BIULETYN INSTYTUTU AUTOMATYKI I ROBOTYKI
NR 23, 2006

Zastosowanie metod grupowania sekwencji czasowych w rozpoznawaniu mowy na podstawie ukrytych modeli Markowa

Tomasz PAŁYS

Zakład Automatyki, Instytut Teleinformatyki i Automatyki WAT,
ul. Kaliskiego 2, 00-908 Warszawa

STRESZCZENIE: Artykuł dotyczy problemu tworzenia ukrytych modeli Markowa na podstawie zarejestrowanych wypowiedzi. Kluczowym problemem jest wyznaczenie zbioru stanów modelu Markowa. Przyjęto, że stany modelu są określone przez skupienia obserwacji. Skupienia te można uzyskać drogą grupowania sekwencji obserwacji sygnału mowy.

SŁOWA KLUCZOWE: HMM, ukryte modele Markowa, estymacja, rozpoznawanie mowy

1. Wprowadzenie

Podstawę rozpoznawania mowy stanowi sygnał próbkowany ze stałą częstotliwością F_s. Przyjmuje się założenie, że właściwości (cechy) sygnału mowy nie zmieniają się w krótkim okresie czasu, w tzw. ramkach czasowych. Ramki mają jednakową szerokość i mogą zachodzić na siebie. Cechy wyznaczone na podstawie ramki o numerze t stanowią obserwację – oznaczaną jako $\mathbf{o}_t$. Cechy wyznaczone w kolejnych, dyskretnych momentach czasu $t = 1, ..., T$ wyznaczają sekwencję obserwacji $O = (\mathbf{o}_1, ..., \mathbf{o}_T)$. Numery ramek czasowych tworzą zbiór chwil czasowych modelu. Obserwacje $\mathbf{o}_t$ są losowe, a ich rozkład prawdopodobieństwa może być różny dla różnych chwil czasowych. Przyjmuje się, że rozkład prawdopodobieństwa obserwacji $\mathbf{o}_t$ jest zależny od stanu q_t procesu Markowa. Stanów procesu Markowa nie można bezpośrednio widzieć (są one ukryte). Istnienie każdego z nich objawia się tylko tym, że generowane ciągi obserwacji są zgodne z różnymi rozkładami prawdopodobieństwa.

Podstawę proponowanego rozwiązania stanowi spostrzeżenie, że stan modelu można kojarzyć ze skupieniami obserwacji w przestrzeni cech. Wyodrębnienie skupień może stanowić punkt wyjścia do określenia rozkładów prawdopodobieństwa, zgodnie z którymi zostały wygenerowane obserwacje.

Osobliwość modeli Markowa wykorzystywanych w zadaniach rozpoznawania mowy polega na założeniu, że trajektoria (będąca sekwencją stanów modelu) nie powraca do stanów, w których model przebywał w poprzednich chwilach czasowych (tzn. może pozostać w aktualnym stanie albo przejść do dowolnego ze stanów następnych). Modele HMM tego typu są określane jako „*lewy – prawy*" (zgodnie z diagramem przejść łańcucha).

2. Ukryte modele Markowa

Przedstawione zostaną podstawowe zagadnienia dotyczące ukrytych modeli Markowa (ang. *Hidden Markov Models*, *HMM*) szczególnego rodzaju: umożliwiające ich łączenie. Metodę łączenia modeli HMM zaproponował Young [9]. Uzupełnił on model HMM o dwa stany: nieemisyjny stan wejściowy – pierwszy stan modelu oraz nieemisyjny stan wyjściowy – ostatni stan modelu o numerze N. Pozostałe stany przyjął jako stany emisyjne i założył, że tylko one mogą generować obserwacje. Takie postępowania spowodowało, że zbiory chwil czasowych są różne dla obserwacji i procesu zmiany stanów. Sekwencję obserwacji, poczynionych w dyskretnych momentach czasu od $t = 1$ do czasu $t = T$, oznaczamy następująco: $O = (\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_t, ..., \mathbf{o}_T)$. Chwile czasu t dla obserwacji tworzą zbiór chwil czasowych: $t \in \{1, ..., T\}$, a dla procesu zmiany stanów zbiór: $\{0, 1, ..., T, T + 1\}$. Cecha ta odróżnia model Younga od zwykłego modelu HMM.

Proces zmiany stanów modelu HMM opisuje się jednorodnym łańcuchem Markowa. Niech zmienna q_t oznacza stan modelu, w którym znajduje się model, w dyskretnym momencie czasu $t \geq 0$. System jest „gotowy do pracy", gdy model znajduje się w nieemisyjnym stanie wejściowym, dlatego zawsze $q_0 = 1$. W dyskretnej chwili czasu $T + 1$ model zawsze znajdzie się w nieemisyjnym stanie wyjściowym, czyli $q_{T+1} = N$. Dla $t \in \langle 1, T \rangle$ model może przebywać w jednym z $N - 2$ różnych stanów emisyjnych, czyli $q_t \in \{2, ..., N - 1\}$. Przyjmuje się, że sekwencja stanów $Q = (q_0, q_1, ..., q_T, q_{T+1})$ nie jest znana. Prawdopodobieństwa przejść pomiędzy kolejnymi stanami opisuje się macierzą przejść $\mathbf{A} = [\, a_{ij} \,]$, gdzie:

$$a_{ij} = P\left(q_{t+1} = j \middle| q_t = i\right), \quad i \in \{1, ..., N\}, \quad j \in \{1, ..., N\}.$$

Początkowy rozkład prawdopodobieństwa stanów jest punktowy: $q_0 = 1$. Dla procesu obserwacji istotny jest rozkład w chwili $t = 1$. Określony

jest on przez prawdopodobieństwo: a_{1j}, gdzie $j \in \{2, ..., N-1\}$ (przejścia z nieemisyjnego stanu wejściowego do dowolnego stanu emisyjnego modelu). Wartość a_{1N} równa jest prawdopodobieństwu pominięcia modelu (jest to prawdopodobieństwo przejścia z nieemisyjnego stanu wejściowego do nieemisyjnego stanu wyjściowego). Prawdopodobieństwo pominięcia końcowych stanów emisyjnych a_{iN}, gdzie $i \in \{2, ..., N-1\}$, może być niezerowe.

Jeżeli **A** jest macierzą trójkątną górną ($a_{ij} = 0$ dla $i > j$), to model nazywa się „*lewy – prawy*". Jest on zwykle stosowany do modelowania jednostek akustycznych [9]. W dostępnej literaturze brak jest informacji na temat stosowania innych typów ukrytych modeli Markowa w zadaniach rozpoznawania mowy.

Zgodnie z przyjętymi powyżej założeniami prawdopodobieństwa przejść spełniają następujące ograniczenia:

1) $0 \le a_{ij} \le 1$, $i \in \{1, ..., N\}$, $j \in \{1, ..., N\}$,
2) $a_{ij} = 0$, jeśli $i > j$,
3) $a_{11} = 0$,
4) $a_{NN} = 1$,
5) $\sum_{j=1}^{N} a_{ij} = 1$, $i \in \{1, ..., N\}$.

Sposób generowania obserwacji zależy od aktualnego stanu. Jest on charakteryzowany przez rodzinę rozkładów prawdopodobieństw **B**. W ogólnym przypadku jest to wektor, którego współrzędne opisują rozkłady prawdopodobieństwa generowania obserwacji przez stany emisyjne modelu HMM:

$$\mathbf{B} = \left[b_2(\mathbf{o}) \quad ... \quad b_{N-1}(\mathbf{o}) \right], \tag{1}$$

gdzie:

$b_j(\mathbf{o})$ – funkcja określająca rozkład prawdopodobieństwa wygenerowania obserwacji **o** przez stan j, gdzie $j \in \{2, ..., N-1\}$,
N – liczba stanów modelu HMM.

Wektor **B** w dalszej części będziemy nazywać wektorem rozkładów obserwacji. Funkcja $b_j(\mathbf{o})$ w zależności od typu rozkładu może być funkcją prawdopodobieństwa lub funkcją gęstości prawdopodobieństwa. W ogólnym przypadku przyjmuje się multimodalne funkcje gęstości, w postaci tzw. kompozycji funkcji gęstości:

$$b_i(\mathbf{o}) = \sum_{m=1}^{M_i} c_{im} \frac{1}{\sqrt{(2\Pi)^D |\mathbf{V}_{im}|}} \exp\left(-\frac{1}{2} (\mathbf{o} - \mathbf{E}_{im}) \mathbf{V}_{im}^{-1} (\mathbf{o} - \mathbf{E}_{im})^{\mathrm{T}} \right), \tag{2}$$

gdzie:

- **o** – wektor wierszowy o D elementach,
- c_{im} – waga składnika kompozycji m stanu i,
- M_i – liczba składników kompozycji stanu i,
- $\mathbf{E}_{im}$ – wartości oczekiwana składnika kompozycji o numerze m dla stanu i,
- $\mathbf{V}_{im}$ – macierz kowariancji składnika kompozycji o numerze m dla stanu i.

Wagi c_{im} spełniają następujący warunek:

$$\sum_{m=1}^{M_i} c_{im} = 1, \quad \text{dla } c_{im} \geq 0 . \tag{3}$$

Rozpatrywany ukryty model Markowa zapisywać będziemy następująco:

$$\lambda = (\mathbf{A}, \mathbf{B}), \tag{4}$$

gdzie:

- **A** – macierz przejść,
- **B** – wektor rozkładów obserwacji.

Z wykorzystaniem modeli HMM związane są następujące problemy:

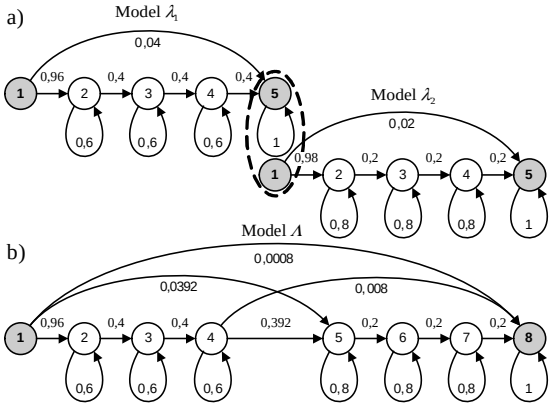
1) **Problem segmentacji** – polega na wyznaczeniu trajektorii, dla której prawdopodobieństwo wystąpienia wraz z zarejestrowaną sekwencją obserwacji jest największe.
2) **Problem ewaluacji HMM** – polega na ocenie możliwości wygenerowania przez model zarejestrowanej sekwencji obserwacji, czyli obliczeniu prawdopodobieństwa wygenerowania przez model zarejestrowanej sekwencji obserwacji.
3) **Problem estymacji parametrów HMM** – polega na zbudowaniu modelu na podstawie zarejestrowanych wcześniej wzorcowych sekwencji obserwacji.

2.1. Konkatenacja modeli HMM

Metodę łączenia, albo inaczej mówiąc konkatenacji, modeli HMM przedstawił Young w [9]. Jak już zaznaczono, Young do modelu HMM dodał dwa stany: nieemisyjny stan wejściowy oraz nieemisyjny stan wyjściowy. Stany te są wykorzystywane do łączenia modeli. Ideę łączenia przedstawiono na rys. 1. Linią przerywaną na rys. 1a oznaczono łączenie modeli. Stany nieemisyjne oznaczono szarym kolorem. W wyniku połączenia otrzymano model, który przedstawiono na rys. 1b. Połączenie modeli jest równoważne z połączeniem nieemisyjnego stanu wyjściowego modelu λ_1 z nieemisyjnym stanem wejściowym modelu λ_2.

Ze względu na mniejszą złożoność obliczeniową, fakt połączenia modeli

uwzględnia się tylko w algorytmach ewaluacji i estymacji parametrów HMM. Stosowne algorytmy można znaleźć w [7] i [9].



Rys. 1. a) sposób łączenia (konkatenacji) modeli HMM
b) wynik konkatenacji modeli HMM

3. Estymacja parametrów modelu wypowiedzi

Estymacji parametrów modeli dokonuje się na podstawie segmentacji zbioru sekwencji uczących. Przyjęto założenie, że segmenty są jednocześnie skupieniami obserwacji w przestrzeni cech. Segmentację wykonuje się wykorzystując wyniki grupowania sekwencji czasowych [5]. Każdą sekwencję uczącą dzielimy na tyle grup z ilu stanów emisyjnych składa się model – każda grupa reprezentuje stan emisyjny modelu wypowiedzi. Najważniejsze własności algorytmu, to:

- estymacja parametrów modeli HMM bez wstępnej segmentacji,
- jednoczesna estymacja parametrów wszystkich modeli HMM.

Jednak najważniejszą własnością algorytmu estymacji parametrów modelu wypowiedzi jest możliwość jednoczesnej estymacji parametrów modeli HMM, które wchodzą w skład modelu wypowiedzi. Uzyskana segmentacja określa przyporządkowanie obserwacji stanom modeli jednostek akustycznych, co umożliwia później estymację parametrów modeli HMM.

Oznaczmy zbiór sekwencji uczących jako $\{O^{r}\}$, $r = 1, ..., R$, gdzie R – liczba sekwencji uczących reprezentujących wypowiedź W. Niech konkatenacja modeli HMM $\Lambda = \lambda_1 \,\&\, ... \,\&\, \lambda_k \,\&\, ... \,\&\, \lambda_K$ oznacza model

wypowiedzi W, gdzie λ_k jest modelem jednostki akustycznej. Estymacji parametrów modeli λ_k dokonuje się na podstawie zbioru sekwencji uczących $\{O^r\}$, $r = 1, ..., R$, gdzie R – liczba sekwencji uczących, które reprezentują wypowiedź W. Każda sekwencja obserwacji O^r składa się z T^r obserwacji (punktów). Każdy model składa się z dwóch stanów nieemisyjnych (wejściowego i wyjściowego) oraz z $N^{(k)} - 2$ stanów emisyjnych. Przez N oznaczmy liczbę stanów emisyjnych modelu wypowiedzi, czyli:

$$N = \sum_{k=1}^{K} N^{(k)} - 2, \tag{5}$$

gdzie:

N – liczba stanów emisyjnych konkatenacji Λ,

$N^{(k)}$ – liczba wszystkich stanów modelu λ_k,

K – liczba modeli.

W modelu HMM przyjmuje się, że tylko stany emisyjne generują obserwacje. Stany te utożsamia się ze skupieniami punktów w przestrzeni cech wygenerowanymi zgodnie z pewnym rozkładem prawdopodobieństwa. Estymacji parametrów modeli dokonuje się na podstawie segmentacji zbioru sekwencji uczących. Segmentację wykonuje się wykorzystując wyniki grupowania sekwencji czasowych. Według przyjętej metody grupowania wyznaczamy drzewo grupowania, a następnie dzielimy sekwencję na N grup. Każda grupa $\mathbf{G}_i^{(k)}$ reprezentuje stan emisyjny modelu wypowiedzi, gdzie: $i = 2, ..., N^{(k)} - 1$; $k = 1, ..., K$; K – liczba modeli konkatenacji Λ. Przed przystąpieniem do opisu algorytmu przyjmijmy następujące oznaczenia:

$\boldsymbol{\xi}^{(k)}(i, j)$ – liczba przejść pomiędzy stanem (i, k) a stanem (j, k), gdzie: $i = 1, ..., N^{(k)}$, $j = 1, ..., N^{(k)}$;

$\Xi^{(k)}(i)$ – liczba stanów (i, k) wygenerowanych przez model wypowiedzi Λ, gdzie: $i = 1, ..., N^{(k)}$.

Kolejne kroki algorytmu estymacji parametrów są następujące.

Struktury danych

Stałe:

$\{O^r\}$ – zbiór sekwencji uczących, $r = 1, ..., R$,

T^r – długość sekwencji obserwacji,

R – liczba sekwencji uczących,

Zmienne:

k – numer modelu konkatenacji Λ,

$N^{(k)}$ – liczba stanów modelu λ_k,

$M_i^{(k)}$ – liczba składowych kompozycji stanu (i, k), gdzie $i = 2, ..., N^{(k)} - 1$,

$\mathbf{G}_n^{(k)}$ – grupa reprezentująca stan emisyjny modelu λ_k, gdzie: $n = 2, ..., N^{(k)} - 1$,

$\mathbf{G}_{im}^{(k)}$ – grupa reprezentująca kompozycję m stanu emisyjnego (i, k), gdzie: $i = 2, ..., N^{(k)} - 1$; $m = 1, ..., M_i^{(k)}$,

$\boldsymbol{\xi}^{(k)}(i, j)$ – liczba przejść pomiędzy stanem (i, k) oraz (j, k), gdzie: $i = 1, ..., N^{(k)}$, $j = 1, ..., N^{(k)}$,

$\xi_r^{(k)}(i, j)$ – liczba przejść pomiędzy stanem (i, k) oraz (j, k), które towarzyszyły sekwencji obserwacji O^r, gdzie: $i = 1, ..., N^{(k)}$, $j = 1, ..., N^{(k)}$, $r = 1, ..., R$,

$\Xi^{(k)}(i)$ – liczba stanów (i, k) wygenerowanych przez model wypowiedzi Λ, gdzie: $i = 1, ..., N^{(k)}$,

$\Xi_r^{(k)}(i)$ – liczba stanów (i, k) modelu wypowiedzi Λ, które towarzyszyły sekwencji obserwacji O^r, gdzie: $i = 1, ..., N^{(k)}$, $r = 1, ..., R$,

$c_{im}^{(k)}$ – waga składnika kompozycji m stanu (i, k), gdzie: $m = 1, ..., M_i^{(k)}$, $i = 2, ..., N^{(k)} - 1$, $k = 1, ..., K$,

$\mathbf{E}_{im}^{(k)}$ – wartość oczekiwana składnika kompozycji m stanu (i, k), gdzie: $m = 1, ..., M_i^{(k)}$, $i = 2, ..., N^{(k)} - 1$, $k = 1, ..., K$,

$\mathbf{V}_{im}^{(k)}$ – macierz kowariancji składnika kompozycji m stanu (i, k), gdzie: $m = 1, ..., M_i^{(k)}$, $i = 2, ..., N^{(k)} - 1$, $k = 1, ..., K$.

Obliczenia wstępne

Dla każdej sekwencji obserwacji O^r, gdzie: $r = 1, 2, ..., R$, należy wykonać przedstawione poniżej czynności.

Według wybranego algorytmu grupowania sekwencji czasowych [5] należy wyznaczyć drzewo grupowania sekwencji O^r, opisane macierzą **Z**. Następnie należy dokonać podziału na N grup, wykorzystując do tego celu algorytm podziału na grupy [5]. W przypadku algorytmu grupowania z nakładaniem, punkty mogą zostać przypisane do więcej niż jednej grupy. Wtedy, wykorzystując metodę największej wiarygodności, należy przyporządkować punkty tylko do jednej grupy tak, aby grupy ułożyły się w niepowracający ciąg.

Należy odnaleźć grupę, w której znajduje się punkt o indeksie równym 1. Wszystkie jej punkty należy przyporządkować do grupy $\mathbf{G}_2^{(1)}$, zapamiętując największy indeks t punktu przyporządkowanego do tego

stanu. Następnie odnaleźć grupę z punktem o indeksie $t+1$ i przyporządkować jej elementy grupie $\mathbf{G}_3^{(1)}$. Postępuje się tak, aż do przyporządkowania elementów grupie $\mathbf{G}_{N^{(1)}-1}^{(1)}$. Następne przypisanie rozpoczyna się od grupy reprezentującej pierwszy stan emisyjny modelu numer dwa, czyli od $\mathbf{G}_2^{(2)}$, itd. Przypisanie elementów grupom, reprezentującym stany emisyjne, kontynuuje się, aż do ich wyczerpania.

Dla każdej sekwencji O^r ($r = 1, ..., R$) określić liczbę przejść pomiędzy stanami modelu, czyli macierz $\xi_r^{(k)}$, a następnie zsumować ją z macierzą $\xi^{(k)}$, czyli:

$$\xi^{(k)}(i,j) = \xi^{(k)}(i,j) + \xi_r^{(k)}(i,j), \tag{6}$$

gdzie:

$i = 1, ..., N^{(k)}$,
$j = 1, ..., N^{(k)}$,
$k = 1, ..., K$.

Wyznaczyć liczbę przyporządkowań obserwacji poszczególnym stanom emisyjnym modelu wypowiedzi, tzn. wyznaczyć macierz $\Xi_r^{(k)}$ i zsumować ją z macierzą $\Xi^{(k)}$:

$$\Xi^{(k)}(i) = \Xi^{(k)}(i) + \Xi_r^{(k)}(i), \tag{7}$$

gdzie:

$i = 2, ..., N^{(k)} - 1$,
$k = 1, ..., K$.

Dla wszystkich modeli należy uwzględnić, że działanie systemu rozpoczyna się w nieemisyjnym stanie wejściowym, a kończy w nieemisyjnym stanie wyjściowym. Jest to powodem zwiększenia odpowiednich wartości elementów macierzy o jeden:

$$\xi^{(k)}(1,j) = \xi^{(k)}(1,j) + 1, \text{ jeśli } o_1 \in \mathbf{G}_j^{(k)},$$
$$\xi^{(k)}(i,N^{(k)}) = \xi^{(k)}(i,N^{(k)}) + 1, \text{ jeśli } o_{T^r} \in \mathbf{G}_i^{(k)},$$
$$\Xi^{(k)}(1) = \Xi^{(k)}(1) + 1,$$
$$\Xi^{(k)}(N^{(k)}) = \Xi^{(k)}(N^{(k)}) + 1.$$

Aby wyznaczyć parametry kompozycji funkcji gęstości dodatkowo dokonuje się podziału grup reprezentujących stany emisyjne modelu. Elementy grupy $\mathbf{G}_i^{(k)}$ należy podzielić na $M_i^{(k)}$ podgrup, wykorzystując do tego celu algorytm rozłącznego grupowania punktów metodą hierarchiczną. W wyniku tego otrzymujemy grupy: $\mathbf{G}_{im}^{(k)}$, gdzie:

$i = 2, ..., N^{(k)} - 1;\ m = 1,...,M_i^{(k)},\ k = 1,...,K$.

Etap wstępny kończy się w momencie pogrupowania wszystkich sekwencji uczących

Etap estymacji

Estymacji parametrów modeli wchodzących w skład konkatenacji Λ dokonuje się następująco. Macierz przejść $\mathbf{A}^{(k)}$ wyznaczamy dla każdego modelu λ_k według wzoru:

$$\mathbf{A}^{(k)}(i,j) = \frac{\boldsymbol{\xi}^{(k)}(i,j)}{\boldsymbol{\Xi}^{(k)}(j)}. \tag{8}$$

gdzie: $i = 1,...,N^{(k)},\ j = 1,...,N^{(k)}, k = 1,...,K$.

Dla każdego składnika kompozycji m stanu emisyjnego (i, k) należy wyznaczyć:

– współczynnik kompozycji:

$$c_{im} = \frac{\left\|\mathbf{G}_{im}^{(k)}\right\|}{\left\|\mathbf{G}_{i}^{(k)}\right\|} \tag{9}$$

gdzie: $i = 2,...,N^{(k)} - 1,\ m = 1,...,M_i^{(k)},\ k = 1,...,K$.

– wartość średnią:

$$\mathbf{E}_{im}^{(k)} = \frac{\sum_{\mathbf{o}_t \in \mathbf{G}_{im}^{(k)}} \mathbf{o}_t^{(r)}}{\left\|\mathbf{G}_{im}^{(k)}\right\|} \tag{10}$$

gdzie: $i = 2,...,N^{(k)} - 1,\ m = 1,...,M_i^{(k)},\ k = 1,...,K$.

– macierz kowariancji:

$$\mathbf{V}_{im}^{(k)} = \frac{\sum_{\mathbf{o}_t \in \mathbf{G}_{im}^{(k)}} \left(\mathbf{o}_t^{(r)} - \mathbf{E}_{im}^{(k)}\right)\left(\mathbf{o}_t^{(r)} - \mathbf{E}_{im}^{(k)}\right)'}{\left\|\mathbf{G}_{im}^{(k)}\right\|} \tag{11}$$

gdzie: $i = 2,...,N^{(k)} - 1,\ m = 1,...,M_i^{(k)},\ k = 1,...,K$.

Etap końcowy

Obliczenia kończymy po wyznaczeniu parametrów wszystkich modeli HMM.

4. Wyniki badań

Wszystkie doświadczenia wykonano przy pomocy laboratoryjnego systemu rozpoznawania mowy, który zbudował autor artykułu. Obliczenia przeprowadzono na komputerze z procesorem AMD Athlon® XP 1700+ z pamięcią RAM 1024 MB.

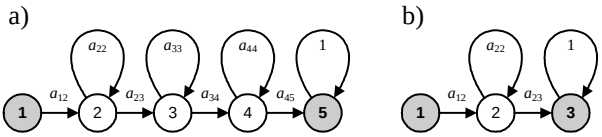
Tab. 1. Słownik systemu

z	Słownik systemu			Liczba sekwencji	
	element słownika W_z	transkrypcja fonetyczna	liczba stanów emisyjnych	uczących	testowych
0	zero	qzeroq	14	90	20
1	jeden	qjedenq	17	95	20
2	dwa	qdwaq	11	94	20
3	Trzy	qtsyq	11	95	20
4	cztery	qcteryq	17	94	19
5	pięć	qpęćq	11	95	20
6	sześć	qseśćq	14	94	19
7	siedem	qśedemq	17	95	20
8	osiem	qośemq	14	94	20
9	dziewięć	qźewęćq	17	95	20

Jako zbiór uczący oraz zbiór testowy użyto plików dźwiękowych z zasobu mowy „Robot". Pliki zawierają szesnastobitowe próbki sygnału mowy, pozyskane z częstotliwością próbkowania $F_s = 22{,}05$ kHz i zakodowane w formacie Microsoft PCM. Sekwencje obserwacji pozyskano stosując ramki czasowe o szerokości 15 ms. Ramki nakładały się na siebie w połowie swojej szerokości. Z każdej ramki wyznaczono 10 współczynników LPC oraz 15 współczynników MFCC [7].

W tab. 1 przedstawiono słownik systemu oraz liczbę sekwencji uczących i testowych. Jako jednostkę akustyczną przyjęto głoskę. Na początku i na końcu każdej wypowiedzi dodano „głoskę q", który oznacza ciszę. Modele jednostek akustycznych to modele „*lewy – prawy*" (rys. 2). Składają się one z trzech stanów emisyjnych oraz dwóch stanów nieemisyjnych: wejściowego i wyjściowego, które na rys. 2a oznaczono szarym tłem. Wyjątkiem jest model „ciszy" (rys. 2b), który składa się z dwóch stanów nieemisyjnych i jednego stanu emisyjnego. Jako zbioru uczącego użyto 941 sekwencji obserwacji. Zbiór

testowy składał się z 198 sekwencji obserwacji.



Rys. 2. Diagram dozwolonych przejść dla: a) modelu głoski, b) modelu „ciszy"

W celu dokonania oceny możliwości estymacji parametrów modeli wypowiedzi przy wykorzystaniu opracowanych metod grupowania sekwencji czasowych przeprowadzono następujący eksperyment. Jako metodę odniesienia przyjęto metodę Bauma – Welcha. Wyznaczono parametry modeli wypowiedzi i przeprowadzono rozpoznawanie sekwencji testowych. Wyniki doświadczenia przedstawiono w rozdz. 4.1. Na ich podstawie dokonano oceny tych możliwości Dodatkowo zbadano wpływ sposobu łączenia w grupy na skuteczność rozpoznawania wypowiedzi. Opis doświadczenia oraz jego wyniki przedstawiono w rozdz. 4.2.

Ocena możliwości wykorzystania do estymacji parametrów modeli wypowiedzi metod grupowania sekwencji czasowych jest pośrednia – na podstawie oceny skuteczności systemu rozpoznawania mowy. Wykorzystane wskaźniki oceny systemu rozpoznawania mowy opracowano na podstawie [6], [8]. Przy omawianiu wyników eksperymentu przyjęto następujące oznaczenia:

- R_z – liczba sekwencji testowych wypowiedzi W_z,
- $\tilde{R}_z$ – liczbę sekwencji obserwacji rozpoznanych jako wypowiedź W_z,
- $\hat{R}_z$ – liczba poprawnie rozpoznanych sekwencji obserwacji wypowiedzi W_z,
- γ_z – stopa błędnego przetworzenia obserwacji pochodzących z wypowiedzi W_z,
- $\hat{\gamma}_z$ – poziom nieufności rozpoznania wypowiedzi W_z,
- ϑ_z – skuteczność rozpoznawania wypowiedzi W_z.

4.1. Reestymacja parametrów modeli wypowiedzi metodą Bauma – Welcha

Eksperyment polegał na zbadaniu możliwości reestymacji modeli jednostek akustycznych przy wykorzystaniu algorytmu Bauma – Welcha.

Dokonano jednokrotnego pomiaru czasu – dla przedstawionego powyżej zbioru uczącego i konfiguracji komputera Początkowe parametry modeli jednostek akustycznych wyznaczono na podstawie równomiernego podziału sekwencji zbioru uczącego na segmenty. Parametry początkowe wyznaczono w czasie równym 13,5 s. Następnie wykonano pięć iteracji algorytmu Bauma – Welcha. Jedna iteracja algorytmu trwała około 18 minut. Po dokonaniu reestymacji parametrów modeli jednostek akustycznych przeprowadzono rozpoznawanie sekwencji ze zbioru sekwencji testowych. Wyniki eksperymentu przedstawiono w tab. 2.

Tab. 2. Wyniki rozpoznawania po reestymacji metodą Bauma – Welcha

Z	R_z	$\tilde{R}_z$	$\hat{R}_z$	γ_z	$\hat{\gamma}_z$	ϑ_z
0	20	20	21	0	0,0476	100,00%
1	20	20	20	0	0	100,00%
2	20	19	19	0,0500	0	95,00%
3	20	20	20	0	0	100,00%
4	19	19	20	0	0,0500	100,00%
5	20	19	19	0,0500	0	95,00%
6	19	19	19	0	0	100,00%
7	20	19	19	0,0500	0	95,00%
8	20	20	20	0	0	100,00%
9	20	20	21	0	0,0476	100,00%
Ogólnie	198	195	198	0,0150	0,0145	98,50%
	SUMA			$\overline{\gamma}$	$\hat{\overline{\gamma}}$	$\overline{\vartheta}$

4.2. Estymacja parametrów modeli wypowiedzi na podstawie grupowania sekwencji czasowych

Doświadczalnie zbadano zastosowanie algorytmów grupowania sekwencji punktów, hierarchicznego grupowania rozłącznego oraz hierarchicznego grupowania z nakładaniem, do estymacji parametrów modeli wypowiedzi. Eksperyment polegał na zbadaniu algorytmu estymacji parametrów modeli wypowiedzi. Pośrednim wskaźnikiem oceny algorytmu jest skuteczność systemu rozpoznawania mowy. Zbadano, czy wybór sposobu określania odległości

pomiędzy punktami oraz grupami ma wpływ na skuteczność systemu rozpoznawania mowy.

Tab. 3. Wyniki rozpoznawania sekwencji testowych po estymacji parametrów na podstawie grupowania rozłącznego przy określaniu odległości pomiędzy: grupami metodą Warda, pomiędzy punktami według metryki Mahalanobisa

z	R_z	$\tilde{R}_z$	$\hat{R}_z$	γ_z	$\hat{\gamma}_z$	ϑ_z
0	20	20	20	0	0	100,00%
1	20	20	20	0	0	100,00%
2	20	20	20	0	0	100,00%
3	20	15	15	0,2500	0	75,00%
4	19	19	24	0	0,2083	100,00%
5	20	19	19	0,0500	0	95,00%
6	19	19	19	0	0	100,00%
7	20	20	21	0	0,0476	100,00%
8	20	20	20	0	0	100,00%
9	20	19	20	0,0500	0,0500	95,00%
Ogólnie	198	191	198	0,0350	0,0306	96,50%
	SUMA			$\overline{\gamma}$	$\hat{\overline{\gamma}}$	$\overline{\vartheta}$

Każdą sekwencję obserwacji, ze zbioru sekwencji uczących, podzielono na grupy. Liczbę grup dobrano do modelu wypowiedzi. Jeżeli sekwencję obserwacji pozyskano z wypowiedzi W_z, to podzielono ją na liczbę grup równą liczbie stanów emisyjnych modelu wypowiedzi Λ_z. Na podstawie tak wydzielonych grup można było estymować parametry modeli HMM, wykorzystując algorytm „estymacji parametrów modelu wypowiedzi".

W pierwszej kolejności dokonano estymacji parametrów modeli wypowiedzi, a następnie dokonano rozpoznawania sekwencji ze zbioru sekwencji testowych. Estymacji parametrów wszystkich modeli jednostek akustycznych, łącznie z grupowaniem sekwencji czasowych, dokonywano w czasie około 3 minut.

Dokonano estymacji parametrów modeli wypowiedzi na podstawie wyników grupowania sekwencji uczących hierarchiczną metodą grupowania rozłącznego. Największą skuteczność rozpoznawania, równą $\overline{\vartheta} = 96\%$

uzyskano dla łączenia w grupy metodą średniej odległości i przy określaniu odległości pomiędzy punktami stosując metrykę Euklidesa. Średnia stopa błędnego przetworzenia obserwacji była równa $\overline{\gamma} = 0,04$, a średni poziom nieufności rozpoznania $\hat{\overline{\gamma}} = 0,037$. Szczegółowe wyniki przedstawiono w tab. 3.

Tab. 4. Wyniki rozpoznawania sekwencji testowych po estymacji parametrów na podstawie grupowania z nakładaniem przy określaniu odległości pomiędzy: grupami metodą średniej odległości, pomiędzy punktami według metryki Euklidesa

z	R_z	$\tilde{R}_z$	$\hat{R}_z$	γ_z	$\hat{\gamma}_z$	ϑ_z
0	20	19	19	0,0500	0	95,00%
1	20	19	21	0,0500	0,0952	95,00%
2	20	20	21	0	0,0476	100,00%
3	20	16	16	0,2000	0	80,00%
4	19	19	22	0	0,1364	100,00%
5	20	19	19	0,0500	0	95,00%
6	19	19	19	0	0	100,00%
7	20	20	22	0	0,0909	100,00%
8	20	20	20	0	0	100,00%
9	20	19	19	0,0500	0	95,00%
Ogólnie	198	190	198	0,0400	0,0370	96,00%
	SUMA			$\overline{\gamma}$	$\hat{\overline{\gamma}}$	$\overline{\vartheta}$

Zbadano również możliwość dokonania estymacji parametrów modeli wypowiedzi na podstawie wyników grupowania sekwencji uczących hierarchiczną metodą grupowania z nakładaniem. W tym przypadku największą skuteczność rozpoznawania, równą $\overline{\vartheta} = 96\%$ uzyskano dla łączenia w grupy metodą średniej odległości i przy określaniu odległości pomiędzy punktami stosując metrykę Euklidesa. Średnia stopa błędnego przetworzenia obserwacji była równa $\overline{\gamma} = 0,04$, a średni poziom nieufności rozpoznania $\hat{\overline{\gamma}} = 0,037$. Wyniki doświadczenia przedstawiono w tab. 4.

5. Podsumowanie

Głównym celem przedstawionego w artykule eksperymentu była ocena możliwości wykorzystania metod grupowania sekwencji czasowych do estymacji parametrów modeli HMM. Jako punkt odniesienia przyjęto metodę Bauma – Welcha. W świetle przedstawionych wyników można sformułować następujące wnioski:

1) zastosowanie metody grupowania sekwencji czasowych pozwala dokonać estymacji parametrów modeli HMM,
2) jakość rozpoznawania mowy przy zastosowaniu modeli uzyskanych na podstawie grupowania obserwacji jest tego samego rzędu, co przy wykorzystaniu metody Bauma – Welcha,
3) czas estymacji jest kilkanaście razy mniejszy niż analogiczny czas obliczeń opartych na algorytmie Bauma – Welcha.

Literatura

[1] Everitt B., Landau S. Leese M.: *Cluster Analysis, 4'th edition*, Edward Arnold Publishers Ltd., London 2001.

[2] Koronacki J., Ćwik J.: *Statystyczne systemy uczące się*, Wydawnictwa Naukowo – Techniczne, Warszawa 2005.

[3] Kwiatkowski W.: *Metody automatycznego rozpoznawania wzorców*, Instytut Automatyki i Robotyki WAT, Warszawa 2001,

[4] Mathworks, Inc: *Statistics Toolbox User's Guide*, http://www.mathworks.com/access/helpdesk/help/pdf_doc/stats/stats.pdf, MathWorks, 2005.

[5] Pałys T.: *Algorytmy grupowania sekwencji czasowych*, Biuletyn Instytutu Automatyki i Robotyki WAT, 23/2006. Warszawa, 2006.

[6] Pałys T.: *Zastosowanie metody grupowania sekwencji czasowych w rozpoznawaniu mowy na podstawie ukrytych modeli Markowa*, Rozprawa doktorska, WAT. Warszawa, 2006.

[7] Wiśniewski A. M.: *Automatyczne rozpoznawanie mowy bazujące na ukrytych modelach Markowa – problemy i metody*, Biuletyn Instytutu Automatyki i Robotyki WAT, 12/2000, ss. 3-83. Warszawa, 2000.

[8] Wiśniewski A. M.: *Metody oceny systemów rozpoznawania mówców*, Biuletyn Instytutu Automatyki i Robotyki WAT, 13/2000, ss. 3-35, Warszawa, 2000.

[9] Young S. J., Evermann G., Hain, T., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P.: *The HTK Book*, Microsoft Corporation, 2000.

Application of Time Sequences Clustering Methods in the Speech Recognition Based on Hidden Markov Models

ABSTRACT: A problem of hidden Markov models formation on the basis of recorded speech is considered in this paper. The key issue is the designation of a Markov model set. The assumption is that each HMM state is associated with clusters of observations. The clusters may be obtained by gathering of observations sequences for a speech signal.

KEYWORDS: HMM, Hidden Markov Model, estimation, speech recognition

Recenzent: prof. dr hab. inż. Włodzimierz KWIATKOWSKI

Praca wpłynęła do redakcji: 28.12.2006