Listy Biometryczne
Nr 42-45 (1974)

Mirosław K r z y ś k o (Poznań)

ZASTOSOWANIE LINIOWEJ METODY IDENTYFIKACJI W PROBLEMACH ROZPOZNAWANIA GŁOSU

1. Od pewnego czasu poszukuje się metod obiektywnie identyfikujących nagrany głos. Badania te pozostają w ścisłym związku z możliwością dopuszczania przez sąd dowodu rzeczowego w postaci nagrania głosu.

Problematyka rozpoznawania głosu obiektywnymi metodami, pozostając ważnym zagadnieniem w kryminalistyce, zaczyna również przenikać do innych dziedzin wiedzy, zwłaszcza do nauk technicznych (między innymi bada się problem stworzenia łączności człowiek-maszyna przy pomocy mowy).

Właściwości indywidualne głosu związane są z osobniczymi cechami fizjologicznymi i psychicznymi. Wydaje się rzeczą oczywistą, że najbardziej trwałe i najmniej podatne na świadome lub nieświadome zafałszowania są właściwości wynikające z uwarunkowań fizjologicznych. Parametrami ściśle zależnymi od budowy fizjologicznej toru głosowego są formanty samogłoskowe. Wielu akustyków uważa, że w parametrach formantowych odnaleźć można cechy osobnicze głosu. Do pomiarowego uchwycenia najlepiej nadają się cztery pierwsze formanty F_1, F_2, F_3, F_4. Sposób dokonywania pomiaru częstotliwości formantowych opisano w pracy [3].

Identyfikacja osób na podstawie częstotliwości formantowych jest możliwa przy zastosowaniu wielowymiarowej analizy statystycznej. W przedstawionej pracy pokazany jest praktyczny sposób identyfikacji głosu ludzkiego przy pomocy liniowej minimaksowej funkcji dyskryminacyjnej.

2. Załóżmy, że danych jest m populacji generalnych $\Pi_1, \ldots, \Pi_m$. W populacji Π_i obserwowane są wartości zmiennej losowej $\underline{X}$ mającej p-wymiarowy rozkład normalny z funkcją gęstości prawdopodobieństwa postaci

$$(2.1) \qquad f(\underline{x}; \underline{\mu}_i, \underline{\Sigma}_i) = (2\pi)^{-\frac{p}{2}} |\underline{\Sigma}_i|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_i)'\underline{\Sigma}_i^{-1}(\underline{x}-\underline{\mu}_i)\right] \quad (i = 1, \ldots, m).$$

Niech $\underline{x}_0$ będzie zaobserwowaną wartością zmiennej losowej $\underline{X}$. Problem identyfikacji polega na wskazaniu, w której z populacji $\Pi_1, \ldots, \Pi_m$ zaobserwowana została wartość $\underline{x}_0$ zmiennej losowej $\underline{X}$.

W artykule tym przedstawiamy rozwiązanie tego problemu opierając się na liniowej funkcji dyskryminacyjnej wprowadzonej przez Andersona i Bahadura w pracy [1].

3. Rozważymy teraz liniową metodę klasyfikacji obserwacji $\underline{x}_0$ do jednej z dwóch populacji Π_1 lub Π_2. Ogólnie metoda ta polega na podziale przestrzeni próby R na dwa nieprzecinające się obszary R_1 i R_2 oraz przyjęciu reguły klasyfikującej obserwację $\underline{x}_0$ do i-tej populacji, gdy $\underline{x}_0 \in R_i$ (i = 1, 2). Zdefiniujemy teraz formalnie metodę liniową.

Niech $\underline{b} \neq \underline{0}$ będzie p-wymiarowym wektorem, zaś c skalarem. p-wymiarową przestrzeń próby R dzielimy (p - 1)-wymiarową hiperpłaszczyzną postaci

$$(3.1) \qquad \underline{b}'\underline{x} + c = 0$$

na dwa obszary identyfikacji R_1 i R_2. Obszar R_1 odpowiada populacji Π_1, a obszar R_2 - populacji Π_2. Hiperpłaszczyzna graniczna (3.1) w rozważanej metodzie identyfikacji spełnia następującą rolę:

$$(3.2) \qquad \begin{aligned} &\text{jeżeli } \underline{b}'\underline{x}_0 + c \leqslant 0, \text{ to decydujemy, że } \underline{x}_0 \in \Pi_1, \\ &\text{jeżeli } \underline{b}'\underline{x}_0 + c > 0, \text{ to decydujemy, że } \underline{x}_0 \in \Pi_2. \end{aligned}$$

Powstaje pytanie, jak znaleźć optymalne wartości $\underline{b}$ i c ?

Gdy $\underline{x} \in \Pi_i$, to $\underline{b}'\underline{x}$ ma jednowymiarowy rozkład normalny ze średnią $E[\underline{b}'\underline{x}] = \underline{b}'\underline{\mu}_i$ i wariancją

$$E\left[\underline{b}'\underline{x} - \underline{b}'\underline{\mu}_i\right]^2 = E\left[\underline{b}'(\underline{x}-\underline{\mu}_i)(\underline{x}-\underline{\mu}_i)'\underline{b}\right] = \underline{b}'\underline{\Sigma}_i\underline{b} \quad (i = 1, 2).$$

Niech $P(\Pi_2|\Pi_1)$ będzie prawdopodobieństwem błędnego zaklasyfikowania obserwacji $\underline{x}_0$ do populacji Π_2, gdy w rzeczywistości $\underline{x}_0$ na-

leży do populacji Π_1, a $P(\Pi_1|\Pi_2)$ - prawdopodobieństwem błędnego zaklasyfikowania obserwacji $\underline{x}_0$ do populacji Π_1, gdy w rzeczywistości $\underline{x}_0$ należy do populacji Π_2. Mamy

$$P(\Pi_2|\Pi_1) = P(\underline{b}'\underline{x} + c > 0) =$$

$$= P\left[\frac{\underline{b}'\underline{x} - \underline{b}'\underline{\mu}_1}{(\underline{b}'\underline{\Sigma}_1\,\underline{b})^{1/2}} > \frac{-c - \underline{b}'\underline{\mu}_1}{(\underline{b}'\underline{\Sigma}_1\,\underline{b})^{1/2}}\right] = \tag{3.3}$$

$$= 1 - \Phi\left[-\frac{c + \underline{b}'\underline{\mu}_1}{(\underline{b}'\underline{\Sigma}_1\,\underline{b})^{1/2}}\right] = 1 - \Phi(t_1),$$

$$P(\Pi_1|\Pi_2) = P(\underline{b}'\underline{x} + c \leqslant 0) =$$

$$= P\left[\frac{\underline{b}'\underline{x} - \underline{b}'\underline{\mu}_2}{(\underline{b}'\underline{\Sigma}_2\,\underline{b})^{1/2}} \leqslant \frac{-c - \underline{b}'\underline{\mu}_2}{(\underline{b}'\underline{\Sigma}_2\,\underline{b})^{1/2}}\right] = \tag{3.4}$$

$$= 1 - \Phi\left[\frac{c + \underline{b}'\underline{\mu}_2}{(\underline{b}'\underline{\Sigma}_2\,\underline{b})^{1/2}}\right] = 1 - \Phi(t_2);$$

Φ oznacza tu dystrybuantę jednowymiarowego standaryzowanego rozkładu normalnego.

Stawiamy sobie zadanie znalezienia wartości $\underline{b}$ i c w taki sposób, by zminimalizować wyrażenia $P(\Pi_2|\Pi_1)$ i $P(\Pi_1|\Pi_2)$ przy zachowaniu ich równości. Inaczej możemy powiedzieć, że chcemy zmaksymalizować wartość wyrażenia $t_1 = t_2$, ponieważ transformacja za pomocą dystrybuanty standaryzowanego rozkładu normalnego $\Phi(t)$ jest ściśle monotoniczna.

Jeżeli przyjmiemy, że strata związana z błędnym zaklasyfikowaniem do populacji Π_1 jest taka sama jak strata związana z błędnym zaklasyfikowaniem do populacji Π_2, to można powiedzieć, że stawiamy sobie zadanie znalezienia liniowej minimaksowej metody identyfikacji. Zadanie to rozwiązali Anderson i Bahadur w [1]. Wyrażenie na c uzyskuje się z równości $t_1 = t_2$:

$$c = -\frac{(\underline{b}'\underline{\Sigma}_1\,\underline{b})^{1/2}\,\underline{b}'\underline{\mu}_2 + (\underline{b}'\underline{\Sigma}_2\,\underline{b})^{1/2}\,\underline{b}'\underline{\mu}_1}{(\underline{b}'\underline{\Sigma}_1\,\underline{b})^{1/2} + (\underline{b}'\underline{\Sigma}_2\,\underline{b})^{1/2}}. \tag{3.5}$$

Tak uzyskane c wstawia się np. do t_2. Otrzymujemy wówczas wyrażenie

$$t_1 = t_2 = \frac{\underline{b}'(\underline{\mu}_2 - \underline{\mu}_1)}{(\underline{b}'\underline{\Sigma}_1\,\underline{b})^{1/2} + (\underline{b}'\underline{\Sigma}_2\,\underline{b})^{1/2}}. \tag{3.6}$$

Szukamy maksimum wyrażenia (3.6) względem wektora $\underline{b}$. Wektor $\underline{b}$ maksymalizujący wyrażenie (3.6) ma postać:

$$\underline{b} = \left[y \underline{\Sigma}_1 + (1 - y) \underline{\Sigma}_2 \right]^{-1} (\underline{\mu}_2 - \underline{\mu}_1), \tag{3.7}$$

gdzie $0 < y < 1$ jest jedynym rozwiązaniem równania

$$\underline{b}' \left[y^2 \underline{\Sigma}_1 - (1 - y)^2 \underline{\Sigma}_2 \right] \underline{b} = 0. \tag{3.8}$$

4. Przejdziemy teraz do rozwiązania problemu identyfikacji sformułowanego w paragrafie 2.

Opierając się na hiperpłaszczyźnie rozgraniczającej $\underline{b}'\underline{x} + c = 0$ można wprowadzić uogólnioną odległość między populacjami normalnymi o różnych macierzach kowariancji [5]. Kwadrat tej odległości wyraża się następującym wzorem:

$$(\Delta^*)^2 = (\underline{\mu}_2 - \underline{\mu}_1)' \left[y \underline{\Sigma}_1 + (1 - y) \underline{\Sigma}_2 \right]^{-1} (\underline{\mu}_2 - \underline{\mu}_1), \tag{4.1}$$

gdzie y jest jedynym rozwiązaniem równania (3.8).

Gdy $\underline{\Sigma}_1 = \underline{\Sigma}_2 = \underline{\Sigma}$, to $\Delta^* = \Delta$, gdzie

$$\Delta^2 = (\underline{\mu}_2 - \underline{\mu}_1)' \underline{\Sigma}^{-1} (\underline{\mu}_2 - \underline{\mu}_1) \tag{4.2}$$

jest kwadratem odległości Mahalanobisa.

W pierwszym kroku proponowanej metody identyfikacji obliczamy wzajemne odległości Δ^* między populacjami π_i $(i = 1, \ldots, m)$. We wzorze na Δ^* występują parametry $\underline{\mu}_i$ oraz $\underline{\Sigma}_i$ $(i = 1, \ldots, m)$, które na ogół nie są znane i zostają ocenione z prób. Niech $\underline{x}_i^{(1)}$, $\underline{x}_i^{(2)}, \ldots, \underline{x}_i^{(Ni)}$ będą wartościami zaobserwowanymi w populacji π_i; N_i oznacza tu liczebność próby $(i = 1, \ldots, m)$. Oceną parametru $\underline{\mu}_i$ jest wektor

$$\bar{\underline{x}}_i = \frac{1}{N_i} \sum_{\alpha=1}^{N_i} \underline{x}_{i\alpha}, \tag{4.3}$$

zaś oceną macierzy $\underline{\Sigma}_i$ jest macierz

$$\underline{S}_i = \frac{1}{N_i - 1} \sum_{\alpha=1}^{N_i} (\underline{x}_{i\alpha} - \bar{\underline{x}}_i)(\underline{x}_{i\alpha} - \bar{\underline{x}}_i)', \quad (i = 1, \ldots, m). \tag{4.4}$$

Niech π_r i π_s będą populacjami maksymalnie od siebie odległymi w sensie miary Δ^*. Dla tej pary populacji budujemy hiper-

płaszczyznę rozgraniczającą $\underline{b}'\underline{x} + c = 0$, gdzie $\underline{b}$ i c dane są wzorami (3.5) oraz (3.7). Po przeprowadzeniu hiperpłaszczyzny rozgraniczającej identyfikowana obserwacja $\underline{x}_0$ znajduje się w obszarze odpowiadającym populacji π_r, bądź w obszarze odpowiadającym populacji π_s. W celu określenia położenia obserwacji $\underline{x}_0$ względem tej hiperpłaszczyzny wystarczy zbadać znaki wyrażenia $\underline{b}'\underline{x} + c$ kolejno dla $\underline{x} = \underline{x}_0$, $\underline{x} = \overline{\underline{x}}_r$ i $\underline{x} = \overline{\underline{x}}_s$. Jeżeli np. znak wyrażenia $\underline{b}'\underline{x} + c$ dla $\underline{x} = \underline{x}_0$ jest zgodny ze znakiem tego wyrażenia dla $\underline{x} = \overline{\underline{x}}_r$, to obserwacja $\underline{x}_0$ znajduje się po tej samej stronie hiperpłaszczyzny rozgraniczającej co populacja π_r. W takim przypadku wykluczamy populację π_s z dalszych rozważań jako tę, do której przynależność obserwacji $\underline{x}_0$ wydaje się najmniej możliwa. Z pozostałych m - 1 populacji wyszukujemy ponownie parę najbardziej odległych w sensie miary Δ^{II} i z kolei z tymi dwiema populacjami postępujemy tak jak poprzednio. Pozwala to nam wyeliminować następną populację. Postępując sukcesywnie w ten sposób wyeliminujemy m - 1 populacji, do których przynależność obserwacji $\underline{x}_0$ wydaje się mało możliwa. Pozostała populacja jest tą, do której przynależy identyfikowana obserwacja $\underline{x}_0$.

Proponowana metoda identyfikacji jest metodą zamykania w jednym obszarze identyfikowanej obserwacji $\underline{x}_0$ oraz środka populacji, do której ta obserwacja należy. Warto zwrócić uwagę, że obszar ten jest tworzony z uwzględnieniem wartości $\underline{x}_0$ a nie, jak w metodach klasycznych, tylko na bazie danych pierwotnych.

5. Materiał doświadczalny stanowi 8 głosów męskich oraz 2 głosy żeńskie charakteryzujące się niską średnią częstotliwością podstawową. Każda z tych osób wypowiadała kolejno samogłoski i, y, e, a, o, u w pięciu powtórzeniach. Nagrania tych wypowiedzi poddano analizie spektrograficznej i pomierzono wartości częstotliwości pierwszych czterech formantów F_1, F_2, F_3, F_4. Wyniki pomiarów pochodzą z Pracowni Fonetyki Akustycznej ZAC IPPT PAN i zamieszczone są w pracy [2]. W ten sposób każda osoba scharakteryzowana została czterowymiarowym wektorem losowym. Rozkład tego wektora dla i-tej osoby traktujemy jako rozkład w populacji π_i, dla i = 1, 2, ..., 10. Przyjmujemy, że w każdej populacji wektor ten ma czterowymiarowy rozkład normalny z wektorem wartości średnich $\underline{m}_i$ i macierzą kowariancji Σ_i (i = 1, 2, ..., 10).

Po upływie trzech lat powtórzono identyczne badania dla dwóch osób spośród nagranych uprzednio z tym, że samogłoski wypowiadano

teraz w 10 powtórzeniach. Wyniki tych pomiarów zawarte są w pracy [4]. Celem tego doświadczenia było stwierdzenie, czy na podstawie uzyskanych pomiarów można poprawnie zidentyfikować osoby, od których pochodzą powtórne nagrania.

Metodę rachunkową zilustrujemy na przykładzie samogłoski e. Identyfikowaną osobą będzie osoba oznaczona inicjałem W J.

Tabela 1

Odległości Δ^{*}

e (F_1, F_2, F_3, F_4)

WJ	10.71								
HN	15.95	3.09							
ZM	3.03	5.50	5.41						
TL	2.70	11.14	11.23	2.28					
KD	24.69	3.02	3.29	12.86	20.74				
RK	15.74	11.08	11.01	6.96	9.62	16.72			
HK	18.41	6.79	10.21	6.08	9.51	4.79	6.47		
BK	20.04	5.97	12.06	16.11	17.76	13.67	24.32	17.22	
HS	23.92	3.52	11.83	14.44	25.04	10.48	25.84	10.64	6.62
	AS	WJ	HN	ZM	TL	KD	RK	HK	BK

Najpierw liczymy oceny z próby odległości Δ^{*} między dziesięcioma osobami. Wyniki obliczeń zawiera tabela 1. Z tabeli tej odczytujemy, że maksymalnie odległymi osobami są R K oraz H S. Dla nich $\Delta^{*} = 25{,}84$. Te dwie osoby rozgraniczamy hiperpłaszczyzną

$$-0{,}14387\,F_1 + 0{,}99754\,F_2 + 0{,}11340\,F_3 - 0{,}00425\,F_4 - 2041{,}22511 = 0$$

Następnie od lewej strony równania hiperpłaszczyzny w miejsce F_i, $i = 1, \ldots, 4$, wstawiamy kolejno wartości średnie formantów osób R K, H S oraz obliczone z powtórnych pomiarów wartości średnie formantów osoby W J i badamy znak uzyskanego wyrażenia. Dla osób H S oraz W J wyrażenie to jest dodatnie, zaś dla osoby R K ujemne. Oznacza to, że osoby W J oraz H S znajdują się po tej samej stronie hiperpłaszczyzny rozgraniczającej, osoba R K zaś znajduje się po stronie przeciwnej. Osobę R K eliminujemy z dalszych rozważań jako tę, od której pochodzenie ponownego nagrania wydaje się najmniej możliwe. W tabeli odległości wykreślamy wszystkie odległości związane z osobą R K. Z pozostałych odległości

znajdujemy maksymalną. Jest to odległość między osobami T L oraz H S i wynosi ona 25,04. Dla osób T L oraz H S budujemy hiperpłaszczyznę rozgraniczającą i stwierdzamy, że osoby W J oraz H S znajdują się po jednej jej stronie, zaś osoba T L po stronie przeciwnej. Osoba T L odpada jako druga z naszych rozważań. W dalszym postępowaniu kolejno odpadają: K D, A S, B K, Z M, H S, H K i H N. Pozostała osoba W J. Wskazanie to jest poprawne.

Identyfikacja przeprowadzona na wartościach średnich formantów z 10 powtórzeń dała wyniki poprawne dla wszystkich samogłosek z wyjątkiem "u". Tam gdzie uzyskano wyniki poprawne, była przeprowadzona identyfikacja dla poszczególnych wypowiedzi. Spośród dziesięciu wypowiedzi samogłoski "i" przez osobę W J było 7 rozpoznań poprawnych, dwa razy wypowiedzi utożsamione zostały z A S i jeden raz z H N. Dla "y" wszystkie rozpoznania były poprawne. Dla "e" uzyskano 9 rozpoznań poprawnych i jedno wskazanie na H N. Podobnie dla "a" było 9 rozpoznań poprawnych i jedno wskazanie na K D. Dla "o" uzyskano 5 rozpoznań poprawnych, po jednym wskazaniu na A S, Z M i H K oraz dwa wskazania na K D. Z uzyskanych wyników widać, że samogłoski "o" oraz "u" są małoprzydatne do rozpoznania osób.

Wyniki tej pracy należy traktować jako wstęp do dalszych badań nad rozróżnialnością głosów.

Literatura cytowana

[1] Anderson, T.W., Bahadur, R.R., Classification into two multivariate normal distributions with different covariance matrices, Ann. Math. Stat. 33 (1962), str. 420-431.

[2] Caliński, T., Jassem, W., Kaczmarek, Z., Investigation of vowel format frequencies as personal voice characteristics by means of multivariate analysis of variance, Speech Analysis and Synthesis Vol. II (1970), str. 7-39.

[3] Jassem W., Vowel format frequencies as cues to speaker discrimination, Speech Analysis and Synthesis Vol.I (1968), str. 9-41.

[4] Jassem, W., Krzyśko, M., Dyczkowski, A., Identyfikacja głosów przy zastosowaniu funkcji dyskryminacyjnych. Prace IPPT, 51 (1971).

[5] Krzyśko, M., Uogólniona odległość między populacjami normalnymi o różnych macierzach kowariancji, Listy Biometryczne Nr 30-33 (1971), str. 49-52.