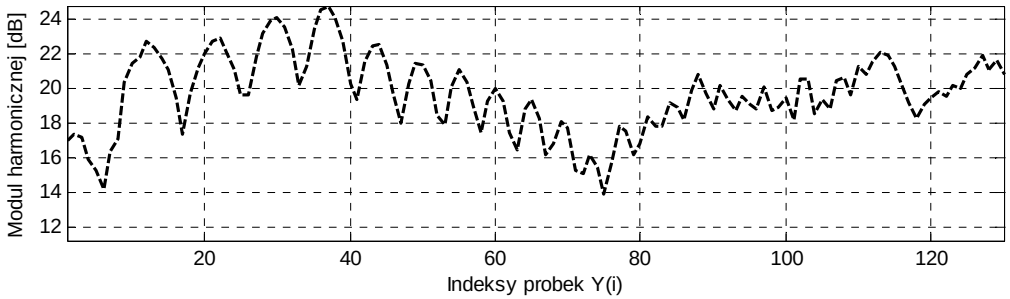
W wyniku zastosowania zależności (8.27) uzyskuje się przybliżone przeskalowane widmo. Na rysunku (8.13) pokazany jest kwadrat modułu widma sygnału w skali logarytmicznej, natomiast na rysunkach (8.14)-(8.15) wykresy współczynników Y_i dla P=26 oraz 130 filtrów.



Rysunek 8.13. Kwadrat modułu widma sygnału w skali logarytmicznej $\log(|(X(k)|^2)=\log(S(k))$.



Rysunek 8.14. Wykres współczynników Y_i dla 26 filtrów.



Rysunek 8.15. Wykres współczynników Y_i dla 130 filtrów.

W celu wyznaczenia współczynników (*cepstrum*) c_n stosuje się transformację kosinusową z równania (8.28). Uzyskane w ten sposób współczynniki są niezależne od poziomu sygnału, ponieważ zmiana poziomu sygnału powoduje przesuwanie się charakterystyki $\log(Y_i)$ o stałą wartość a, która podlega redukcji w transformacji kosinusowej, co zostało pokazane w zależności (8.28).

$$\begin{aligned} c_n &= \sum_{j=1}^{M}(Y_j + a)\cos(\frac{n\pi}{M}(j-0.5)) = \qquad , \quad n = 1,2,...,P \\ &= \sum_{j=1}^{M} Y_j \cos(\frac{n\pi}{M}(j-0.5)) + a\sum_{j=1}^{M}\cos(\frac{n\pi}{M}(j-0.5)) = \sum_{j=1}^{M} Y_j \cos(\frac{n\pi}{M}(j-0.5)) \end{aligned} \tag{8.28}$$

Można znaleźć pracę wprowadzające modyfikacje opisanej metody polegające na kompensacji przybliżonego widma Y_j np.: w pracy [125] została wprowadzona metoda kompensacji podana równaniem (8.29). Kompensacja została wprowadzona w celu poprawy wyników rozpoznawania w przypadku, gdy dźwięk jest nagrany z szumem w tle.

$$m_j = \alpha_j \log\{1 + \beta \max[(Y_j - \hat{N}_j), \gamma Y_j]\} \tag{8.29}$$

Współczynniki $\beta, \gamma \in (0,1)$ są niezależne od indeksu filtru j, natomiast parametr $\alpha_j \in (0,1)$ jest wyznaczany dla każdego z filtrów według zależności (8.30). Współczynniki $\hat{N}_j$ reprezentują przybliżone widmo sygnału szumu wyznaczane w taki sam sposób jak Y_j z równania (8.27). Po kompensacji współczynniki *Yj* zastępowane są przez m_j w transformacji kosinusowej.

$$\alpha_j = \frac{\log(1 + Y_j / \hat{N}_j)}{\sum_{k=1}^{M}\log(1 + Y_j / \hat{N}_j)}, \qquad \sum_{k=1}^{M}\alpha_j = 1 \tag{8.30}$$

8.4. Pomiar odległości pomiędzy widmami sygnału

Istnieje wiele metod pomiaru odległości pomiędzy widmami sygnału. W tym rozdziale została opisana zastosowana w programie metoda (*cepstral distance*), oraz metoda uzyskaną po skalowaniu widma do skali „*Mel*".

Podczas porównywania sygnału w oknach porównuje się wygładzone widma sygnału. Wygładzanie stosowane jest w celu uwydatnienia częstotliwości o najwyższym poziomie (*formantów*).

Odległość dwóch sygnałów w oknach o szerokości N, dla których kwadraty modułów widma oznaczone są przez $S(k)$, $S'(k)$ wyznacza się z zależności (8.31).

$$d_2^2 = \frac{1}{2N-1}\sum_{k=-N+1}^{N-1} |\log S(k) - \log S'(k)|^2 \tag{8.31}$$

gdzie:

$$\begin{aligned} S(k) &= \sum_{n=-N+1}^{N+1} r(n) e^{-jnk\frac{2\pi}{2N-1}}, \quad -N+1 \le k \le N-1, \\ r(n) &= 1/(2N-1)\sum_{k=-N+1}^{N-1} S(k) e^{jnk\frac{2\pi}{2N-1}}, \quad -N+1 \le n \le N-1 \end{aligned} \tag{8.32}$$

$r(n)$ jest to korelacja sygnału $x(n)$ opisana zależnościami (8.33).

$$\begin{aligned} r(n) &= \sum_{k=0}^{N-1} x(k)x(k-n), \quad -N+1 \le n \le N-1 \\ r(n) &= 0, \qquad\qquad -N+1 > n > N-1 \end{aligned} \tag{8.33}$$

Szereg Fouriera wyrażenia $\log(S(k))$ można zapisać jak w zależności (8.34), wówczas współczynniki szeregu c_n nazywa się współczynnikami (*cepstrum*) są odpowiednikami korelacji $r(n)$ sygnału. Widmo współczynników (*cepstrum*) jest widmem korelacji sygnału w skali logarytmicznej.

$$\log S(k) = \sum_{n=-N+1}^{N-1} c_n e^{-jnk\frac{2\pi}{2N-1}}, \quad -N+1 \le k \le N-1 \tag{8.34}$$

Korzystając z (8.34) zależność na odległość dwóch sygnałów (8.31) można zapisać jak w równaniu (8.35).

$$\begin{aligned} d_2^2(S,S') &= \frac{1}{2N-1}\sum_{k=-N+1}^{N-1} |\log S(k) - \log S'(k)|^2 = \frac{1}{2N-1}\sum_{k=-N+1}^{N-1} \Big| \sum_{n=-N+1}^{N-1} c_n e^{-jnk\frac{2\pi}{2N-1}} - \sum_{n=-N+1}^{N-1} c_n' e^{-jnk\frac{2\pi}{2N-1}} \Big|^2 = \\ &= \frac{1}{2N-1}\sum_{k=-N+1}^{N-1} \Big| \sum_{n=-N+1}^{N-1} (c_n - c_n') e^{-jnk\frac{2\pi}{2N-1}} \Big|^2 = \frac{1}{2N-1}\sum_{k=-N+1}^{N-1} \Big| \sum_{n=-N+1}^{N-1} (c_n - c_n') e^{-jnk\frac{2\pi}{2N-1}} \Big|^2 = \\ &= \frac{1}{2N-1}\sum_{k=-N+1}^{N-1} \Big(\sum_{n=-N+1}^{N-1} (c_n - c_n') e^{-jnk\frac{2\pi}{2N-1}} \sum_{m=-N+1}^{N-1} (\bar{c}_m - \bar{c}_m') e^{jmk\frac{2\pi}{2N-1}} \Big) = \\ &= \frac{1}{2N-1}\sum_{k=-N+1}^{N-1} \Big(\sum_{n=-N+1}^{N-1} \sum_{m=-N+1}^{N-1} (c_n - c_n')(\bar{c}_m - \bar{c}_m') e^{-j(n-m)k\frac{2\pi}{2N-1}} \Big) = \\ &= \sum_{n=-N+1}^{N-1} \sum_{m=-N+1}^{N-1} (c_n - c_n')(\bar{c}_m - \bar{c}_m') \frac{1}{2N-1} \sum_{k=-N+1}^{N-1} e^{-j(n-m)k\frac{2\pi}{2N-1}} = \sum_{n=-N+1}^{N-1} \sum_{m=-N+1}^{N-1} (c_n - c_n')(\bar{c}_m - \bar{c}_m') \delta(n-m) = \\ &= \sum_{n=-N+1}^{N-1} |c_n - c_n'|^2 = \sum_{n=-N+1}^{N-1} (c_n - c_n')^2 \end{aligned} \tag{8.35}$$

Jeżeli sygnał jest reprezentowany przez model predykcyjny wówczas $S(k)$ odpowiada w modelu predykcyjnym $\sigma^2/|A(e^{j\omega})|^2$. Po rozwinięciu wyrażenia $1/|A(e^{j\omega})|$ w szereg (*Laurenta*) (8.36) a następnie zróżniczkowaniu tego wyrażenie ze względu na z^{-1} i przyrównaniu współczynników o takim samym wykładniku z^{-1} otrzymuje się rekurencyjną zależność na współczynniki szeregu jak w zależności (8.37).

$$\log\left[\sigma / A(z^{-1})\right] = \log[\sigma] + \sum_{n=1}^{\infty} c_n z^{-n} \tag{8.36}$$

$$\begin{aligned} c_n &= -a_n - \frac{1}{n}\sum_{k=1}^{n-1} k c_k a_{n-k}, \qquad n > 0 \\ c_0 &= \log\sigma^2 \end{aligned} \tag{8.37}$$

Po wyznaczeniu współczynników (*cepstrum*) stosuje się dodatkową procedurę (8.38).

$$C_n = c_n\left(1 + \frac{Q}{2}\sin(\pi n / Q)\right), \quad 0 \le n \le Q \tag{8.38}$$

W praktyce podczas porównywania widma sygnału porównuje się L tj. 10-30 współczynników (*cepstrum*) c_n bez współczynnika c_0 odpowiadającego poziomowi sygnału, wówczas zależność (8.31) przyjmuje postać (8.39).

$$d_2^2 = \sum_{n=1}^{L}(c_n - c_n^{'})^2 \tag{8.39}$$

Istnieje możliwość przeskalowania współczynników c_n w taki sposób, żeby reprezentowały dźwięk w skali (*Mel*).

$$Mel(f) = 1127.01048\log(1 + f/700) \tag{8.40}$$

Po wprowadzeniu funkcji skalującej (8.40) oraz uwzględnieniu symetrii widma oraz jego dyskretnego charakteru można wypisać zależności (8.41). Zależności te opisują przyporządkowanie indeksów dla widma przed i po skalowaniu, w związku z tym z ich pomocą można skalować dyskretne widmo. W wyniku skalowania zostają opuszczone niektóre próbki dla wysokich częstotliwości, natomiast dla niskich powielone.

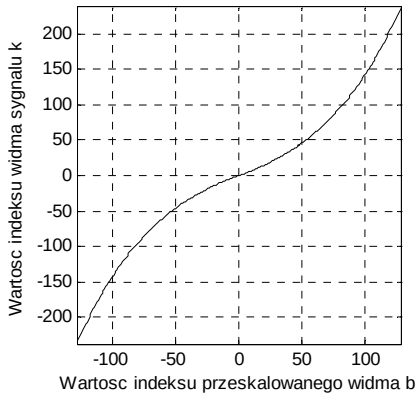
$$\begin{aligned} Mel(k) &= -((2N\text{-}1)/f_p)1127.01048\log(1 + (-kf_p/(2N\text{-}1))/700)), \quad -N+1 \le k < 0 \\ Mel(k) &= ((2N\text{-}1)/f_p)1127.01048\log(1 + (kf_p/(2N\text{-}1))/700)), \qquad 0 \le k \le N-1 \end{aligned}$$

$$\begin{aligned} Mel^{-1}(b) &= -(e^{-bf_b/(1127.01048(2N\text{-}1))} - 1)(700/(f_p/(2N\text{-}1))), \quad -B \le b < 0 \\ Mel^{-1}(b) &= (e^{bf_b/(1127.01048(2N\text{-}1))} - 1)(700/(f_p/(2N\text{-}1))), \qquad 0 \le b \le B \end{aligned} \tag{8.41}$$

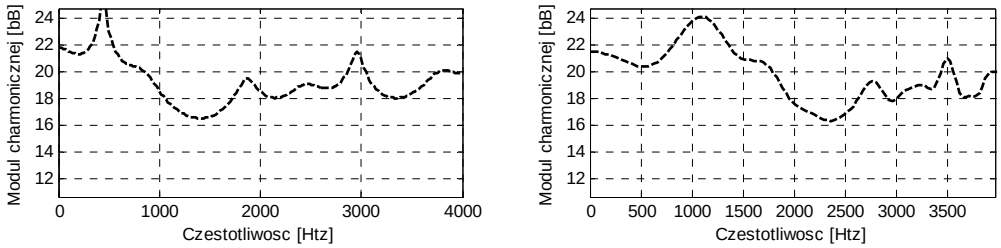
$$\begin{aligned} \theta(b) &= Mel^{-1}(b), \quad -B \le b < 0 \\ \theta(b) &= Mel^{-1}(b), \quad 0 \le b \le B \end{aligned}$$

$$B = \left\lfloor ((2N\text{-}1)/f_p)1127.01048\log(1 + ((N\text{-}1)f_p/(2N\text{-}1)/700)) \right\rfloor$$

Wykres $\theta(b)$ dla częstotliwości próbkowania 8000*KHtz* i N=240 pokazany jest na rysunku (8.16). Widmo modelu predykcyjnego $\log\left[\delta^2 \Big/ \left|A(e^{j\omega})\right|^2\right]$ $0 \le k \le N$ przed skalowaniem oraz po skalowaniu dla $0 \le b \le B$ przedstawione są na rysunkach (8.17).



Rysunek 8.16. Wykres funkcji skalującej $\theta(b)$ dla fp=8000, N=256.



Rysunek 8.17. Kwadrat moduł widma transmitancji w skali logarytmicznej $\log(\delta^2|H(k)|^2)$ modelu predykcyjnego dla p=26 przed i po skalowaniu.

Wprowadzając przeskalowane widmo do zależności (8.31) uzyskuje się zależności (8.42).

$$\begin{aligned}\tilde{d}_2^2(S,S') &= \frac{1}{2B-1}\sum_{b=-B}^{B}|\log S(\theta(b))-\log S'(\theta(b))|^2 = \\ &= \sum_{n=-N+1}^{N-1}\sum_{m=-N+1}^{N-1}\left(c_n-c_n'\right)\left(c_m-c_m'\right)\frac{1}{2B-1}\sum_{b=-B}^{B}e^{-j(n-m)\theta(b)\frac{2\pi}{2N-1}} = \\ &= \sum_{n=-N+1}^{N-1}\sum_{m=-N+1}^{N-1}\left(c_n-c_n'\right)\left(c_m-c_m'\right)w_{nm}\end{aligned} \tag{8.42}$$

Funkcja $\theta(b)$ jest to dyskretna funkcja o wartościach ciągłych. Po wprowadzeniu limitu L na ilość porównywanych współczynników (*cepstrum*) oraz oznaczenia $\boldsymbol{c}=[c_1\ c_2\ \ldots\ c_L]$, $\boldsymbol{W}=[w_{mn}]$ można uzyskać równanie (8.43).

$$\begin{aligned}\tilde{d}_2^2 &= \sum_{n=-L}^{L}\sum_{m=L}^{L}\left(c_n-c_n'\right)\left(\bar{c}_m-\bar{c}_m'\right)w_{nm} = (\boldsymbol{c}-\boldsymbol{c}')\boldsymbol{W}(\boldsymbol{c}-\boldsymbol{c}')^T = \\ &= (\boldsymbol{c}-\boldsymbol{c}')\boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}(\boldsymbol{c}-\boldsymbol{c}')^T = (\boldsymbol{c}\boldsymbol{W}^{1/2}-\boldsymbol{c}'\boldsymbol{W}^{1/2})(\boldsymbol{c}\boldsymbol{W}^{1/2}-\boldsymbol{c}'\boldsymbol{W}^{1/2})^T\end{aligned} \tag{8.43}$$

Symetryczna macierz $\boldsymbol{W}^{1/2}$ może być wyznaczona jeden raz i wykorzystana do skalowania współczynników c_n. Wartości współczynników macierzy $\boldsymbol{W}$ można wyznaczyć z zależności (8.44).

$$w_{mn} = \frac{1}{2B-1}\sum_{b=-B}^{B}e^{-j(n-m)\theta(b)\frac{2\pi}{2N-1}} \tag{8.44}$$

Jeżeli dźwięk jest reprezentowany przez wektory uzyskane metodą (MFCC) wówczas odległość (8.31) można zapisać w postaci (8.45).

$$d_c^2 = \sum_{i=1}^{P} (c_i - c_i^{'})^2 \tag{8.45}$$

W metodzie (*cepstral distance*) mirą odległości pomiędzy widmami sygnału reprezentowanymi przez wektory cech jest suma kwadratów różnicy składowych wektorów (*cepstrum*).

Inna stosowaną miarą odległości jest np. iloczynu skalarny wektorów (*cepstrum*).

8.5. Pochodne cepstrum

W celu poprawy efektywności reprezentacji dźwięku przez wektory (*cepstrum*) uwzględnia się dynamikę tj.: pierwsze i drugie pochodne wektorów w czasie. Pierwsza pochodna zapisana jest równaniem (8.46).

$$\frac{\partial \log S(\omega,t)}{\partial t} = \sum_{i=-\infty}^{\infty} \frac{\partial c_i(t)}{\partial t} e^{-j\omega} \tag{8.46}$$

W celu wyznaczenia pochodnej wykorzystuje się aproksymację współczynników (*cepstrum*) $c(t)$ wielomianem drugiego rzędu $h_1+h_2t+h_3t^3$. Współczynniki wielomianu dobiera się przez minimalizację błędu E opisanego równaniem (8.47).

$$E = \sum_{t=-M}^{M} [c(t) - (h_1 + h_2 t + h_3 t^2)]^2 \tag{8.47}$$

Z równań (8.48) uzyskanych dzięki zróżniczkowaniu błędu E po parametrach h_1, h_2, h_3 i przyrównaniu pochodnych do zera można wyznaczyć współczynniki wielomianu (8.49).

$$\begin{aligned}
&\sum_{t=-M}^{M} [c(t) - (h_1 + h_2 t + h_3 t^2)] = 0 \\
&\sum_{t=-M}^{M} [c(t)t - h_1 t - h_2 t^2 - h_3 t^3] = 0 \\
&\sum_{t=-M}^{M} [c(t)t - h_1 t^2 - h_2 t^3 - h_3 t^4] = 0
\end{aligned} \tag{8.48}$$

$$\begin{aligned}
h_1 &= \frac{1}{2M}\left[\sum_{t=-M}^{M} c(t) - h_3 T_M\right] \\
h_2 &= \frac{\sum_{t=-M}^{M} tc(t)}{T_M} \\
h_3 &= \frac{T_M \sum_{t=-M}^{M} c(t) - (2M+1)\sum_{t=-M}^{M} t^2 c(t)}{T_M^2 - (2M+1)\sum_{t=-M}^{M} t^2}
\end{aligned} \tag{8.49}$$

gdzie:

$$T_M = \sum_{t=-M}^{M} t^2$$

Po wyznaczeniu współczynników pochodne (*cepstrum*) wyznacza się z zależności (8.50).

$$\frac{\partial c_i(\tau+t)}{\partial t} \approx h_2 = \Delta c_i, \quad \frac{\partial^2 c_i(\tau+t)}{\partial t^2} \approx 2h_3 = \Delta^2 c_i, \tag{8.50}$$

W systemach rozpoznawania mowy stosuje się obserwacje $\mathbf{o}=(c_1,\ldots c_L,\ \Delta c_1,\ldots,\Delta c_L,\ \Delta^2 c_1,\ldots,\Delta^2 c_L)$ złożone z współczynników (*cepstrum*) oraz ich pierwszych i drugich pochodnych.

8.6. Metoda modyfikacji algorytmu przetwarzania sygnału

Porównując charakterystyki częstotliwościowe sygnału mowy można zauważyć, że dla tych samych dźwięków wypowiadanych przez różnych mówców występują przesunięcia maksimów modułu widma transmitancji tj. formantów. Przesunięcia formantów są spowodowane różnymi charakterystykami torów głosowych mówców.

W pracy [174] zostały omówione metody poprawy efektywności porównywania charakterystyk częstotliwościowych. Poprawę efektywności uzyskano dzięki zastosowaniu skalowania widma w dziedzinie częstotliwości, oraz modyfikacji biegunów transmitancji modelu predykcyjnego.

W tym podrozdziale zostanie omówiona metoda, która powinna dać podobne wyniki do metod opisanych w pracy [174]. Metoda ta polega na wprowadzeniu dodatkowej funkcji do skali (*Mel*). Celem wprowadzenia funkcji jest wyeliminowanie przesunięć charakterystyk częstotliwościowych mówcy rozpoznawanego w stosunku do mówcy bazowego. Za mówcę bazowego rozumie się tutaj mówcę, dla którego wzorców dźwięku trenowane były modele HMM systemu rozpoznawania mowy.

Jeżeli wartości średnie dwóch pierwszych formanty dla ustalonego dźwięku mówcy bazowego wynoszą odpowiednio f_1, f_2, natomiast mówcy rozpoznawanego wynoszą f'_1, f'_2, wówczas można wprowadzić dodatkową funkcję skalującą $K(f)$ do skali (*Mel*). Zmodyfikowana skala opisana jest zależnością (8.51).

$$Mel(f) = 1127.01048\log(1 + K(f)/700) \tag{8.51}$$

W podrozdziale została zaproponowana funkcja skalująca $K(f)$ w postaci funkcji sklejanej opisana zależnością (8.52).

$$\begin{aligned} K(f) &= a_1 f^3 + a_2 f^2 + a_3 f, && 0 \le f \le f_1' \\ K(f) &= b_1 f + b_2, && f_1' \le f \le f_2' \\ K(f) &= c_1 f^3 + c_2 f^2 + c_3 f + c_4, && f_2' \le f \le f_p/2 \end{aligned} \tag{8.52}$$

Współczynniki a, b oraz c można wyznaczyć korzystając z zależności (8.53).

$$p=\left(f_2'-f_1'\right)/\left(f_2-f_1\right),\quad p_1=f_1'/f_1,\quad p_2=\left(f_p/2-f_2'\right)/\left(f_p/2-f_2\right)$$

$$\mathbf{A}_1=\begin{bmatrix} f_1^3 & 3f_1^2 & 0 \\ f_1^2 & 2f_1 & 0 \\ f_1 & 1 & 1 \end{bmatrix},\quad \mathbf{b}_1=\begin{bmatrix} f_1' & p & p_1 \end{bmatrix};$$

$$\mathbf{A}_2=\begin{bmatrix} f_1 & f_2 \\ 1 & 1 \end{bmatrix},\quad \mathbf{b}_2=\begin{bmatrix} f_1' & f_2' \end{bmatrix}$$

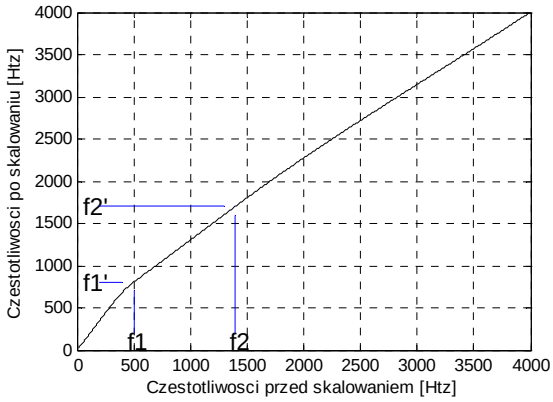
$$\mathbf{A}_3=\begin{bmatrix} f_1^2 & \left(f_p/2\right)^3 & 3f_2^2 & 3\left(f_p/2\right)^2 \\ f_2^2 & \left(f_p/2\right)^2 & 2f_2^2 & 2\left(f_p/2\right) \\ f_2 & \left(f_p/2\right) & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix},\quad \mathbf{b}_3=\begin{bmatrix} f_2' & \left(f_p/2\right) & p & p_2 \end{bmatrix} \qquad (8.53)$$

$$\begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}=\mathbf{b}_1\mathbf{A}_1^{-1}$$

$$\begin{bmatrix} b_1 & b_2 \end{bmatrix}=\mathbf{b}_2\mathbf{A}_2^{-1}$$

$$\begin{bmatrix} c_1 & c_2 & c_3 & c_4 \end{bmatrix}=\mathbf{b}_3\mathbf{A}_3^{-1}$$

Przykładowa funkcja skalująca pokazana jest na rysunku (8.18).



Rysunek 8.18. Wykres funkcji $K(f)$ dla f_1=500Htz, f_2=1400Htz, f_1'=800Htz, f_2'=1700.

Opisaną modyfikację można wprowadzić do metody LPC podczas wyznaczania macierzy ***W*** oraz do metody MFCC. W metodzie MFCC spowoduje ona specyficzne dla mówcy rozmieszczenie zespołu filtrów trójkątnych. Podobne funkcje skalujące prawdopodobnie są stosowane, ale nie są one wyznaczane na podstawie formantów.

8.7. Kwantyzacja wektorowa (VQ)

Modele HMM z dyskretnymi rozkładami prawdopodobieństw w stanach mogą być stosowane jedynie do modelowania sygnału mowy w postaci próbek albo symboli uzyskanych metoda kwantyzacji wektorowej wektorów cech dźwięku. W wyniku kwantyzacji sygnał dźwięku w oknie reprezentowany jest przez pojedynczą wartość.

Ponieważ sieciowe modele HMM opisane w rozdziale (4) zostały zaimplementowane dla dyskretnych rozkładów prawdopodobieństwa w stanach, dlatego sekwencje treningowe musiały zastać poddane kwantyzacji.

W metodzie kwantyzacji odległość widm sygnału była wyznaczana według zależności (8.39).

Do reprezentacji dźwięku została zastosowana jedna książka kodowa o maksymalnej wielkości równaj 1024 wektory.

Zastosowany algorytm kwantyzacji polegał na początkowym losowym wyborze wektorów a następnie iteracyjnej ich modyfikacji mającej na celu zmniejszenie błędu kwantyzacji.

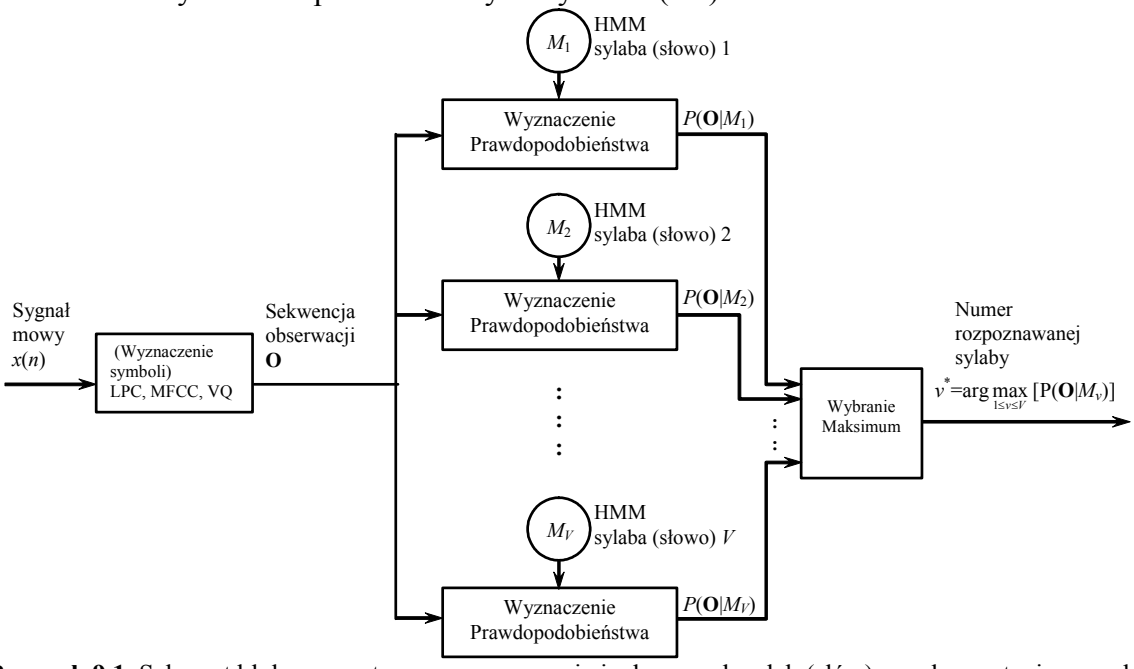
W stosowanym algorytmie zostały zastosowane dwa kryteria stopu algorytmu tj. liczba kroków oraz przekroczenie dopuszczalnej wartości relatywnego błędu kwantyzacji.

9. Opis systemu rozpoznawania

W ramach pracy został zaimplementowany system umożliwiający badanie modeli HMM. Ponieważ zaproponowany algorytm optymalizacji modelu sieciowego nie jest zastosowany w systemach takich jak np. HTK, który został umieszczony na stronie [182], dlatego konieczne było napisanie własnego oprogramowania. Oprogramowanie dołączone jest na załączonej płycie CD-ROM. Na dołączonej płycie umieszczone są następujące katalogi i pliki:

Literatura – katalog zawiera literaturę dotycząca tematyki związanej z pracą,
Nagrania – katalog zawiera nagrania izolowanych liczb i sylab,
Program – katalog zawiera skompilowany program optymalizacyjny *Program.exe*, biblioteki Borlanda Buildera 6.0, pliki tekstowe parametrów oraz dodatkowe katalogi:
Kod Programu – katalog zawiera kod programu napisanego w C++,
Pliki – katalog zawiera pliki z danymi potrzebne do poprawnego uruchomienia programu,
Pomoc - katalog zawiera plik pomocy programu,
Wyniki zamieszczone w pracy – katalog ten zawiera dane, oraz wyniki zamieszczone w rozprawie tj: pliki z danymi, wykresy oraz tabele,
Opis Programu.pdf – plik zawiera opis programu,
Rozprawa Doktorska Karol Szostek.pdf – tekst rozprawy.

System umożliwia rozpoznawanie przyjętej liczby V słów. W działaniu systemu można wyróżnić dwa etapy:
1) Etap optymalizacji. Dla każdej sylaby v ze słownika, generowany jest model HMM. Wykorzystywane są algorytmy estymacji parametrów modeli HMM.
2) Etap rozpoznawania. W etapie rozpoznawania dla każdej rozpoznawanej sylaby został zastosowany schemat przedstawiony na rysunku (9.1).



Rysunek 9.1. Schemat blokowy systemu rozpoznawania izolowanych sylab (słów) z wykorzystaniem modeli HMM.

Do trenowania modeli HMM zostały zaimplementowane algorytmy Bauma-Welcha oparte na kryterium (EM). Dla modelu podstawowego został zastosowany algorytm opisany w podrozdziale (3.3.4.1). Dla modelu sieciowego został zaimplementowany algorytm opisany w podrozdziale (4.2). Algorytm dla modelu sieciowego jest to uniwersalny algorytm, który jest również wykorzystywany do optymalizacji modeli z jawną gęstością rozkładu prawdopodobieństwa pozostawania w tym samym stanie opisanego w podrozdziale (3.4).

Poza algorytmami Bauma-Welcha zostały zastosowane dwa algorytmy sztucznej inteligencji opisane w rozdziale (6).

Napisany program pozwala na optymalizowanie modeli HMM do wielokrotnych sekwencji obserwacji. Ponieważ sekwencje obserwacji mogą mieć różne długości, dlatego została wprowadzona zależność wyznaczania prawdopodobieństwa sekwencji na K próbek opisana równaniem (9.1) oraz (9.2).

$$P_K(\mathbf{O}/M) = \left(\sqrt[T]{P(\mathbf{O}/M)}\right)^K \tag{9.1}$$

W przypadku logarytmu prawdopodobieństwa wykorzystana jest zależności (9.2).

$$\log P_K(\mathbf{O}/M) = \frac{K}{T}\log P(\mathbf{O}/M) \tag{9.2}$$

gdzie: T - długość sekwencji $\mathbf{O}$, K – parametr normalizujący.

W programie została przyjęta wartość K=600, ponieważ sekwencje odpowiadające sylabom w postaci próbek przy częstotliwości próbkowania 8000 Htz najczęściej posiadały taką długość.

Wprowadzenie zależności (9.1) oraz (9.2) było niezbędne w celu wyznaczenia błędów modelowana. Program umożliwia wyznaczanie zarówno błędów modelowania oraz błędów rozpoznawania dla zastosowanych modeli HMM. Błędy modelowania występują wówczas, gdy dla ustalonego modelu M uzyskuje się większe prawdopodobieństwo $P(\mathbf{O}|M)$ dla sekwencji niezgodnej z modelem niż dla sekwencji zgodnej z modelem. Błędy rozpoznawania występują wówczas, gdy dla ustalonej sekwencji $\mathbf{O}$ uzyskuje się większe prawdopodobieństwo dla modelu nie zgodnego z sekwencją niż dla modelu zgodnego z sekwencją.

W programie istnieje możliwość badania zjawiska przeuczenia się modeli HMM w przypadku wykonywania zbyt dużej ilości iteracji do niewystarczającej ilości danych. Wykresy efektów przeuczenia modeli HMM przy niewystarczających danych umieszczone są w rozdziale (10.4). Na wykresach pokazane są przebiegi kryterium oceny modeli podczas optymalizacji. Na jednym wykresie narysowany jest przebieg kryterium dla zbioru uczącego oraz zbioru oceny (*podobnie jak dla sieci neuronowych stosuje się zbiór walidacyjny*).

Ponieważ algorytmy wykorzystują wielokrotne sekwencje obserwacji oraz zbiory uczący oraz oceny mogą być różnej wielkości, dlatego w celu porównania kryterium dla zbioru uczącego oraz zbioru oceny zostało wprowadzone kryterium na sekwencje opisane zależnością (9.3)

$$\log P(\mathbf{O}^{(k)} \mid \mathrm{M})_{Na\ Sekwencje} = \log \prod_{k=1}^{L} P_k \Big/ L \tag{9.3}$$

Zależność ta została zastosowana z założeniem, że prawdopodobieństwo $P(\mathbf{O}^{(1)},...,\mathbf{O}^{(L)}|M)$ jest to iloczyn L wartości.

Wyniki przeprowadzonych badań zastosowania modeli HMM do rozpoznawania sekwencji dźwięku zostały przedstawione w rozdziale (10).

10.Wyniki obliczeń komputerowych

W rozdziale tym zostały umieszczone wyniki przeprowadzonych badań. Badania zostały przeprowadzonych przy pomocy napisanego przez autora systemu rozpoznawania opisanego w rozdziale (9). W systemie zostały zaimplementowane algorytmy optymalizacji oparte na kryterium (EM) opisane w rozdziałach (3), (4), (6) dla modeli z dyskretnymi rozkładami prawdopodobieństw obserwacji w stanach. Badania zostały przeprowadzone dla 15 sylab oraz cyfr od 0 do 9. Sylaby zostały wyodrębnione z nagranych zdań przy pomocy programu Cool Edit 2000. Wykorzystane zostały sekwencje w postaci próbek dźwięku oraz symboli uzyskanych metodą kwantyzacji wektorowej wektorów cech wyznaczonych metodami LPC oraz MFCC.

W pierwszej części zostały przedstawione wyniki optymalizacji modeli HMM do sekwencji, dla których istnieje optymalny model HMM dający prawdopodobieństwo wygenerowania sekwencji równe jeden.

W drugiej części przedstawione są efekty przeuczenia się modeli HMM dla małej ilości danych treningowych.

W trzeciej części zostały przedstawione wyniki optymalizacji modeli HMM do sekwencji uczących w postaci sylab.

W czwartej części zostały umieszczone wyniki porównania modelu podstawowego HMM, modelu z czasem przebywania oraz modelu sieciowego.

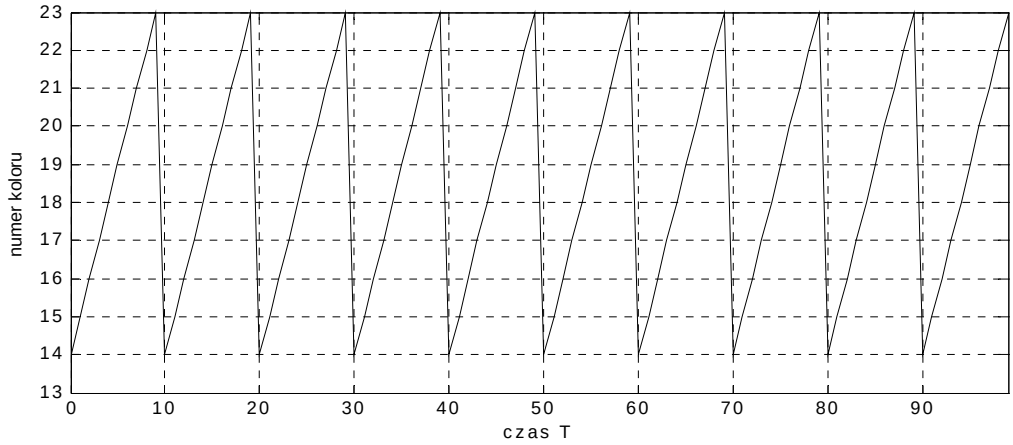
W ostatniej części rozdziału zostały omówione wyniki zastosowania algorytmów sztucznej inteligencji do optymalizacji modeli HMM.

10.1. Optymalizacja modeli do sekwencji, dla których istnieje optymalny model HMM

W przeprowadzonych badaniach zostały zastosowane modele HMM z dyskretnymi rozkładami w stanach o 32 wartościach.

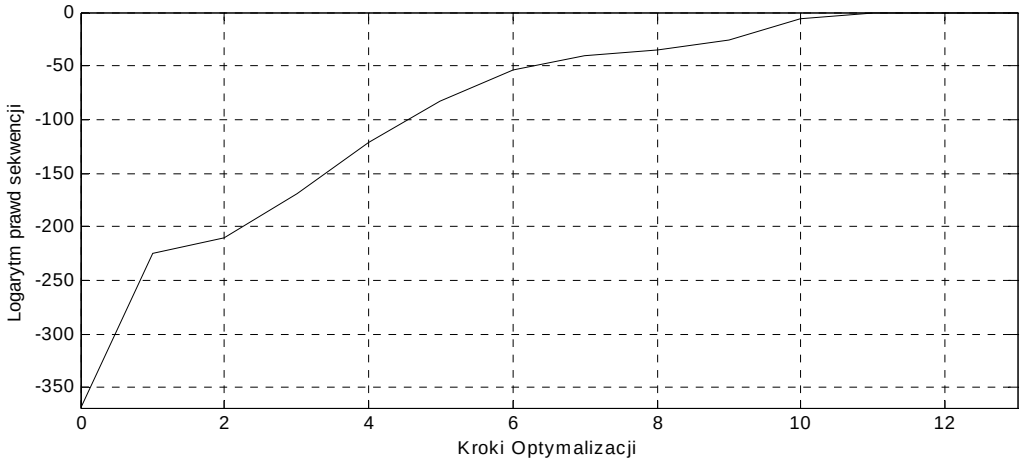
10.1.1. Optymalizacja modelu podstawowego HMM

Dla modelu podstawowego przyjęto liczbę stanów 10 oraz minimalną wartość dla prawdopodobieństw modelu równą 1e-2000. Optymalizacja została przeprowadzona dla sekwencji z rysunku (10.1).



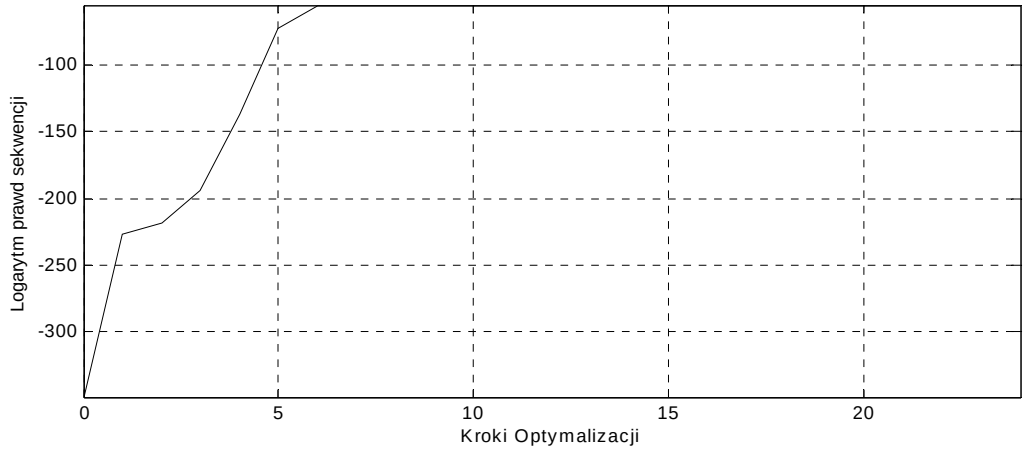
Rysunek 10.1. Sekwencja ucząca modelu podstawowego, powtarzający się dziesięć razy sygnał wzrastający od 14 do 23. Długość sekwencji wynosi 100 próbek.

W zależności od początkowego modelu HMM w wyniku optymalizacji znajdowane jest optimum globalne albo algorytm zatrzymuje się w pewnym krytycznym punkcie i nie następuje dalsza poprawa wartości funkcji oceny modelu opisana równaniem (9.2). Czym większe rozmiary modelu HMM tym trudniejsze jest znalezienie rozwiązania optymalnego. W przypadku wykorzystanego modelu rozwiązanie to było znajdywane 30 razy na 100 rozwiązań początkowych.



Rysunek 10.2. Przebieg procesu optymalizacji modelu podstawowego HMM. Wykres przedstawia logarytm prawdopodobieństwa wygenerowania sekwencji obserwacji w kolejnych krokach optymalizacji (*W procesie optymalizacji zostało znalezione rozwiązanie optymalne*).

Na rysunku (10.2) przedstawiony jest przebieg optymalizacji, w którym zostało znalezione rozwiązanie optymalne. Rozwiązanie optymalne zostało znalezione w 13 krokach algorytmu. Rozwiązanie początkowe miało wartość: (P)=1.80802e-160, log(P)=-367821e+02.

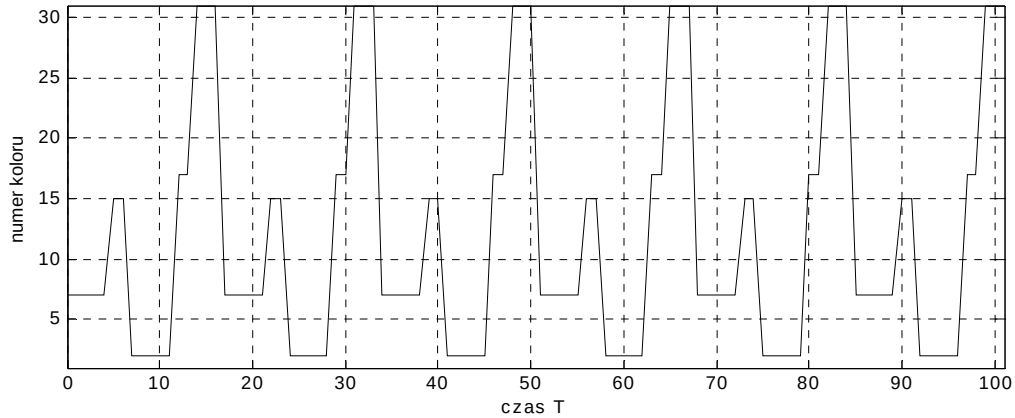


Rysunek 10.3. Przebieg procesu optymalizacji modelu podstawowego HMM. Wykres przedstawia logarytm prawdopodobieństwa wygenerowania sekwencji obserwacji w kolejnych krokach optymalizacji (*W procesie optymalizacji zostało znalezione rozwiązanie optymalne*).

Na rysunku (10.3) pokazany jest przebieg optymalizacji w czasie, której nie zostało znalezione rozwiązanie optymalne. Rozwiązanie początkowe miało wartość: (P)= 1.02023e-151, log(P)=-3.47670e+02, natomiast prawdopodobieństwo dla najlepszego znalezionego rozwiązania wynosi (P)=8.27180e-25, log(P)=-5.54517e+01.

10.1.2. Optymalizacja modelu z jawnym rozkładem prawdopodobieństwa pozostawania w tym samym stanie

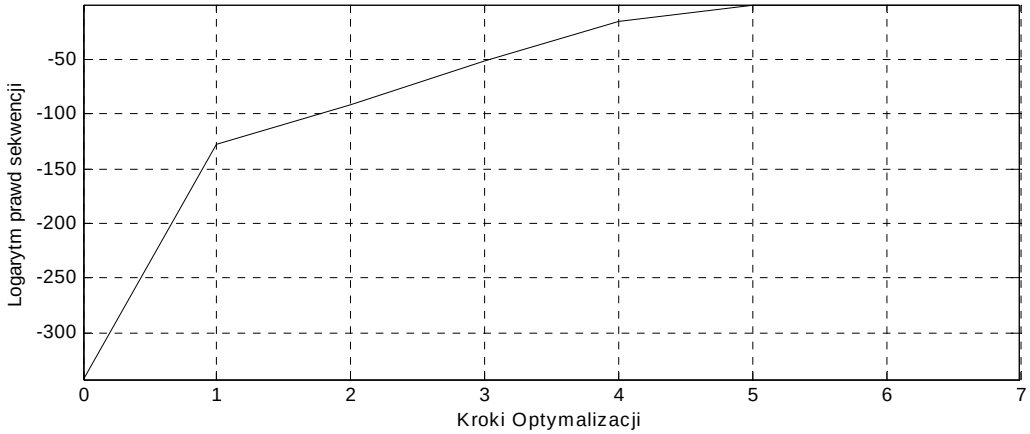
Dla modelu z jawną gęstością prawdopodobieństwa pozostawania w tym samym stanie przyjęto liczbę stanów równą 5, maksymalny czas pozostawania w danym stanie 5 oraz minimalną wartość dla prawdopodobieństw modelu równą 1e-500. Optymalizacja została przeprowadzona dla sekwencji z rysunku (10.4).



Rysunek 10.4. Sekwencja do uczenia modelu z jawną gęstością prawdopodobieństwa pozostawania w tym samym stanie. Wykres przedstawia sześciokrotnie powtarzający się sygnał składający się z pięciu siódemek, dwóch piętnastek, pięciu dwójek, dwóch siedemnastek, oraz trzech trzydziestek jedynek. Długość sekwencji wynosi 102 próbki.

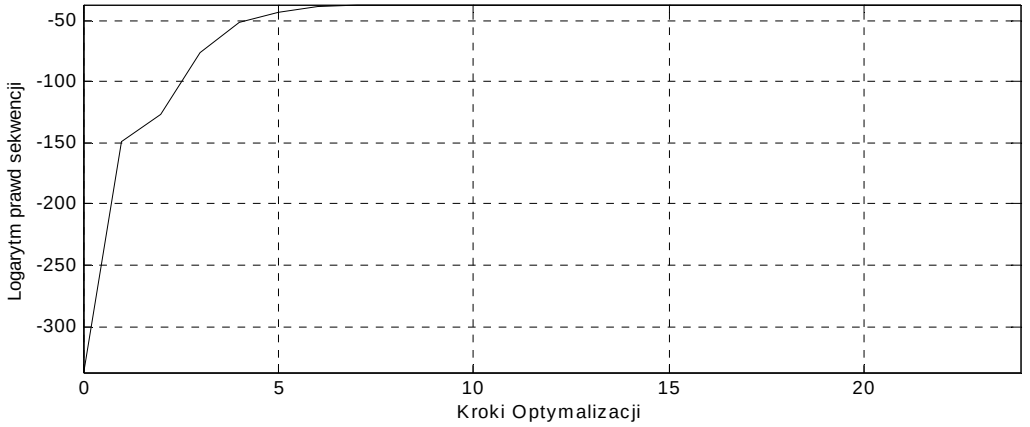
Podobnie jak w modelu podstawowym w zależności od wartości początkowych modelu HMM w wyniku optymalizacji znajdowane jest optimum globalne albo algorytm Bauma-Welcha zatrzymuje się przy w krytycznym punkcie. Wyniki pokazane są dla modelu z pięcioma stanami, ponieważ dla modeli z większą liczbą stanów znalezienie rozwiązania optymalnego jest

kłopotliwe. W tym przypadku rozwiązanie optymalne było znajdywane 20 razy na 100 rozwiązań początkowych.



Rysunek 10.5. Przebieg procesu optymalizacji modelu HMM z jawną gęstością prawdopodobieństwa pozostawania w tym samym stanie. Wykres przedstawia logarytm prawdopodobieństwa wygenerowania sekwencji obserwacji w kolejnych krokach optymalizacji (*W procesie optymalizacji zostało znalezione rozwiązanie optymaln*e).

Na rysunku (10.5) przedstawiony jest przebieg optymalizacji, w którym zostało znalezione rozwiązanie optymalne w 7 krokach algorytmu, rozwiązanie początkowe miało wartość: *P*= 5.47158e-149, log(*P*)= -3.41385e+02.

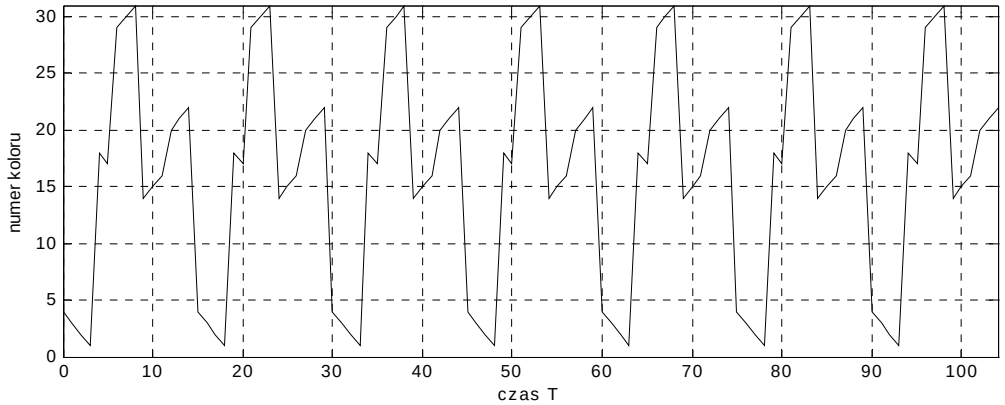


Rysunek 10.6. Przebieg procesu optymalizacji modelu HMM z jawną gęstością prawdopodobieństwa pozostawania w tym samym stanie. Wykres przedstawia logarytm prawdopodobieństwa wygenerowania sekwencji obserwacji w kolejnych krokach optymalizacji (W procesie optymalizacji nie zostało znalezione rozwiązanie optymalne).

Na rysunku (10.6) pokazany jest przebieg optymalizacji dla modelu HMM, dla którego nie zostało znalezione rozwiązanie optymalne. Rozwiązanie początkowe miało wartość *P*= 7.79229e-147, log(*P*)=-3.36426e+02, natomiast prawdopodobieństwo najlepszego znalezionego rozwiązania wynosi: *P*=1.0155995666841607092e-16, log(*P*)=-3.68258e+01.

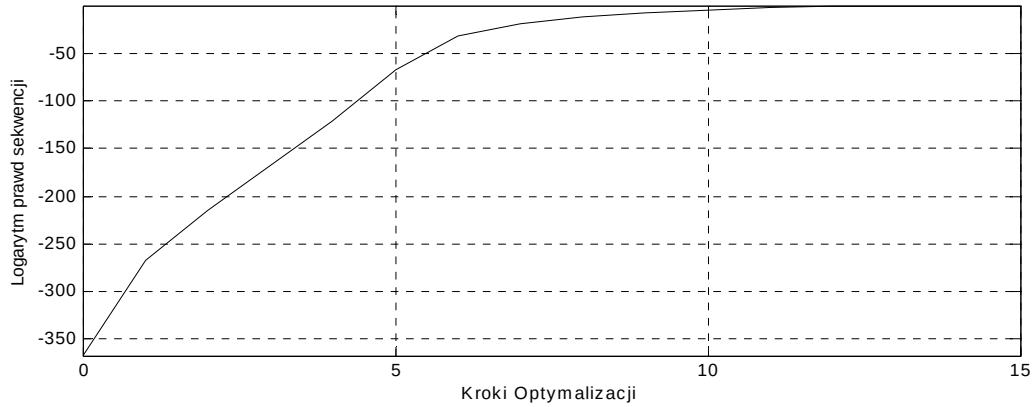
10.1.3. Optymalizacja modelu sieciowego

Dla modelu z sieciowego przyjęto liczbę węzłów 3, maksymalny czas pozostawania w danym węźle 5, liczbę stanów pod modeli równą 3, minimalną wartość prawdopodobieństw modelu równą 1e-500. Optymalizacja została przeprowadzona dla sekwencji z rysunku (10.7).



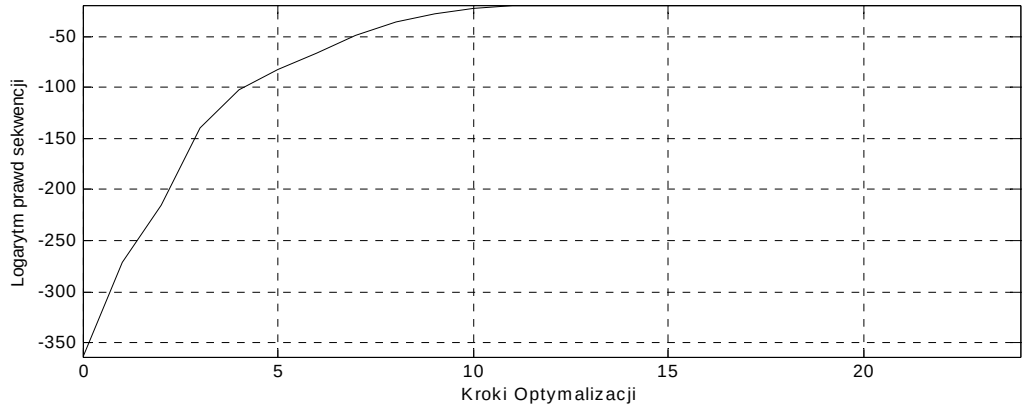
Rysunek 10.7. Sekwencja do uczenia modelu sieciowego. Wykres przedstawia siedmiokrotnie powtarzający się sygnał składający się z liczb od 4 do 1, od 18 do 17, od 29 do 31, od 14 do 16, oraz od 20 do 22. Długość sekwencji wynosi 102 próbek.

Podobnie jak w modelu podstawowym znalezienie rozwiązanie zależy od doboru początkowego modelu HMM.



Rysunek 10.8. Przebieg procesu optymalizacji modelu sieciowego HMM. Wykres przedstawia logarytm prawdopodobieństwa wygenerowania sekwencji obserwacji w kolejnych krokach optymalizacji (W procesie optymalizacji zostało znalezione rozwiązanie optymalne).

Na rysunku (10.8) przedstawiony jest przebieg optymalizacji, w którym zostało znalezione rozwiązanie optymalne w 15 krokach algorytmu, wartość funkcji oceny dla rozwiązania początkowego wynosiła P=9.92553e-160, log(P)=-3.66118e+02.



Rysunek 10.9. Przebieg procesu optymalizacji modelu sieciowego HMM. Wykres przedstawia logarytm prawdopodobieństwa wygenerowania sekwencji obserwacji w kolejnych krokach optymalizacji (*W procesie optymalizacji nie zostało znalezione rozwiązanie optymalne*).

Na rysunku (10.9) pokazany jest przebieg optymalizacji dla modelu HMM, dla którego nie zostało znalezione rozwiązanie optymalne. Wartość funkcji oceny dla rozwiązania początkowego wynosiła P= 3.80024e-158, logP=-3.62473e+02, natomiast wartość funkcji oceny dla najlepszego znalezionego rozwiązania wynosiła P= 3.72529e-09, log(P)=-1.94081e+01.

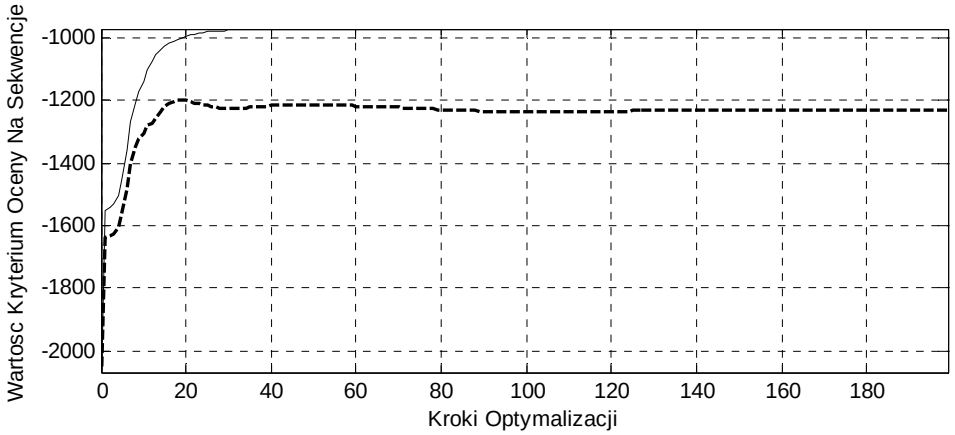
Z przeprowadzonych badań widać, że algorytmy poprawy nie zawsze znajdują zadawalające rozwiązania nawet dla prostych modeli oraz sekwencji, dla których wiemy, jakie jest rozwiązanie optymalne. Wartość prawdopodobieństwa dla najlepszego znalezionego modelu nie zależy od wartości prawdopodobieństwa modelu początkowego. W procesie optymalizacji dla modelu podstawowego prawdopodobieństwo modelu początkowego, dla którego zostało znalezione rozwiązanie optymalne wynosiło 1.80802e-160 natomiast dla drugiego modelu początkowego prawdopodobieństwo było większe i wynosiło 1.02023e-151 pomimo to algorytm nie znalazł dla niego rozwiązania optymalnego. Z powyższych badań wynika, że znalezienie rozwiązania optymalnego zależy w dużej mierze od parametrów modelu początkowego natomiast nie jest tak istotna wartość prawdopodobieństwa dla modelu początkowego.

Dla modeli, w których prawdopodobieństwa mają takie same wartości wag tzn.: $\delta j=1/n$, $\gamma ij=1/n$, $\pi_{ik}=1/K$, nie następuje modyfikacja wartości tych współczynników, dlatego parametry początkowe modeli powinny być wyznaczane w sposób losowy. Wynika to z własności estymatora oraz zostało sprawdzone eksperymentalnie.

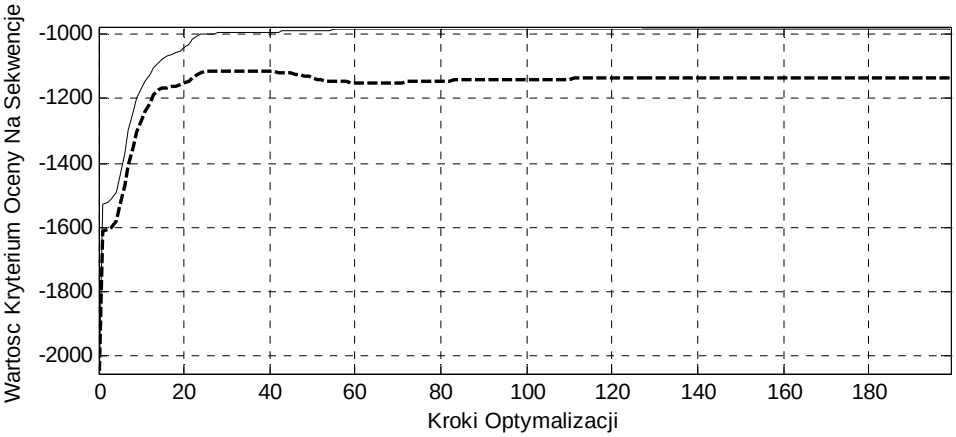
Po przeprowadzonej analizie można zaproponować metodę pozwalającą znaleźć średnią liczbę optimów lokalnych dla modelu HMM na podstawie częstości pojawiania się rozwiązana najlepszego podobnie jak dla modelu, dla których istniało rozwiązanie optymalne.

10.2. Efekty przeuczenia modeli HMM dla nielicznych danych treningowych

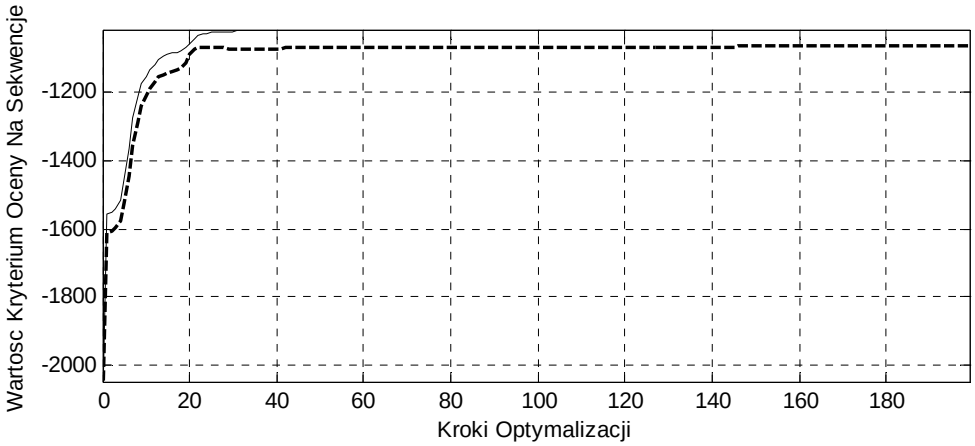
W podrozdziale tym zostały pokazane efekty przeuczenia się modeli HMM podczas optymalizacji. Efekty te są szczególnie widoczne w przypadku niewystarczającej liczby danych treningowych. Badania zostały przeprowadzone dla modeli podstawowych HMM z dziesięcioma węzłami. Efekty przeuczenia są szczególnie widoczne w przypadku sygnału dźwięku nie poddanego obróbce tj. w postaci czasowej. Modele optymalizowano do zbiorów sekwencji odpowiadających sylabie „*ce*” w postaci czasowej przez 500 iteracji. Na rysunkach (10.10), (10.11), (10.12), (10.13) przedstawione są wykresy przebiegu optymalizacji modeli HMM. Na rysunkach pokazana jest wartość kryterium oceny (9.2) w kolejnych krokach optymalizacji dla zbioru uczącego oraz sekwencji oceny odpowiednio linia ciąga oraz linia przerywana. Wykres pierwszy odpowiada optymalizacji do jedne sylaby „*ce*”, wykres drugi do dwóch sekwencji, ostatni do czterech sylab.



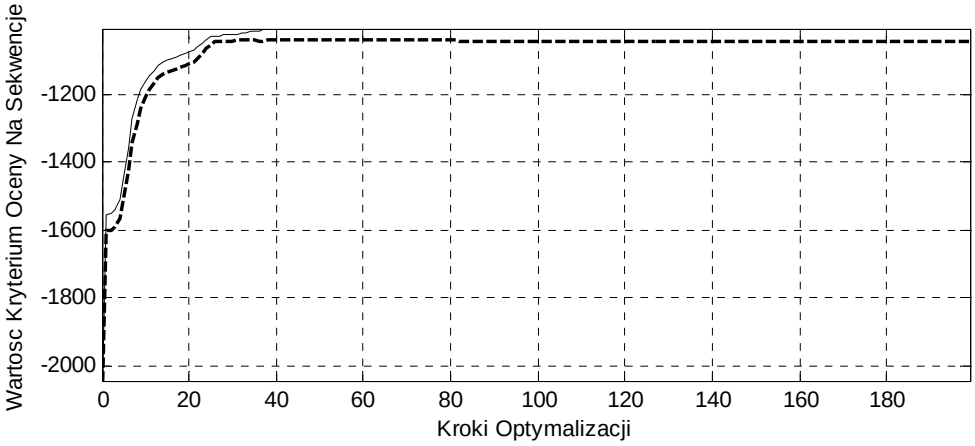
Rysunek 10.10. Optymalizacja modelu HMM do jednej sekwencji sylaby „*ce*”.



Rysunek 10.11. Optymalizacja modelu HMM do dwóch sekwencji sylaby „*ce*”.



Rysunek 10.12. Optymalizacja modelu HMM do trzech sekwencji sylaby „*ce*”.



Rysunek 10.13. Optymalizacja modelu HMM do czterech sekwencji sylaby „*ce*".

Z rysunków widać, że model na początku optymalizacji uczy się cech wspólnych wyróżniających sylabę, co jest widoczne w postaci poprawy obydwu kryteriów oceny zarówno dla sekwencji uczących jak i dla sekwencji oceny rysunki (10.10), (10.11). Po pewnej liczbie kroków optymalizacji kryterium dla zbioru oceny zaczyna się pogarszać, jest to spowodowane tym, że podczas zbyt długiej nauki przeważająca staje się poprawa cech indywidualnych sekwencji uczących. Ilość iteracji algorytmu, po której następuje przegięcie krzywej oceny zależy w dużej mierze od wielkości zbioru uczącego. Dla większej ilości zbioru uczącego efekty są mniej widoczne.

10.3. Porównanie wyników rozpoznawania sylab w postaci przebiegów czasowych oraz po przetwarzaniu metodą LPC oraz MFCC.

W tym podrozdziale przedstawione są wyniki optymalizacji modeli podstawowych HMM do wielokrotnych sekwencji obserwacji reprezentujących sylaby ze zdania „*Optymalizacja modeli procesów losowych*". Badania zostały przeprowadzone z wykorzystaniem sekwencji reprezentujących czternaście sylab. Każda sylaba został nagrana w sześciu wersjach. Pierwsze pięć wersji wykorzystane zostały do nauki modeli HMM, natomiast szósta wersja do oceny.

Sposób tworzenia sekwencji treningowych jest opisany w rozdziale (8). Sylaby były reprezentowane przez sekwencje dyskretnych obserwacji w postaci próbek dzwięku, oraz w postaci symboli uzyskanych metodą kwantyzacji wektorowej wektorów cech wyznaczanych metodami przetwarzania LPC oraz MFCC. W metodach przetwarzania zostały wyznaczone wektory o piętnastu składowych a następnie poddane kwantyzacji wektorowej.

Wykorzystane podczas badań modele HMM miały po 10 stanów oraz dyskretne rozkłady prawdopodobieństw w stanach o 32 wartościach. W tabelach (10.1)-(10.3) przedstawione są wyniki rozpoznawania sylab. Optymalizacja została przeprowadzana przez 2000 iteracji dla sekwencji czasowych oraz przez 200 iteracji dla sekwencji w postaci symboli.

		Sylaby													
		ce	cja	de	li	lo	ma	mo	op	pro	so	sów	ty	wych	za
Modele sylab	ce	7.64e-465	1.16e-544	2.55e-607	3.01e-708	2.42e-655	1.10e-881	3.61e-571	1.18e-737	3.78e-680	2.67e-588	4.35e-595	1.11e-759	2.06e-646	2.21e-780
	cja	2.03e-485	7.90e-519	4.30e-574	2.11e-743	9.62e-685	2.63e-717	1.39e-509	5.57e-664	1.15e-625	1.65e-508	1.56e-544	8.88e-761	1.40e-587	1.05e-705

	de	1.76e-485	2.37e-551	1.013e-513	2.19e-750	8.27e-647	1.16e-843	2.78e-507	1.66e-660	1.70e-616	1.76e-473	3.96e-536	3.69e-761	1.03e-692	1.55e-706
	li	2.59e-717	6.78e-776	7.26e-793	5.08e-640	1.07e-771	2.03e-845	1.57e-683	4.07e-892	1.83e-787	8.00e-839	2.64e-675	8.53e-663	5.29e-608	6.74e-989
	lo	1.38e-504	4.08e-524	2.93e-619	4.86e-663	9.79e-643	5.13e-707	1.60e-504	1.50e-694	2.05e-627	3.70e-557	1.49e-556	2.72e-746	5.88e-583	1.56e-695
	ma	2.14e-644	5.08e-610	1.34e-662	5.18e-693	6.55e-697	4.03e-673	5.69e-548	6.17e-767	5.69e-672	5.89e-560	3.62e-652	8.60e-849	2.37e-569	2.13e-768
	mo	2.59e-490	8.42e-536	1.27e-616	6.10e-785	2.43e-700	5.40e-744	1.20e-476	4.45e-719	2.32e-654	5.56e-522	2.67e-592	1.89e-829	4.89e-584	2.83e-704
	op	1.35e-670	3.70e-727	4.16e-707	5.56e-916	5.97e-890	6.24e-834	1.43e-585	1.18e-702	1.07e-705	1.05e-596	8.55e-732	7.18e-982	9.77e-606	2.59e-834
	pro	5.82e-490	2.72e-546	8.26e-589	4.79e-715	2.49e-669	1.40e-712	2.78e-502	7.36e-682	8.67e-620	1.74e-543	1.33e-595	1.48e-783	2.21e-559	2.57e-667
	so	5.19e-516	9.82e-529	3.09e-643	9.91e-769	2.77e-680	3.18e-743	3.37e-564	5.44e-710	1.94e-680	2.67e-544	1.88e-582	1.66e-802	4.04e-669	2.08e-708
	sów	3.00e-469	5.83e-569	5.08e-566	1.28e-707	4.10e-660	7.11e-818	4.23e-522	1.83e-684	7.41e-632	1.57e-555	8.51e-553	2.71e-707	6.43e-572	4.66e-769
	ty	7.20e-564	7.67e-650	3.49e-643	2.53e-693	1.29e-731	3.14e-940	2.58e-635	1.81e-773	1.82e-736	3.77e-686	1.14e-569	1.36e-664	1.71e-647	7.09e-920
	wych	5.06e-529	8.70e-625	2.78e-615	6.88e-729	2.16e-711	5.50e-809	1.75e-530	3.05e-765	1.18e-653	2.76e-591	1.90e-622	6.41e-815	1.12e-552	2.86e-857
	za	1.15e-531	4.96e-574	8.75e-587	2.24e-747	2.29e-688	4.99e-818	3.34e-525	9.85e-655	1.62e-607	1.72e-520	1.41e-578	1.51e-813	1.78e-592	8.76e-717

Tabela 10.1. Wyniki rozpoznawania sylab reprezentowanych przez przebiegi czasowe.

		Sylaby													
		ce	cja	de	li	lo	ma	mo	op	pro	so	sów	ty	wych	za
Modele sylab	ce	1.43e-652	3.09e-1148	2.86e-1163	1.64e-2401	1.64e-2401	1.63e-2401	1.64e-2401	1.60e-2401	1.64e-2401	4.05e-1854	5.41e-2196	4.81e-2311	1.64e-2401	4.85e-1214
	cja	2.19e-1290	1.44e-318	2.43e-1830	1.62e-2401	1.62e-2401	1.69e-2239	8.23e-2244	1.61e-2401	4.12e-2269	2.98e-1662	2.41e-2045	1.71e-2259	1.62e-2401	4.29e-1311
	de	2.14e-1632	2.16e-1843	2.81e-539	1.61e-2401	1.09e-2213	7.56e-1671	7.38e-2026	1.65e-2401	1.61e-2401	1.61e-2401	1.61e-2401	2.31e-2260	1.97e-2151	1.41e-1604
	li	1.60e-2401	1.60e-2401	4.10e-2065	4.28e-76	1.60e-2401	1.59e-2401	1.60e-2401	1.58e-2401	1.60e-2401	1.60e-2401	1.60e-2401	4.65e-903	1.60e-2401	1.60e-2401
	lo	1.62e-2401	1.62e-2401	6.31e-1951	1.62e-2401	2.10e-289	7.50e-1489	1.08e-1587	1.87e-1472	2.01e-1152	3.82e-1807	2.47e-1637	1.70e-2401	4.82e-1981	1.62e-2401
	ma	1.58e-2401	1.58e-2401	1.24e-1866	1.58e-2401	1.72e-1461	5.33e-407	1.46e-2051	1.57e-2401	1.00e-1564	1.60e-2002	1.58e-2401	1.57e-2401	1.13e-2140	1.58e-2401
	mo	2.53e-1914	6.95e-1720	3.32e-1847	1.61e-2401	1.97e-2221	1.32e-1418	1.73e-112	6.95e-1410	7.18e-2364	9.19e-1498	2.79e-1611	1.65e-2401	1.23e-2114	1.61e-2401
	op	1.57e-2401	1.57e-2401	1.57e-2401	1.57e-2401	1.57e-2401	1.57e-2401	4.83e-1412	1.78e-49	1.57e-2401	1.24e-1963	5.06e-1907	1.59e-2401	1.57e-2401	1.57e-2401
	pro	1.60e-2401	2.76e-2296	1.60e-2401	1.60e-2401	1.61e-866	3.73e-1916	1.05e-1896	1.61e-2401	4.03e-296	1.05e-1744	1.63e-2251	1.61e-2401	1.62e-2113	1.60e-2401
	so	3.24e-1884	5.81e-1532	3.24e-1884	1.66e-2401	6.30e-630	1.19e-1042	2.59e-1329	1.11e-1550	1.10e-655	2.21e-404	9.25e-1212	1.61e-2401	4.08e-2113	1.66e-2401
	sów	8.70e-2210	1.92e-2196	1.64e-2401	1.64e-2401	1.87e-2270	4.78e-2245	1.16e-1849	6.33e-996	1.32e-2222	2.00e-1655	9.81e-486	1.62e-2401	2.21e-1836	1.64e-2401
	ty	2.99e-2228	3.94e-2052	3.18e-1904	2.54e-1608	1.62e-2401	1.61e-2401	1.62e-2401	1.62e-2401	1.62e-2401	1.62e-2401	1.62e-2401	2.67e-407	1.62e-2401	7.57e-2095
	wych	1.58e-2401	1.07e-2324	3.51e-1869	1.58e-2401	1.96e-2221	1.23e-1468	1.25e-1857	1.58e-2401	4.33e-2254	1.79e-2290	4.33e-2254	1.58e-2401	3.78e-435	1.58e-2401
	za	1.80e-771	1.45e-1856	9.42e-871	1.61e-2401	2.62e-2247	8.95e-1839	2.26e-2113	1.67e-2401	1.61e-2401	1.61e-2401	1.61e-2401	4.82e-2267	2.30e-2234	1.26e-128

Tabela 10.2. Wyniki rozpoznawania sylab reprezentowanych przez symbole wyznaczone z wykorzystaniem metody LPC.

		Sylaby													
		ce	cja	de	li	lo	ma	mo	op	pro	so	sów	ty	wych	za
Modele	ce	4.96e-625	2.44e-643	1.56e-1051	1.69e-2401	3.26e-2399	1.67e-2401	2.85e-2113	1.64e-2401	2.55e-2091	2.54e-1543	8.60e-1881	1.38e-2302	1.30e-1999	2.29e-1048
	cja	6.19e-1032	3.26e-819	2.38e-1169	2.08e-1203	7.53e-2398	1.79e-2401	1.06e-2079	8.29e-2007	9.23e-2265	1.51e-1630	1.38e-1841	1.73e-2401	2.12e-1615	6.60e-1261

	de	1.05e-1064	1.66e-1428	3.27e-400	1.63e-2401	3.17e-2230	1.64e-2401	9.29e-1864	1.63e-2401	4.52e-2039	9.17e-2177	8.44e-2109	7.89e-2261	9.99e-519	6.43e-781
	li	1.58e-2401	1.58e-2401	7.34e-2283	8.51e-122	1.44e-2251	1.58e-2401	1.58e-2401	1.59e-2401	1.86e-2266	1.58e-2401	1.58e-2401	2.25e-941	1.76e-2190	1.58e-2401
	lo	1.62e-2401	1.62e-2401	3.13e-2377	1.62e-2401	2.28e-727	1.73e-1962	2.99e-2220	2.91e-1826	4.09e-923	1.62e-2401	1.40e-1745	1.62e-2401	1.18e-1862	3.14e-2377
	ma	9.51e-2301	1.76e-1938	1.49e-2099	1.61e-2401	9.04e-2209	3.27e-661	1.32e-1859	2.21e-2318	4.32e-2217	1.87e-1833	1.32e-2257	1.65e-2401	1.61e-2401	5.98e-1511
	mo	4.30e-2102	4.47e-1932	3.18e-1850	1.64e-2401	2.53e-2251	9.42e-1868	3.67e-997	1.34e-1468	1.53e-2072	1.45e-1972	3.30e-1851	1.75e-2401	4.58e-1538	5.89e-2168
	op	1.57e-2401	1.57e-2401	1.57e-2401	1.57e-2401	3.04e-2041	1.57e-2401	1.44e-1982	2.01e-514	9.66e-2374	6.57e-1496	8.79e-2174	1.59e-2401	1.57e-2401	8.98e-2157
	pro	1.03e-2267	1.65e-2401	1.93e-2287	1.65e-2401	2.55e-1517	9.81e-1890	2.58e-2208	1.57e-1690	1.06e-462	1.65e-2401	1.15e-1663	1.72e-2401	1.12e-1980	2.18e-2267
	so	1.49e-1651	1.09e-1331	1.27e-1866	1.69e-2401	2.45e-1890	1.68e-2401	2.99e-1871	1.69e-1180	7.17e-1359	1.033e-409	9.52e-1424	1.66e-2401	6.02e-1401	8.82e-2038
	sów	2.67e-2226	2.47e-2271	1.63e-2401	1.63e-2401	5.44e-1966	1.65e-2401	6.18e-1848	4.45e-1165	1.24e-2046	1.37e-1157	3.35e-1089	1.69e-2401	1.63e-2401	7.56e-2111
	ty	1.58e-1836	1.76e-2236	1.19e-1997	2.12e-2206	1.58e-2401	1.58e-2401	1.58e-2401	1.58e-2401	1.58e-2401	1.58e-2401	1.76e-2236	5.48e-821	1.58e-1836	1.65e-2036
	wych	5.86e-1809	1.95e-2207	4.38e-1455	1.60e-2401	4.64e-1761	2.10e-1316	1.36e-1819	1.64e-2401	9.18e-1763	1.60e-2401	1.79e-1890	2.91e-2195	1.58e-259	9.62e-1681
	za	2.72e-1333	1.22e-1851	2.57e-652	1.65e-2401	1.43e-2085	1.66e-2401	1.77e-1916	1.72e-2401	3.94e-1846	1.65e-2401	1.65e-2233	6.78e-2256	2.47e-1197	5.98e-619

Tabela 10.3. Wyniki rozpoznawania sylab reprezentowanych przez symbole wyznaczone z wykorzystaniem metody MFCC.

Dla dźwięków w postaci czasowej tabela (10.1) uzyskano znacznie gorsze wyniki rozpoznawania maksymalnie na poziomie ponad 55%, natomiast dla dźwięków po przetworzeniu sygnału tabele (10.2)-(10.3) uzyskane wyniki rozpoznawania są gorsze dla metody MFCC, dla której zdarzają się błędy pomimo reprezentacji widma w skali (*Mel*) uzyskano wyniki na poziomie ponad 95% natomiast dla metody LPC na poziomie 100%. Polami szarymi w tabelach zaznaczone są największe wartości w kolumnach.

Poza rozpoznawaniem sylab zostały przeprowadzone badania rozpoznawania cyfr od 0 do 9. W przypadku rozpoznawania liczb zostały uzyskane dużo lepsze wyniki rozpoznawania na poziomie 100% dla wyraźnie wypowiadanych nagrań. Uzyskanie lepszych wyników dla cyfr jest spowodowane tym, że jest ich mniejsza liczba niż badanych sylab, mają one większą długość oraz są bardziej zróżnicowane.

Można zauważyć, że ze zmniejszaniem liczby rozpoznawanych symboli, zwiększeniem ich długość oraz zróżnicowanie widm rozpoznawanych symboli, poprawia się uzyskiwane wyniki. Dlatego w przypadku wyznaczania podziału „*fonetycznego*" należy dążyć do uzyskania kompromisu pomiędzy trzema warunkami:

- liczba „*fonemów*" powinna być jak najmniejsza,
- zróżnicowanie widm „*fonemów*" powinno być jak największe,
- wyrazy powinny być reprezentowane przez jak najmniejszą liczbę „*fonemów*".

Pewnym rodzajem podziału językowego „*fonetycznego*" jest kwantyzacja wektorowa. Dobierając wielkość książki kodowej można określić z góry ilość „*fonemów*", dodatkowo spełniony jest warunek drugi, ponieważ wektory książki kodowej dobierane są w taki sposób, żeby maksymalizować odległość pomiędzy tymi wektorami, natomiast nie jest spełniony warunek trzeci, wyrazy zazwyczaj reprezentowane są przez dużą liczbę symboli (kodów). W związku z powyższym w przypadku wyznaczania reprezentacji słów w postaci stanów, co zostało zaproponowane w algorytmie podziału fonetycznego podrozdział (11.2), może się zdarzyć, że jedno słowo będzie miało bardzo wiele odmiennych reprezentacji co nie będzie korzystne.

W przypadku, jeżeli korzysta się z wyznaczonego podziału fonetycznego na podstawie znajomości języka zdarza się, że fonemy są do siebie podobne oraz zawierają widmo o podobnej charakterystyce, pomimo, że jest spełniony warunek trzeci.

Przeprowadzenie odpowiedniego podziału „*fonetycznego*" z pewnością ma istotny wpływ na uzyskiwane wyniki w systemie rozpoznawania mowy.

10.4. Porównanie wyników rozpoznawania cyfr dla modeli HMM

W podrozdziale tym zostały przedstawione wyniki rozpoznawania cyfr od 1 do 9 dla modelu podstawowego HMM, modelu z czasem przebywania oraz modelu sieciowego.

W badaniach do przetwarzania dźwięku został wykorzystana metoda LPC. Wektory zostały zakodowane metodą kwantyzacji wektorowej w postaci 256 symboli.

Modele początkowe zostały utworzone w sposób losowy, minimalna wartość dla współczynników modeli wynosiła 10e-4. Proces optymalizacji został przeprowadzany przez 50 iteracji. Każdy model był trenowany do pięciu sekwencji uczących, ocena rozwiązania została przeprowadzona również do pięciu sekwencji oceny.

10.4.1. Model podstawowy HMM

Zostały zastosowane modele HMM składające się z dziesięciu stanów. W tabelach (10.4), (10.5) przedstawione są wyniki rozpoznawania.

		Cyfry									
Modele cyfr		cztery	dwa	dziewięć	jeden	osiem	pięć	siedem	sześć	trzy	zero
	cztery	2.70e-647	1.63e-2352	4.03e-2407	4.02e-2407	4.02e-2407	1.25e-2372	4.97e-2334	1.52e-2382	3.06e-1971	1.74e-2197
	dwa	3.18e-2363	9.82e-438	8.81e-2342	1.02e-2234	3.04e-2240	3.62e-2407	5.19e-2187	3.62e-2407	3.63e-2407	8.85e-2140
	dziewięć	1.90e-2312	8.23e-2167	7.48e-565	2.07e-2362	3.52e-2385	5.39e-1204	1.45e-2363	1.63e-2073	4.38e-2407	4.49e-2407
	jeden	3.70e-2407	3.35e-2243	8.77e-2358	1.99e-494	3.70e-2407	3.70e-2407	2.72e-1884	3.71e-2407	3.69e-2407	3.70e-2407
	osiem	3.61e-2407	3.58e-2407	1.51e-2387	3.59e-2407	2.97e-475	7.49e-2331	7.41e-1774	3.62e-2407	3.56e-2407	3.60e-2407
	pięć	3.58e-2407	3.54e-2407	2.95e-1621	3.56e-2407	2.41e-2393	2.93e-459	2.04e-2363	6.20e-2149	3.52e-2407	3.57e-2407
	siedem	5.18e-2313	1.97e-2293	2.53e-2316	4.88e-1870	3.14e-1751	1.04e-2254	3.13e-582	1.39e-2367	4.01e-2407	4.13e-2407
	sześć	4.70e-2367	4.00e-2407	7.51e-2220	4.01e-2407	2.09e-2344	2.70e-2211	2.65e-2305	3.55e-534	7.08e-1912	4.01e-2407
	trzy	3.17e-2096	3.52e-2407	3.52e-2407	3.52e-2407	3.52e-2407	9.62e-2379	3.52e-2407	8.47e-2070	4.90e-428	3.52e-2407
	zero	3.58e-1956	5.85e-2173	3.29e-2407	3.29e-2407	3.29e-2407	3.29e-2407	3.30e-2407	3.30e-2407	3.28e-2407	1.87e-423

Tabela 10.4. Tabela pokazująca, która cyfra jest najlepiej modelowana przez poszczególne modele oraz które modele zostaną wybrane w procesie rozpoznawania sylab zaciemnione pola, (Tabela dla sekwencji uczących).

		Cyfry									
ele		cztery	dwa	dziewięć	jeden	osiem	pięć	siedem	sześć	trzy	zero

	cztery	1.79e-919	4.01e-2407	4.03e-2407	1.41e-2318	4.02e-2407	4.01e-2407	4.02e-2407	1.18e-2379	5.18e-2209	1.56e-2079
	dwa	1.07e-2357	1.38e-479	6.07e-2336	1.31e-2373	8.59e-2199	3.63e-2407	7.89e-2237	3.62e-2407	3.63e-2407	4.77e-2138
	dziewięć	1.25e-2329	3.85e-2243	1.24e-605	4.47e-2407	5.72e-2355	2.61e-1132	1.23e-2365	2.91e-2118	4.35e-2407	4.49e-2407
	jeden	2.16e-2371	1.68e-2286	2.24e-2353	1.02e-597	3.70e-2407	3.69e-2407	7.41e-1951	3.70e-2407	3.68e-2407	3.70e-2407
	osiem	3.60e-2407	3.57e-2407	3.62e-2407	3.59e-2407	2.89e-514	3.57e-2407	6.52e-1656	3.61e-2407	3.55e-2407	2.14e-2386
	pięć	3.57e-2407	3.54e-2407	1.84e-1763	3.56e-2407	6.38e-2357	7.93e-484	2.95e-2356	2.81e-2130	3.50e-2407	3.57e-2407
	siedem	1.47e-2245	1.99e-2326	1.01e-2330	4.04e-1917	3.30e-1687	4.04e-2407	1.08e-562	2.28e-2360	3.96e-2407	3.76e-2384
	sześć	1.81e-2320	4.00e-2407	2.28e-2372	4.01e-2407	3.47e-2255	3.80e-2204	1.03e-2354	1.39e-674	4.06e-1749	4.01e-2407
	trzy	1.22e-2203	3.52e-2407	3.52e-2407	3.52e-2407	3.52e-2407	3.52e-2407	3.52e-2407	1.36e-2108	1.39e-468	3.52e-2407
	zero	8.92e-1952	6.29e-2230	3.30e-2407	3.29e-2407	3.29e-2407	3.29e-2407	3.30e-2407	3.30e-2407	3.28e-2407	3.31e-465

Tabela 10.5. Tabela pokazująca, która cyfra jest najlepiej modelowana przez poszczególne modele oraz które modele zostaną wybrane w procesie rozpoznawania sylab zaciemnione pola, (Tabela dla sekwencji oceny).

W pierwszej tabeli (10.4) zostały umieszczone wyniki dla nagrania uczącego natomiast w drugiej (10.5) dla nagrania oceny. Widać, że nie występują błędy rozpoznawania ani dla nagrania uczącego ani dla nagrania oceny.

10.4.2. Model HMM z jawnym rozkładem prawdopodobieństwa pozostawania w tym samym stanie

Zostały zastosowane modele HMM składające się z dziesięciu stanów. Maksymalny czas pozostawania w stanie wynosił 5. W tabelach (10.6), (10.7) przedstawione są wyniki rozpoznawania.

		Cyfry									
Modele cyfr		cztery	dwa	dziewięć	jeden	osiem	pięć	siedem	sześć	trzy	zero
	cztery	5.44e-479	1.73e-2413	5.86e-2410	2.25e-2360	2.24e-2411	3.26e-2265	1.44e-2252	3.64e-2305	1.09e-1981	1.23e-2322
	dwa	3.14e-2338	8.11e-396	8.49e-2340	2.27e-2194	3.20e-2313	4.17e-2414	8.08e-2256	1.69e-2412	3.29e-2416	4.56e-2246
	dziewięć	3.46e-2414	1.06e-2133	4.16e-457	2.81e-2371	5.01e-2390	2.46e-1297	1.00e-2369	5.23e-2198	1.68e-2417	6.19e-2415
	jeden	1.21e-2353	4.53e-2171	3.35e-2408	8.54e-420	8.00e-2409	3.87e-2409	1.55e-1935	2.38e-2408	3.02e-2410	6.79e-2409
	osiem	6.44e-2414	3.13e-2416	9.10e-2389	2.48e-2415	1.35e-469	6.04e-2341	1.26e-1988	9.79e-2337	2.36e-2417	1.27e-2414
	pięć	1.80e-2373	7.41e-2410	4.06e-1643	6.69e-2409	4.08e-2391	3.55e-451	7.97e-2408	3.01e-2113	1.97e-2411	2.42e-2408
	siedem	1.19e-2261	3.53e-2236	1.29e-2303	5.89e-1993	6.96e-1586	3.97e-2335	6.03e-543	2.31e-2347	3.42e-2417	9.86e-2415
	sześć	3.39e-2373	1.92e-2416	4.42e-2150	1.16e-2415	7.24e-2303	1.61e-2177	2.48e-2335	7.66e-538	2.41e-2090	5.86e-2415
	trzy	4.21e-2174	1.38e-2408	2.23e-2407	4.42e-2408	9.24e-2408	6.95e-2269	1.74e-2407	1.53e-2025	8.77e-344	8.19e-2408
	zero	1.09e-1994	1.04e-2091	1.47e-2410	7.75e-2245	3.92e-2412	8.16e-2413	3.46e-2364	5.88e-2411	5.40e-2415	4.29e-438

Tabela 10.6. Tabela pokazująca, która cyfra jest najlepiej modelowana przez poszczególne modele oraz które modele zostaną wybrane w procesie rozpoznawania sylab zaciemnione pola, (Tabela dla sekwencji uczących).

		Cyfry									
		cztery	dwa	dziewięć	jeden	osiem	pięć	siedem	sześć	trzy	zero
Modele cyfr	cztery	2.06e-609	8.86e-2414	2.43e-2410	2.42e-2240	1.47e-2411	8.86e-2414	6.44e-2153	1.58e-2245	2.44e-2115	5.06e-2010
	dwa	1.68e-2298	4.88e-713	3.66e-2373	2.84e-2293	1.05e-2311	2.38e-2415	1.27e-2164	3.46e-2413	5.87e-2417	1.98e-2260
	dziewięć	4.03e-2415	3.06e-2262	2.73e-774	1.80e-2415	3.68e-2389	7.03e-1094	3.52e-2321	1.91e-2236	2.31e-2418	6.19e-2415
	jeden	3.93e-2306	1.00e-2256	2.16e-2408	4.54e-575	6.79e-2409	8.17e-2410	5.62e-1851	1.23e-2408	9.47e-2411	6.79e-2409
	osiem	8.59e-2415	1.70e-2416	1.91e-2413	3.79e-2415	9.36e-554	1.70e-2416	1.42e-1923	1.42e-2381	1.96e-2418	1.27e-2414
	pięć	1.86e-2408	3.54e-2410	4.88e-1777	9.85e-2409	2.49e-2361	2.30e-528	5.13e-2355	1.85e-2095	3.36e-2413	2.42e-2408
	siedem	5.81e-2210	2.13e-2320	6.77e-2319	1.64e-2072	2.38e-1510	1.29e-2416	2.03e-844	8.77e-2375	2.26e-2418	9.86e-2415
	sześć	9.09e-2369	1.25e-2416	6.22e-2317	1.92e-2415	3.54e-2312	6.52e-2100	1.08e-2313	5.89e-742	8.79e-1993	5.86e-2415
	trzy	1.29e-2213	8.17e-2409	1.84e-2407	5.39e-2408	8.19e-2408	8.17e-2409	1.39e-2407	2.93e-2039	2.46e-573	8.19e-2408
	zero	7.60e-2010	2.88e-2180	4.56e-2411	1.47e-2251	2.72e-2412	3.84e-2414	4.79e-2303	1.08e-2411	5.48e-2416	2.66e-623

Tabela 10.7. Tabela pokazująca, która cyfra jest najlepiej modelowana przez poszczególne modele oraz które modele zostaną wybrane w procesie rozpoznawania sylab zaciemnione pola, (Tabela dla sekwencji oceny).

W pierwszej tabeli (10.6) zostały umieszczone wyniki dla nagrania uczącego natomiast w drugiej (10.7) dla nagrania oceny. Widać, że nie występują błędy rozpoznawania ani dla nagrania uczącego ani dla nagrania oceny.

W przypadku modelu z czasem przebywania nie występują widoczne różnice w porównaniu z modelem podstawowym HMM.

10.4.3. Model sieciowy HMM

Zostały zastosowane modele sieciowe składające się z trzech węzłów. Każdy z pod modeli składał się z trzech stanów. Maksymalny czas pozostawania w węźle wynosił 5. W tabelach (10.8), (10.9) przedstawione są wyniki rozpoznawania.

		Cyfry									
		cztery	dwa	dziewięć	jeden	osiem	pięć	siedem	sześć	trzy	zero
Modele cyfr	cztery	9.38e-589	2.38e-2416	4.79e-2413	1.18e-2360	1.22e-2414	3.28e-2284	1.01e-2241	2.87e-2319	5.78e-2043	5.48e-2337
	dwa	6.78e-2343	2.03e-464	1.85e-2344	2.72e-2204	9.83e-2252	6.56e-2415	1.77e-2208	3.72e-2413	6.73e-2417	3.54e-2215
	dziewięć	1.92e-2413	1.52e-2177	9.31e-604	2.51e-2365	2.43e-2388	2.21e-1335	1.60e-2351	7.84e-2187	1.54e-2416	4.01e-2414
	jeden	3.78e-2362	1.35e-2184	3.75e-2413	1.68e-574	8.96e-2415	1.83e-2415	6.74e-1943	1.42e-2413	1.29e-2417	6.18e-2415
	osiem	3.86e-2415	7.87e-2417	4.51e-2390	3.46e-2417	3.21e-645	3.70e-2347	1.05e-2042	1.03e-2350	2.36e-2416	8.71e-2417
	pięć	1.26e-2388	2.54e-2416	5.59e-1702	1.89e-2415	7.53e-2389	5.55e-591	1.28e-2413	6.91e-2136	2.21e-2417	9.95e-2415

	siedem	4.06e-2221	1.42e-2273	2.95e-2283	1.59e-2035	5.23e-1679	2.11e-2322	9.10e-649	4.32e-2344	4.49e-2417	1.63e-2414
	sześć	1.02e-2376	2.59e-2417	8.13e-2121	3.88e-2416	2.87e-2250	5.92e-2194	5.92e-2273	1.44e-620	7.72e-2118	2.25e-2415
	trzy	5.50e-2204	2.68e-2416	5.18e-2413	1.87e-2415	1.35e-2414	1.48e-2289	1.19e-2413	6.70e-2058	7.12e-478	9.45e-2415
	zero	3.81e-1940	5.88e-2126	4.60e-2413	1.09e-2256	1.14e-2414	2.36e-2415	3.15e-2369	1.76e-2413	1.83e-2417	4.40e-537

Tabela 10.8. Tabela pokazująca, która cyfra jest najlepiej modelowana przez poszczególne modele oraz które modele zostaną wybrane w procesie rozpoznawania sylab zaciemnione pola, (Tabela dla sekwencji uczących).

		Cyfry									
		cztery	dwa	dziewięć	jeden	osiem	pięć	siedem	sześć	trzy	zero
Modele cyfr	cztery	1.70e-710	1.31e-2416	1.43e-2413	4.33e-2216	8.53e-2415	1.31e-2416	9.22e-2117	2.85e-2244	1.12e-2145	1.18e-2053
	dwa	2.45e-2301	1.16e-661	5.29e-2377	2.18e-2310	2.02e-2248	3.81e-2416	5.47e-2116	7.47e-2414	8.79e-2418	1.56e-2255
	dziewięć	2.81e-2414	4.12e-2251	4.31e-868	1.31e-2414	8.22e-2388	6.06e-1153	3.36e-2325	9.61e-2251	1.90e-2417	4.01e-2414
	jeden	4.11e-2314	3.79e-2240	1.09e-2413	4.63e-726	6.18e-2415	8.94e-2417	1.48e-1896	2.51e-2414	1.32e-2418	6.18e-2415
	osiem	2.64e-2416	3.83e-2420	8.28e-2415	6.12e-2415	5.54e-730	3.83e-2420	1.23e-1940	8.06e-2387	8.26e-2419	8.71e-2417
	pięć	6.89e-2415	1.47e-2416	3.52e-1835	2.96e-2415	3.58e-2358	8.24e-641	2.34e-2336	2.64e-2120	3.12e-2418	9.95e-2415
	siedem	2.20e-2158	4.84e-2300	4.03e-2316	2.10e-2099	9.87e-1643	2.83e-2416	6.82e-951	4.69e-2372	5.22e-2418	1.63e-2414
	sześć	6.97e-2373	3.23e-2417	2.09e-2292	3.71e-2416	2.16e-2298	1.70e-2099	2.67e-2300	9.67e-813	2.07e-2034	2.25e-2415
	trzy	6.17e-2237	1.56e-2416	1.56e-2413	2.84e-2415	9.45e-2415	1.56e-2416	5.05e-2414	8.81e-2083	9.74e-619	9.45e-2415
	zero	8.98e-1982	2.66e-2205	1.36e-2413	4.88e-2259	7.89e-2415	1.19e-2416	5.21e-2320	3.16e-2414	2.01e-2418	5.82e-672

Tabela 10.9. Tabela pokazująca, która cyfra jest najlepiej modelowana przez poszczególne modele oraz które modele zostaną wybrane w procesie rozpoznawania sylab zaciemnione pola, (Tabela dla sekwencji oceny).

W pierwszej tabeli (10.8) zostały umieszczone wyniki dla nagrania uczącego natomiast w drugiej (10.9) dla nagrania oceny. Widać, że nie występują błędy rozpoznawania ani dla nagrania uczącego ani dla nagrania oceny.

Dla wyraźnych nagrań dźwięku nie występują albo występują bardzo rzadko błędy rozpoznawania dla zastosowanych modeli HMM. W przypadku nagrania ze zmiennym szumem w tle albo niewyraźnej wypowiedzi błędy rozpoznawania pojawiając się częściej.

W przypadku omówionych wyników rozpoznawania izolowanych dźwięków nie można zauważyć znaczącej różnicy pomiędzy modelem podstawowym modelu sieciowy oraz model z czasem przebywania. W badaniach zostały zastosowane modele o podobnej liczbie parametrów.

10.5. Zastosowanie algorytmów sztucznej inteligencji do optymalizacji modeli HMM

Algorytmy opisane w podrozdziale (6) zostały zaimplementowane i wykorzystane do optymalizacji modeli HMM. Dzięki zastosowaniu algorytmów nie zostały uzyskane lepsze rozwiązania niż w przypadku algorytmu Bauma-Welcha. Uzyskane tymi metodami modele HMM nie dają dobrych wyników rozpoznawania.

Istnieje kilka przyczyn, które powodują nieprawidłowe działanie zaproponowanych algorytmów zarówno algorytmu ewolucyjnego jak i algorytmu (ILS).

Optymalizowany model HMM dla 10 stanów oraz 512 kolorów ma dokładnie 6120 parametrów postaci liczb *long double* posiadających precyzję 80 bitów. W przypadku reprezentacji bitowej uzyskuje się 489600 bitów reprezentujących modeli HMM, dodatkowo parametry modelu powiązane są ze sobą warunkami ograniczającymi.

Z powodu dużej liczby parametrów modelu HMM oraz ich powiązania przez warunki ograniczające konieczne jest wykorzystanie licznej populacji albo stosowanie algorytmu przez dużą ilość iteracji, natomiast złożoność obliczeniowa algorytmu wyznaczającego wartość funkcji oceny *forward-backward* powoduje, że czas obliczeń algorytmu jest nie akceptowalny.

W pracy zostały zastosowane dwa algorytmy tj.: algorytm genetycznym oraz algorytm (ILS), w przypadku zastosowanego tych algorytmów konieczne okazało się wprowadzenie kroku, o jaki modyfikowane są parametry. W zastosowanych algorytmach parametry modyfikowane są o ustaloną wartość, która maleje w kolejnych iteracjach.

Parametry modelu po optymalizacji zmieniają się o bardzo niewielką wartość w porównaniu z parametrami początkowymi. Dlatego z powodu wprowadzenia kroku w algorytmach, o który zmieniane są parametry modeli HMM utrudniony jest dokładny dobór wartości tych parametrów na poziomie precyzji liczb *long double*, a taki dobór parametrów zapewnia algorytm Bauma-Welcha.

W wyniku zastosowania zaproponowanych algorytmów nie udało się znaleźć minimum lokalnego, które jest znajdowane w wyniku zastosowania algorytmu Bauma-Welcha. Należy zaznaczyć, że do optymalizacji modeli HMM wykorzystywane są metody gradientowe, które stosowane są w algorytmach dyskryminacyjnych. Algorytmy gradientowe dla jednego modelu działają wolniej niż algorytm Bauma-Welcha, natomiast dają takie same wyniki końcowe.

11.Podsumowanie i wnioski

11.1. Otrzymane wyniki

Obecnie stosowane metody rozpoznawania mowy bazują na porównywaniu zarejestrowanych dźwięków z wzorcami w postaci modeli HMM. W pracy [8] wydanej przez (*Massachusetts Institute of Technology*) [178] uczelni, na której intensywnie prowadzone są badania nad rozpoznawaniem mowy można znaleźć zdanie mówiące, że stosowane wówczas systemy miały zdolność rozpoznawania porównywalną z kilkuletnim dzieckiem, sytuacja ta nie zmieniła się radykalnie w obecnej chwili.

Pomimo, że wyrażenie czy zdanie jest zrozumiałe dla dorosłego człowieka zdarza się, że system nie może sobie z nim poradzić. Człowiek potrafi odtworzyć usłyszane zdanie, pomimo, że nie słyszał wyraźnie pewnych jego części domyślając się niewyraźnie usłyszanych fragmentów z kontekstu, czego nie potrafią stosowane systemy.

W celu stosowania systemów wprowadzania tekstu do komputera takich jak (*Via Voice*), firmy IBM, (*Dragon*) firmy ScanSoft lub narzędzi alternatywnego wprowadzania danych Microsoftu konieczne jest nauczenie się wyraźnego mówienia do komputera w celu uzyskania zadawalających rezultatów. Poza tym podczas wprowadzania tekstu konieczne jest częste korzystanie z klawiatury w celu np. wybory alternatyw wyrazów, co jest bardzo uciążliwe.

Przez autora nie została znaleziona praca, w której omówione są wszystkie stosowane metody związane z zastosowaniem modeli HMM w systemach rozpoznawania mowy, prawdopodobnie zarówno z powodu złożoności stosowanych metod jak również z powodu pojawiania się prac zawierających nowe pomysły. W trzecim rozdziale pracy zostały omówione podstawowe algorytmy poprawiające dostępne w literaturze związane z wykorzystaniem modeli HMM.

W pracy zostały zrealizowane przyjęte tezy. Składają się na następujące wyniki cząstkowe:

1. przebadano skuteczność rozpoznawania izolowanych sylab oraz liczb w języku polskim z wykorzystaniem wzorców w postaci modeli HMM.
2. zastosowano algorytmy związanych z zaproponowanym modelem sieciowym HMM tj. algorytmu optymalizacji modelu sieciowego oraz algorytmu wyznaczania prawdopodobieństwa obserwacji dla tego modelu.
3. zaprojektowano dwa algorytmy sztucznej inteligencji do optymalizacji modeli HMM.

Wynikami autora są:

-Zastosowanie metody wyznaczania rozkładu „*prior*" stosowanego w algorytmie Bayesa tj. podstawowego algorytmu optymalizacji modeli HMM. Wymiernym wynikiem zastosowana metody jest zredukowanie ilości parametrów „*prior*" oraz zmniejszenie liczby działań w algorytmie w porównaniu z pracą np. [9] (podrozdział 3.5.2).

-Wniosek (3.108) (podrozdział 3.5.2).

-Wyznaczenie pochodnych rozkładu „*prior*" dla modelu niestacjonarnego opisanego w pracy [77] zaproponowaną metodą podziału (podrozdział 3.6.3).

-Opracowanie modelu sieciowego oraz algorytmów wyznaczenia prawdopodobieństwa obserwacji oraz algorytmu (EM) optymalizacji tego modelu (podrozdziały 4.1, 4.2).

-Pokazanie, że model sieciowy jest uogólnieniem teorii modeli HMM i algorytmy zastosowane do tego modelu sprowadzają się w skrajnych przypadkach do modelu podstawowego oraz modelu z jawnym prawdopodobieństwem pozostawania w danym stanie (podrozdziały 4.3 4.4).

-Opracowanie metod uczenia modelu sieciowego dla wielokrotnych sekwencji obserwacji oraz metod skalowania dla tego modelu (podrozdziały 4.6 4.7).

-Oszacowanie złożoności obliczeniowej dla zastosowanych algorytmów poprawiających (rozdział 5).

-Opracowanie algorytmów sztucznej inteligencji optymalizacji modeli HMM tj. algorytmu Ewolucyjnego oraz algorytmu ILS (rozdział 6).

-Propozycja zastosowania funkcji skalującej (8.53) w algorytmach przetwarzania sygnału (podrozdział 8.6).

-napisanie programu optymalizującego modele HMM, w którym zostały zastosowane algorytmy optymalizacji oparte na kryterium (ML) dla wielokrotnych sekwencji obserwacji dla modelu podstawowego oraz dla modelu sieciowego HMM. W programie zostały zastosowane dwa algorytmy przetwarzania sygnału tj: LPC, MFCC oraz kwantyzacja wektorowa VQ.

11.2. Propozycje dalszych badań

Przedstawiony w rozdziale (4) model sieciowy jest uogólnieniem wiążącym ze sobą modele HMM wykorzystywane do modelowania fonemów oraz model językowy. Złożoność obliczeniowa algorytmów estymacji parametrów zaproponowanego modelu sieciowego jest większa niż w metodach, w których do doboru parametrów modelu językowego stosuje się metody oparte na analizie tekstu. Natomiast złożoność w procesie rozpoznawania mowy ciągłej z wykorzystaniem Algorytmu Viterbiego jest identyczna.

Z wprowadzenia modelu sieciowego wynika kilka faktów:

a) Model ten jest uogólnieniem teorii dotyczącej modeli HMM.

b) Model sieciowy może zostać potraktowany jak model lingwistyczny pierwszego albo drugiego rzędu tj. (*Unigram model*) albo (*Bigram model*), wówczas nie jest konieczne rozróżnienie pomiędzy modelem językowym a modelami HMM reprezentującymi „*fonemy*".

c) W stosowanych systemach rozpoznawania mowy uwzględnia się czynniki statystyczne takie jak częstotliwość następstwa słów „*fonemów*" po sobie w modelu językowym, uwzględnia się również prawdopodobieństwo wystąpienia danego widma sygnału dźwięku oraz prawdopodobieństwa następstwa kolejnych widm po sobie w modelach HMM, co jest pokazane np. w pracy [9] na schemacie ze strony 19. Po wprowadzeniu modelu sieciowego oczywiste wydaje się wprowadzenie dodatkowego parametru rozkładu prawdopodobieństwa długości fonemów w zależności od ich następstwa. Prawdopodobieństwo to reprezentowane jest w modelu sieciowym przez rozkład prawdopodobieństwa pozostawania w danym węźle.

d) W stosowanych systemach rozpoznawania mowy stosuje się podział językowy na „*fonemy*" podział taki jest na początku dokonywany przez operatorów. Wyodrębnione „*fonemy*" są stosowane do wstępnego uczenia modelu językowego. W wyniku nauki model językowy nabiera zdolności podziału dźwięku (segmentacji) na „*fonemy*", do których był uczony, wówczas możliwe jest prowadzenie dalszej nauki bez udziału operatora. Jednak w przypadku błędów w klasyfikacji fonemów może nastąpić pogorszenie własności systemu, dlatego proces uczenia na etapie klasyfikacji (segmentacji) „*fonemów*" z wykorzystaniem algorytmu Viterbiego powinien być kontrolowany przez operatora. Zaproponowany model sieciowy może być uczony do dźwięków bez jakiegokolwiek wstępnego podziału jak również tworzenia modelu

językowego na podstawie analizy statystycznej tekstu. W wyniku nauki nastąpi automatyczny podział cech dźwięku pomiędzy stany. Dlatego model ten może zostać zastosowany do automatycznego podziału językowego.

Do przeprowadzenia automatycznego podziału językowego można zaproponować algorytm, w którym model początkowy złożony jest z pojedynczych stanów w węzłach sieci. Przy pomocy zaproponowanego modelu możliwe będzie przeprowadzenie podziału dla dowolnej przyjętej liczby pseudo „*fonemów*" tj. jednostek językowych odpowiadających węzłom. W dalszym procesie uczenia powinien być możliwy przyrost liczby stanów w węzłach. Przyrost liczby stanów powinien być zależny od błędów występujących w danym stanie w procesie uczenie. W wyniku zastosowania takiego algorytmu możliwe jest wyznaczenie podziału językowego na dowolną przyjętą ilość jednostek językowych.

W przypadku zastosowania powyższego algorytmu możliwe jest wprowadzenie kryterium podziału, dodatkowo w wyniku zastosowania algorytmu uzyskane modele jednostek językowych w węzłach sieci mogą mieć różna wielkość w odróżnieniu od systemu opisanego w pracy [9], w którym wszystkie modele HMM „*fonemów*" składają się z trzech stanów.

W wyniku zastosowania zaproponowanego algorytmu podziału fonetycznego możliwe będzie sprawdzenie jaki podział językowy tzn. na ile jednostek językowych da najlepsze wyniki rozpoznawania.

Po przeprowadzeniu automatycznego podziału konieczne będzie utworzenie słownika, w którym wyrazy zostaną zakodowane w postaci alternatyw sekwencji stanów.

Kolejnym interesującym zagadnieniem powinno być wykazanie słuszności tezy, że opisana metoda MLLR opisana w podrozdziale (3.9) stosowa w celu adaptacji do mówcy polegająca na wyznaczeniu przekształcenia liniowego wartości średniej oraz macierzy kowariancji jest równoważna doborowi odpowiedniej funkcji skalującej $K(f)$ na poziomie przetwarzania sygnału. Przykładowa funkcja skalująca dobierana na podstawia formantów została zaproponowana w podrozdziale (8.6).

12. Literatura

[1] Rabiner L., Schafer R.: *Digital Processing of Speech Signals*, New Jersey, Prentice Hall 1978 ISBN: 0132136031.

[2] Tadeusiewicz R.: *Sygnał Mowy*, Warszawa, Wydaw. Komunikacji i Łączności, 1988. ISBN: 8320607051.

[3] Rabiner L.R., Juang B.H.: *Fundamentals of Speech Recognition*, New Jersey, Prentice Hall 1993. ISBN: 0130151572.

[4] Deller J., Hansen J., Proakis J.: *Discrete-Time Processing of Speech Signals*, Macmilan Publisher Company, 1993. ISBN: 0023283017.

[5] Foster P., Schalk T.: *Speech Recognition: The Complete Practical Reference Guide*, Telecom Library, inc 1993. ISBN: 0936648392.

[6] Ian L. MacDonald and Walter Zucchini: *Hidden Markov and Other Models for Discrete- Valued Time Series*, Chapman and Hall 1997. ISBN 0412558505.

[7] Jelinek F.: *Statistical Methods for Speech Recognition*, The MIT Press. January 1998. ISBN: 0262100665.

[8] Manning C., Schutze H.: *Foundations of Statistical Natural Language Processing*, The MIT Press. 1999. ISBN: 0262133601.

[9] Becchetti C., Becchetti L.P., Ricotti L.P.: *Speech Recognition: Theory and C++ Implementation*, New York, J. Wiley 1999. ISBN: 0471977306.

[10] Jurafsky D., Martin J. H.: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, New Jersey, Prentice Hall 2000. ISBN: 0130950696.

[11] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon: *Spoken Language Processing*, Prentice Hall PTR, April 2001, ISBN: 0130226165.

[12] Zielinski T. P. *Od teorii do cyfrowego przetwarzania sygnałów.* Kraków Wydział EAIiE AGH, 2002. ISBN: 8388309552.

[13] Chou W., Juang B. H., Juang B. H.: *Pattern Recognition in Speech and Language Processing*, CRC Pr I Llc, January 2003, ISBN: 0849312329.

[14] Minker W., Bennacef S.: *Speech and Human-Machine Dialog*, Springer, May 2004, ISBN: 1402080360.

[15] Divenyi P.: *Speech Separation by Humans and Machines*, October 2005, ISBN: 1402080018.

[16] Bristow G.: *Electronic speech recognition: techniques, technology and applications*, McGraw-Hill, October 1986, ISBN: 0070079137.

[17] Yannakoudakis E. J., Hutton P. J.: *Speech synthesis and recognition*, Ellis Horwood Ltd, January 1987, ISBN: 0470209593.

[18] Ainsworth W. A.: *Speech recognition by machine*, Peter Peregrinus Ltd, July 1988, ISBN: 0863411150.

[19] Furui S.: *Digital Speech Processing, Synthesis and Recognition*, Marcel Dekker Inc, 1989, ISBN: 0824704525.

[20] Waibel A., Kai-Fu Lee, Kai-Fu: *Readings in Speech Recognition*, Morgan Kaufmann Pub, January 1990, ISBN: 1558601244.

[21] Huang X. D.: *Hidden Markov models for speech recognition*, Edinburgh Univ Pr, February 1991, ISBN: 0748601627.

[22] Laface P., Mori R D, North Atlantic Treaty Organization*: Speech Recognition and Understanding: Recent Advances, Trends and Applications*, January 1992, ISBN: 3540540326.

[23] Acero A.: *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Pub, January 1993, ISBN: 0792392841.

[24] Keller E.: *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art and Future Challenges*, John Wiley & Sons Inc, January 1994, ISBN: 0471944491.

[25] Antonio J. Rubio Ayuso, Juan M. Lopez Soler, Antonio J. Rubio Ayuso: *Speech Recognition and Coding: New Advances and Trends*, Springer Verlag, June 1995, ISBN: 3540600981.

[26] Jean-Claude Junqua, Jean-Paul Haton: *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, Kluwer Academic Pub, June 1996,. ISBN: 0792396464

[27] Lee C. H., Soong F. K., Paliwal K. K.: *Automatic Speech and Speaker Recognition: Advanced Topics*, Kluwer Academic Pub, June 1996, ISBN: 0792397061.

[28] Carson-Berndsen J.: *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition*, Springer, December 1997, ISBN: 0792348834.

[29] Renato De Mori: *Spoken Dialogues With Computers*, Academic Pr, January 1998, ISBN: 0122090551.

[30] Balentine B., Morgan D. P.: *How to Build a Speech Recognition Application: A Style Guide for Telephony Dialogues*, Enterprise Integration Group Inc, April 1999, ISBN: 0967127815.
[31] Deller J., Hansen J., Proakis J.: *Discrete-Time Processing of Speech Signals*, IEEE, January 2000, ISBN: 0780353862.
[32] Weinschenk S., Barker D. T.: *Designing Effective Speech Interfaces*, John Wiley & Sons Inc, January 2000, ISBN: 0471375454.
[33] Holmes J. N., Holmes J., Holmes W.: *Speech Synthesis and Recognition*, Taylor & Francis, January 2001, ISBN: 0748408568.
[34] Furui S.: *Digital Speech Processing, Synthesis, and Recognition*, Marcel Dekker Inc, February 2001, ISBN: 0824704525.
[35] Clark L.: *Voice Recognition With Software Applications*, Glencoe/McGraw-Hill Post Secondary, January 2002, ISBN: 0078226414.
[36] Kotelly B.: *The Art and Business of Speech Recognition: Creating the Noble Voice*, Addison-Wesley, January 2003, ISBN: 0321154924.
[37] Goronzy S.: *Robust Adaptation to Non-Native Accents in Automatic Speech Recognition*, Springer Verlag, January 2003, ISBN: 3540003258.
[38] Schroeder M. R.: *Computer Speech: Recognition, Compression, Synthesis*, Springer Verlag, October 2004, ISBN: 3540212671.
[39] Minker W., Buhler D., Dybkjaer L.: *Spoken Multimodel Human-computer Dialogue In Mobile Environments*, Kluwer Academic Pub, December 2005, ISBN: 1402030746.
[40] Gąciarz T.: *Hidden Markov Model (HMM) – opis modelu i algorytmów pod kątem wykorzystania w problemach rozpoznawania mowy i pisma*, Elektrotechnika, Tom 17, Zeszyt 1, Uczelniane Wydawnictwo Naukowo - Dydaktyczne AGH, Kraków, str. 17-31, 1998.
[41] Basztura Cz.: *Jak rozmawiać z komputerem ludzkim głosem,* Wrocław, Wydawnictwo Format 1993.
[42] Shomali M., Kapusta M., Gajer M.: *Zastosowanie niejawnych modeli Markowa w systemach automatycznego rozpoznawania mowy*, Elektrotechnika i Elektronika, Tom 18, Zeszyt 3, Uczelniane Wydawnictwo Naukowo - Dydaktyczne AGH, Kraków, str. 89-98, 1999.
[43] Szostek R., Suraj Z.: *Wykorzystanie niejawnych łańcuchów Markowa do analizy temporalnych systemów informacyjnych*, Zeszyty Naukowe Politechniki Rzeszowskiej, Matematyka z. 26, str. 263-273, Rzeszów 2002.
[44] Baum L.E., Petrie T.: *Statistical interference for probabilistic functions of finite state Markov chain*, Ann. Math. Stat., 37 pp. 1554-1563, 1966.
[45] Baum L.E., Egon J.A.: *An inequality with applications to statistical estimation for probabilistic functions of Markov process and to a model for ecology*, Bull. Amer. Meteorol. Soc., 73: pp. 360-363, 1967.
[46] Baum L.E., Sell G.R.: *Growth functions for transformations on manifolds*, Pac. J. Math., 27 (2): pp. 211-227, 1968.
[47] Baum L.E., Petrie T., Soules G., Weiss N.: *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains*, Ann. Math. Stat., 41 (1): pp. 164-171, 1970.
[48] Baum L. E., *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes*, Inequalities, 3: pp. 1-8, 1972.
[49] Jelinek F.: *Continuous Speech Recognition by Statistical Methods,* Proceedings of the IEEE, Vol. 64, pp. 532-536, 1976.
[50] Jelinek F., Mercer R. L.: *Interpolated estimation of Markov source parameters from sparse data*, in Patrtern Recognition in Practice, E.S. Gelesma and L.N. Kanal, Eds. Amsterdam, The Nerherlands: North-Holland, pp. 381-397, 1980.
[51] Liporace L. R.: *Maximum likelihood estimation for multivariate observations of Markov sources*. IEEE Transactions on Information Theory, IT-28: pp. 729-734, 1982.
[52] Poritz A. B.: *Linear predictive hidden Markov models and the speech signals*, In Proc. ICASSP 82 (Paris, France), pp. 1291-1294, May 1982.
[53] Bahl L.R., Jelinek F., Marcer R.L.: *A Maximum Likelihood Approach to Continuous Speach Recognition,* IEEE Trans. Pattern Analisis, Machine Intelligence, Vol. 5, pp. 179-190, 1983
[54] Redner R. A., Walker H. F.: *Mixture densities, maximum likelihood and the EM algorithm*. SIAM Review, vol. 26, pp. 195-239, 1984.
[55] Juang B. H.: *Maksimum likelihood estimation for mixture multivariate stochastic observations of Markov chains,* AT&T Tech, J., 64 (6): pp. 1235-1249, July-Aug. 1985.
[56] Juang B. H., Rabiner L. R.: *Moxture autotegressive hiddeen Markovmodels for speech signals*, IEEE Trans. Acoustics, Signal Proc., ASSP-33 (6) pp. 1404-1413, December 1985.
[57] Juang. B. H., Levinson S. E., Sondhi M. M.: *Maximum likelihood estimation for multivariate mixture observations of Markov chains*, IEEE Trans. Information Theory, Vol. IT-32, No.2, pp.307-309, March 1986.

[58] Bahl L. R., Brown P. F., deSouza P. V., Mercer R. L.: *A new algorithm for the estimation of hidden Markov model parameters*, Proc. ICASSP 88, pp. 493-496, New York, April 1988.
[59] Deng L., *A generalized HMM with state-conditioned. trend functions of time for the speech signal*. Signal Process., vol. 27, no. 1, pp. 65-78, April 1992.
[60] Formey G.D.: *The Viterbi algorithm,* Proceeding of the IEEE, Vol. 61, pp. 268-278, March 1973.
[61] Rabiner L.R., Wilpon J. G., Juang B.H.: *A Segmental k-means Training Procedure for Connected Word Recognition*, AT&T Technical Journal, pp. 21-40, May-June 1986.
[62] Juang. B. H., *The Segmental k-Means Algorithm for Estimating Parameters of Hidden Markov Models*, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 38, No. 9, pp. 1639-1641, September 1990.
[63] Deng L., Aksmanovic M., Sun X., Wu J.: *Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states*. IEEE Trans. on Speech & Audio Process., vol.2, no.4, pp.507-520, 1994.
[64] Dempster A. P., Laird N. M., Rubin D. B.: *Maximum likelihood from incomplete data via the EM algorithm*. J. of the Royal Statistical Society, Series B, pp. 34:1-38, 1977.
[65] Brown P. F., Lee C.H., Spohrer J. C.: *Bayesian adaptation in speech recognition*, Proc. ICASSP, pp. 761-764, 1983.
[66] Ferretti M., Scarci S.: *Large-vocabulary speech recognition with speaker-adapted codebook and HMM parameters*, In Proc. Eurospeech89, pp. 154-156, Paris, Sept 1989.
[67] Lee C.H., C. H. Lin C.H., Juang B.H.: *A study on speaker adaptation of the parameters of continuous density hidden Markov models*, IEEE Trans. Acoustics, Speech, and Signal Processing, vol.39, no.4, pp.806-814, April 1991.
[68] Gauvain J. L., Lee C. H.: *Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models*. Proc. DARPA Speech and Natural Language Workshop, Pacific Grove, CA, pp. 272-277, Feb. 1991.
[69] Gauvain J. L., Lee C.H.: *Bayesian learning for hidden Markov model with Gaussian mixture state observation densities*. Speech Communication, 11: pp. 205-213, 1992.
[70] Gauvain J. L., Lee C.H.: *MAP estimation of continuous density HMM: Theory and applications*. In Proc. DARPASpeech and Natural Language Workshop, pp. 185-190, Feb. 1992.
[71] Lee C.H., Gauvain J. L.: *Speaker adaptation based on MAP estimation of HMM parameters*. In Proceedings of IEEE Conference on Audio Speech and Signal Processing, volume 2, pp. 558-561, 1993.
[72] Lee C. H., Gauvain J.L.: *Bayesian Adaptive Learning and MAP estimation of HMM*. Kluwer Academic Publishers, Boston 1993.
[73] Lee C. H., Gauvain J. L.: *Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains*, IEEE Trans. On Speech and Audio Processing. vol. 2 no. 2 April 1994.
[74] Huo Q., Chan C., Lee C. H.: *Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition*", Proc. IEEE Trans. Speech and Audio Processing, Vol. 3, No. 5, pp. 334-345, 1995.
[75] Buttigieg V., P.G. Farrell P.G.: *A maximum a-posteriori (MAP) decoding algorithm for variable-length error-correcting codes,* in Codes and cyphers: Cryptography and coding IV, Essex, England, The Inst. of Mathematics and its, pp. 103-119, April 1995.
[76] Kenny P., Gilles B., Pierre D.: *Bayesian adaptation revisited*, In *ASR-2000*, pp. 112-119, 2000.
[77] Rathinavelu C., Deng L.: *A maximum a posteriori approach to speaker adaptation using the trended hidden Markov model*, IEEE Transactions on Speech and Audio Processing, Vol.9, No.5, pp. 549-557, July 2001.
[78] Heckerman D., Geiger D., Chickering D.M.: *Learning Bayesian networks: The combination of knowledge and statistical data*, Machine Learning 20 pp. 197-243, also available as Microsoft Research Technical Report MSR-TR-94-09, 1995.
[79] Heckerman D.: *A Bayesian approach to learning causal networks*. In Proc. Eleventh Conf. on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA, 1995.
[80] Heckerman, D.: *A tutorial on learning with Bayesian networks*. In Jordan, M. I., ed., Learning in Graphical Models, Kluwer 1996.
[81] Chickering D. M., Heckerman D.: *Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables*. Machine Learning, 29(2-3): pp. 181-212, 1996.
[82] Chickering D., Heckerman D., Meek C.: *A Bayesian approach to learning Bayesian networks with local structure*. In Proceedings of UAI-97, pp. 80-89, 1997.
[83] Geoffrey G. Z., Stuart J. R.: *Speech recognition with dynamic bayesian networks*. 1998.
[84] Deviren M., Daoudi K.: *Structural learning of dynamic bayesian networks in speech recognition*," in Eurospeech, 2001.
[85] K. Livescu, J. Glass, and J. Bilmes, *Hidden feature models for speech recognition using dynamic Bayesian networks*. Eurospeech, Geneva 2003.
[86] Korkmazsky F., Deviren M., Fohr D., Illina I.: *Hidden Factor Dynamic Bayesian Networks for Speech Recognition*. Proceedings of International Conference on Spoken Language Processing, ICSLP'04, 2004.

[87] Deviren M., Daoudi K.: *Continuous Speech Recognition Using Dynamic Bayesian Networks: A Fast Decoding Algorithm*. In Advances in Bayesian Networks. Physica Verlag (Springer), 2004.

[88] Biing-Hwing J., Rabiner L.R.: *Mixture autoregressive hidden Markov models for speach signals*. IEEE Trans. on Acoustics, Speech, Signal Processing, ASSP-33 (6), pp. 1404-1413, December 1985.

[89] Juang B. H.,.Rabiner L.R.: *Mixture Autoregressive Hidden Markov Models for Speaker Independent Isolated Word Recognition*, Conference Record 1986 IEEE International Conference on Acoustics, Speech, and Signal Processing, Paper 2.1, pp. 41-44, April 1986.

[90] Katagiri S., Lee C. H., and Juang B. H.: New *Discriminative Training Algorithms Based on the Genralized Probabilistic Descent Merhod*, Proc. 1991 Workshop Neural Networks for Signal Processing, pp. 299-308, IEEE Signal Processing Society, Princeton, September 1991.

[91] Bagos P.G, Liakopoulos T., Hamodrakas S.J.: *Faster gradient descent training of Hidden Markov Models, using individual learning rate adaptation*, ICGI 2004 pp. 40-52, 2004.

[92] Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L.: *Maximum mutual information estimation of hidden Markov model parameters for speech recognition*. In Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 49-52, Tokyo 1986.

[93] Yen-Lu Chow: *Maximum Mutual Information estimation of HMM parameters for continuous speech recognition using the N-best algorithm*, in Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing, pp. 701-704, 1990.

[94] Normandin Y.: *Hidden Markov Models, Maximum Mutual Information Estimation and the Speech Recognition Problem*, Ph.D. Thesis, McGill University, Montreal 1991.

[95] Normandin Y.: *Maximum Mutual Information Estimation of Hidden Markov Models, in Automatic Speech and Speaker Recognition",* (Lee C.H., Soong F., Paliwal K.). Kluwer Academic Publisher, pp. 57-82, 1996.

[96] Singer Y., Warmuth M.K.: *Training algorithm for hidden Markov models using entopy based distance function.* In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advence in natural Information Processing System*s 9 (*NIPS*), pp. 641-647. Morgan Kaufman Publisher, 1996.

[97] Valtchev V., Odell J.J., Woodland P.C., Young S.J.: *MMIE Training of Large Vocabulary Speech Recognition Systems*, Speech Communication, Vol. 22, pp. 303-314, 1997.

[98] Zheng J., Butzberger J., Franco H., Stolke A.: *Improved maximum mutual information estimation training of continuous density HMM's*, in Proc. 7th Eur. Conf. Speech Communication and Technology, Aalborg, Denmark, September 2001.

[99] Gravier G., Axelrod S., Potamianos G. Neti C.: *Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR*. in IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, 2002.

[100] Mohamed Kamal Omar, Hasegawa-Johnso M.: *Maximum Conditional Mutual Information Projection For Speech Recognition*, In EUROSPEECH-2003, pp. 505-508, 2003.

[101] Droppo J., Mahajan M., Gunawardana A., Acero A.:. *How to Train a Discriminative Front End with Stochastic Gradient Descent and Maximum Mutual Information*, in Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding. Puerto Rico, December 2005.

[102] Droppo J. Acero A.: *Maximum Mutual Information SPLICE Transform for Seen and Unseen Conditions*, in *Proc. of the Interspeech Conference*. Lisbon, Portugal, September 2005.

[103] Ephraim Y., Dembo A., Rabiner L. R.: *A minimum discrimination information approach for hidden Markov modeling*, IEEE Trans. Inform. Theory, vol. 35, pp. 1001-1013, 1989.

[104] Ljolje A., Ephraim Y., Rabiner, L. R.: *Estimation of Hidden Markov Model Parameters by Minimizing Empirical Error Rate*. Proceedings of the IEEE, ICASSP-90, Vol. 2, pp. 709-712, 1990.

[105] Juang B.H., Katagiri S.: *Discriminative Learning for Minimum Error Classification*, IEEE Transactions on Signal Processing, Vol. 40, No. 12, pp. 3043-3054, December 1992.

[106] McDermott E., Katagiri S.: *Prototype-based minimum classification error-generalized probabilistic descent training for various speech units*, Computer Speech and Language, vol.8, No.8, pp.351-368, 1994.

[107] Rathinavelu C. Deng L.: *Speech trajectory discrimination using the minimum classification error learning*. In IEEE Trans. on Speech and Audio Processing, 1997.

[108] Markov K. Nakagava S. Nakamura S.: *Discriminative Training Of HMM Using Maximum Normalized Likelihood Algorithm*. ICASSP2001, Vol.1, pp. 497-500, 2001.

[109] Wang X., Paliwal K.K.: *Generalized MCE Training Algorithm for Feature Dimensionality Reduction*, Microelectronic Engineering Research Conference, Brisbane, Australia, Nov. 2001.

[110] Qi-Li, Juang B.H.: *A new algorithm for fast discriminative training*, Proc. ICASSP-2002, IEEE, Orlando, FL., May 2002.

[111] Wang X., Paliwal K.K.: *A Modified Minimum Classification Error (MCE) Training Algorithm for Dimensionality Reduction*, Springer The Journal of VLSI Signal Processing, Volume 32, Numbers 1-2, August 2002, pp. 19-28(10), 2002.

[112] Juang, B.H., Q. Li.: *A new algorithm for fast discriminative training,* Proc. ICASSP-2002, IEEE, Orlando, FL., May 2002.

[113] Loggeter C.J., Woodland P.C.: *Speaker adaptation of continuous density HMMs using multivariate linear regression*. Proc. ICSLP 1994, pp. 451-454 1994.
[114] Loggeter C.J., Woodland P.C.: *Speaker adaptation of HMMs using linear regression*. Cambridge University, Cambridge, UK, Tech. Rep. CUED/F-INFENG/ TR. 181, June 1994.
[115] Digalakis V., Neumeyer L.: *Speaker Adaptation Using Combined Transformation and Bayesian Methods*, Proc. ICASSP '95 1995.
[116] Gales M.J.F., Woodland P.C.: *Mean and Variance Adaptation Within the MLLR Framework*, *Computer Speech & Language*, Vol. 10, pp. 249-264, Woodland 1996.
[117] Gales M.J.F.: *Maximum Likelihood Linear Transformations for HMM-based Speech Recognition*, Tech. Rep., CUED/FINFENG/TR291, Cambridge Univ., 1997.
[118] Sam-Joo Doh, Richard M. Stern.: *Weighted Principal Component MLLR For Speaker Adaptation.* Pittsburg, PA 15213, USA 1999.
[119] Legetter C. Woodland P.: *Flexible speaker adaptation using maximum likelihood linear regression*, Speech and Audio Processing, IEEE Transaction, Volume: 7, Issue: 3 pp. 253-261 May 1999.
[120] Doh, S. J., Stern R.: *Inter-class MLLR for speaker adaptation Acoustics*, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on, Volume: 3, On pp: 1543-1546 vol.3, 2000.
[121] *Au-Yeung* S.K. Siu M.: *Maximum Likelihood Linear Regression Adaptation for the Polynomial Segment Models*, To appear in *IEEE Signal Processing Letters*, October 2006.
[122] Markel J.D., Gray .A.H.: *Linear Prediction of speech*, Berlin, Springer-Varlang 1976.
[123] Dautrich B. A., Rabiner L. R., Martin T. B..: *On the Use of Filter Bank Features for Isolated Word Recognition, Conference Record 1983 International Conference on Acoustics, Speech, and Signal Processing,* Paper 22.11, pp. 1061-1064, April 1983.
[124] Davis S.B., Mermelstein P.: *Comparision of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*, IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-28 (4): pp. 357-366, August 1980.
[125] Eric H. C. Choi: *On Compensation the Mel-Frequency Cepstral Coefficients for Noisy Speech Recognition.* Australia Australian Computer Society, 2006.
[126] H. Hermansky H.: *Perceptual linear prediction (plp) analysis of speech*, Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738-1752, April 1990.
[127] H. Hermansky H., Morgan N., Bayya A., Kohn P.: *Rasta-PLP speech analysis*. In Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing, volume 1, pp. 121-124, San Francisco 1992.
[128] Hönig F,. Stemmer G., Hacker, C., Brugnara, F.: *Revising Perceptual Linear Prediction (PLP)*, In INTERSPEECH-2005, pp. 2997-3000, 2005.
[129] Krishnan M., Neophytou C., Prescott G.: *Wavelet transform speech recognition using vector quantization*, dynamic time wraping and articicial neural networks. Preprint, 1994.
[130] Choueiter G. Glass J.: *A Wavelet and Filter Bank Framework for Phonetic Classification*, Proc. ICASSP, Philadelphia, March 2005.
[131] Chavy C., Gas B., Zarader J. L.: *Discriminative coding with predictive neural networks*. International Conference on Artificial Neural Networks (ICANN 99'), Edimburg, Scottland, pp 216-220 September 1999.
[132] Chavy C, Gas B et Zarader J.L.: *Neural predictive coding applied to noisy phoneme*. International Joint Conference on Neural Networks (IJCNN 99') USA Washington, July 1999.
[133] Gas B., Jean-Luc Zarader J., Chavy C.: *Neural predictive coding for speech discriminant feature extraction: The DFE-NPC*. ESANN 2002: pp. 275-280, 2002.
[134] Haeb-Umbach R., Ney H.: *Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition*, In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pp. 13-16, 1992.
[135] Aubert X., Haeb-Umbach R., Ney H.: *Continuous Mixture Densities and Linear Discriminant Analysis for Context -Dependent Acoustic Models*, In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pp. 648-651, 1993.
[136] Haeb-Umbach R., Geller D., Ney H.: *Improvements in Connected Digit Recognition using Linear Discriminant Analysis and Mixture Densities*, In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pp. 239-242, 1993.
[137] Eisele T., Haeb-Umbach R., Langmann D.: *A Comparative Study of Linear Feature Transformation Techniques for Automatic Speech Recognition*, " Proc. Int'l Conf. Spoken Language Processing, vol. I, pp. 252-255, October 1996.
[138] Ruske G., Faltlhauser R., Pfau T.: *Extended linear discriminant analysis (ELDA) for speech recognition*, In *ICSLP-1998*, paper 0100 Thilo 1998.

[139] Yongxin Li, Gao Yuqing, Erdogan Hakan: *Weighted pairwise scatter to improve linear discriminant analysis*, In *ICSLP-2000*, vol.4, pp. 608-611, 2000.
[140] Marcel Katz, Hans-Günter Meier, Hans Dolfing, Dietrich Klakow: *Robustness of Linear Discriminant Analysis in Automatic Speech Recognition*. ICPR (3): pp. 371-374, 2002.
[141] Burget, Lukas: *Combination of speech features using smoothed heteroscedastic linear discriminant analysis*, In *INTERSPEECH-2004*, pp. 2549-2552, 2004.
[142] Bing Zhang., Spyros Matsoukas.: *Minimum Phoneme Error Based Heteroscedastic Linear Discriminant Analysis for Speech Recognition*, IEEE International Conference on Acoustics, Speech, and Signal Processing, March 2005.
[143] H Erdogan H.: *Regularizing linear discriminant analysis for speech recognition*, Interspeech - Eurospeech, Lisbon Portugal, September 2005.
[144] Georg Stemmer, Fabio Brugnara: *Integration of Heteroscedastic Linear Discriminant Analysis (HLDA) into Adaptive Training*, IEEE International Conference on Acoustics, Speech, and Signal Processing May 2006.
[145] Doh S. J., Stern R. M.: *Weighted principal component MLLR for speaker adaptation*, Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 88, 1999.
[146] Huseyin M., Ertunc, Kenneth A. Loparo, Hasan Ocak.: *Tool wear condition monitoring in drilling operations using hidden Markov models (HMMs),* International Journal of Machine Tools & Manufacture 41 1363-1384, 2001.
[147] Michale M. Yin, Jason T. L. Wang.: *Effective hidden Markov models for detecting splicing junction sites in DNA sequences,* Information Sciences, pp. 139-163, 2001.
[148] Wei W., Wang B., Towsley D.: *Continuous-time hidden Markov models for network performance evaluation,* Performance Evaluation 49 pp. 129-146, 2002.
[149] Dehert M., Helm W.E, Hutt M.-Th.: *A discrete autoregressive process as a model for short-range correlations in DNA sequences.* Physica. A 327 pp. 505-553, 2003.
[150] Fang-Sheng Chen, Chih-Ming Fu, Chung-Lin Hung.: *Hand gesture recognition using a real-time tracking method and hidden Markov models*, Image And Vision Computing 21 745- 758 2003
[151] Comprox A. C., Gautier R., Tuffery P.: *A Hidden Markov Model Derived Structural Alphabet for Proteins,* K. Mol. Biol 339, pp. 591-605, 2004.
[152] Woo Sung Kim, Rae-Hong Park: *Off-line recognition of handwritten Korean and alphanumeric characters using hidden Markov models,* Pattern Recognition, Vol. 29, No. 5, pp. 845-858, 1996.
[153] Se-Cahang Oh, Jin-Young Ha, Jin-H. Kim.: *Context dependent search in interconnected hidden Markov model for unconstrained handwriting recognition,* Patetern Recognition Vol. 28, No. 1, pp. 1693-1704, 1995.
[154] Bunke H., Roth M., Schukat-Talamazzini E.G.: *Off-line cursive handwriting recognition using hidden markov models,* Patetern Recognition Vol. 28, No. 9, pp. 1399-1413, 1995.
[155] Hee-Seon Park, Seong-Whan Lee: *Off-line recognition of large-set handwritten characters with multiple hidden Markov models,* Pattern Recognition, Vol. 29, No. 2, pp. 231-144, 1996.
[156] Hang Joon Kim, Kyung Hyun Kim, Sang Kyoon Kiom, Jong Kook Lee: *On-line recognition of handwritten Chinese characters based on hidden Markov models,* Pattern Recognition, Vol. 30, No. 9, pp. 1489-1500, 1997.
[157] Dehghan M., Faez K., Ahmadi M., Shridhar M.: *Unconstrained Farsi handwritten word recognition using fuzzy vector quantization and hidden Markov model,* Pattern Recognition Letters 22 pp. 209-214, 2001.
[138] Said Ramdane, Bruno Taconet, Abderrazak Zahour.: *Classification of forms with handwritten fields by planar hidden Markov models,* Patetern Recognition 36 pp. 1045-1060, 2003.
[159] Somaja Alma'added, Colin Higgins, Dave Elliman: *Off-line recognition of handwritten Arabic words using multiple hidden Markov models,* Knowledge-Based Systems 17 pp. 75-79, 2004.
[160] Thanaruk Theeramunkong, Chaint Wongtapan: *Off-line isolated handwritten Thai OCR using island-based projection with n-gram model and hidden Markov,* Information Processing and Managment 41 pp. 139-160, 2005.
[161] Schafer R. W., Rabiner L. R.: *System for Automatic Formant Analysis of Voiced Speech, Journal of the Acoustical Society of America,* Vol. 47, No. 2, pp. 634-648, February 1970.
[162] Parham Zolfaghari, Tony Robinson. *Formant Analysis Using Mixtures of Gaussians*. Cambridge University Engineering Department. International Conference on Speech and Language Processing, 1996.
[163] Schafer R. W., Jackson K., Dubnowski J. J., Rabiner L. R.: *Detecting the Presence of Speech Using ADPCM Coding, IEEE Trans. on Communications,* Vol. COM-24, pp. 563-567, May 1976.
[164] Linde Y., Buzo A., Gray R.M.: *An Algorithm for vector quantizer design,* IEEE Trans. Communications, COM-28: pp. 84-95, January 1980.

[165] L. R. Rabiner, M. M. Sondhi and S. E. Levinson.: *A Vector Quantizer Incorporating Both LPC Shape and Energy*. Conference Record 1984 International Conference on Acoustics, Speech, and Signal Processing, pp. 17.1.1 - 17.1.4, March 1984.

[166] Bahl L.R., Gapalakrishnan P.S.: *Fast Match Techniques, in Automatic Speech and Speaker Recognition,* (Lee C.H., Soong F., Paliwal K.). Kluwer Academmic Publisher, pp. 413-428, 1996.

[167] Juang, B.H., Li Q., Lee C.H., Zhou Q., Soong F.K: *On speaker authentication*, IEEE Workshop on Automatic Identification Advanced Technologies, Stony Brook, NY, pp.3-6, November 1997.

[168] Juang, B.H., Li Q., Lee C.H., Zhou Q., Soong F.K.: *Recent advancements in automatic speaker authentication*, IEEE Robotics & Automation Magazine, March 1999.

[169] Juang, B.H., Peter Li, Book Chapter.: *Speaker Authentication, Pattern Recognition in Speech and Language Processing*, Wu and Juang (ed.), CRC Press., 2002.

[170] Ueberla Joerg: *Analizing and improving Statistical Language Models For Speech Recognition*. A theses submitted in partial fulmfillmnet of the requirements for the degree of Doctor of Philosophy in the School of Computing Science in Simon Fraser University, May 1994.

[171] Tabelskis Joe: *Speech Recognition using Neural Networks*, Submitted in partial fulmfillmnet of the requirements for a degree of Doctor of Philosophy in School of Computer Science Carnegie Mellon University, May 1995.

[172] Juliam James Odell: *The Use of Context in Large Vacabulaty Speech Recognition*, Dissertation submitted to the University of Cambridge for the degree of Doctor of Philosophy, March 1995.

[173] Ashvin Kannan: *Adaptation of Spectral Trajectory Models For Large Vacabulaty Continuous Speech Recognition*. Submitted in partial fulmfillmnet of the requirements for the degree of Doctor of Philosophy, 1997.

[174] Mike Lincoln: *Characterization of speakesr for Improved Automatic Speech Recognition*, A theses submitted for the Degree of Doctor of Philosophy in the School of Information Systems, University of East Anglia, Nortwich, 1999.

[175] Goodman Graham Laurence.: *A Comparison of Informative and Discriminative Estmation of Parameters for Classifier Training*. Submited for examination for the degree of Doctor of Philosophy University of South Australia, 2000.

[176] Kevin Patrick Murphy: *Dynamic Bayesian Networks*. A dissertation submitted in partial satisfaction of the requrements for the degree of Doctor of Philosophy in Computer Science of the University of California, Berkeley, 2002.

[177] Nguyen Chien Phu: *A Study on Effective Algorithms for Temporal Decomposition of Speech*, Submitted to Japan Advenced Institiute of Science and Technology in partial fulfillmnet of the requirements for the degree of Doctor of Philosophy, 2003.

[178] Massachusetts institute of technology.
http://groups.csail.mit.edu/sls/sls-blue-noflash.shtml

[179] Microsoft Research group.
http://research.microsoft.com/research/srg/

[180] University of Washington.
http://ssli.ee.washington.edu/

[181] IBM Research on HMM (Hidden Markov Model).
http://www.research.ibm.com/trl/projects/speech/index_e.htm

[182] The center of spoken language University of Colorado.
http://cslr.colorado.edu/

[183] Strona z narzędziami do HMM w C++.
http://htk.eng.cam.ac.uk/