BIULETYN INSTYTUTU AUTOMATYKI I ROBOTYKI
NR 22, 2005

Metoda weryfikacji mówcy na podstawie nieuzgodnionej wypowiedzi

Leszek GRAD

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

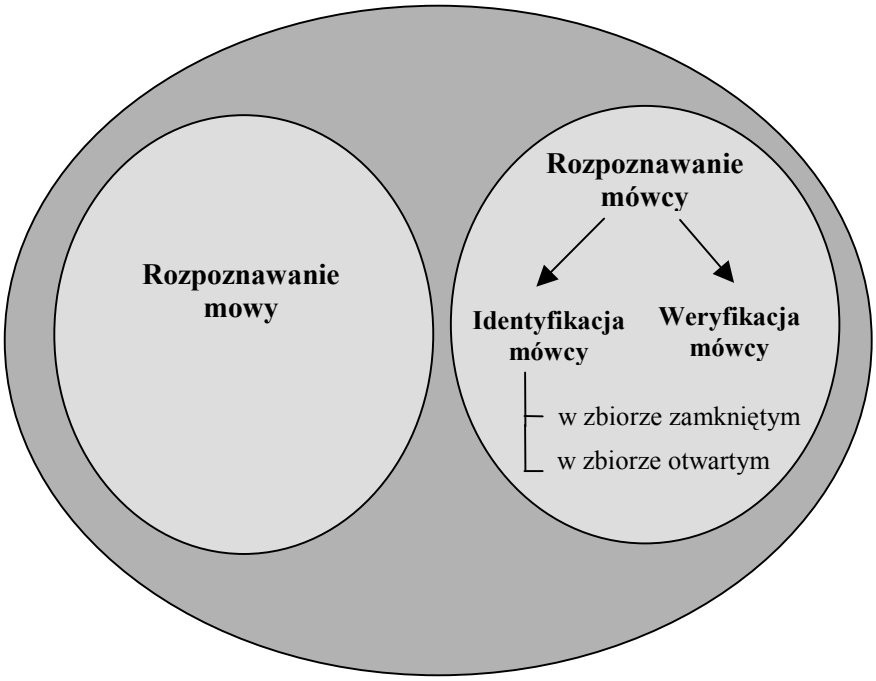
STRESZCZENIE: W artykule została przedstawiona metoda weryfikacji mówcy na podstawie nieuzgodnionej wypowiedzi (ang. *text independent*). Metoda ta oparta jest na metodzie niezależnej detekcji klas, zaliczanej do metod klasyfikacji minimalnoodległościowych.

Zagadnienie weryfikacji mówcy cieszy się obecnie dużym zainteresowaniem, głównie ze względu na fakt praktycznego wykorzystania do zdalnego uwierzytelniania osób w systemach informatycznych. Współczesne systemy silnego uwierzytelniania, bazujące na cechach biometrycznych, dla zapewnienia dużej niezawodności zmuszone są do realizacji weryfikacji tożsamości na podstawie większej liczby cech, np. obrazu siatkówki oka, linii papilarnych. Takie podejście sprawia, że celowe staje się opracowywanie i rozwijanie metod weryfikacji na podstawie głosu, nawet w przypadku kiedy stosowane samodzielnie nie gwarantują wystarczającej niezawodności.

Do podstawowych metod wykorzystywanych w zadaniach rozpoznawania osób na podstawie głosu należy zaliczyć: metodę ukrytych modeli Markowa [8] oraz metodę sztucznych sieci neuronowych. Zaproponowana w niniejszym artykule metoda niezależnej detekcji klas należy do grupy metod klasyfikacji minimalnoodległościowych. Jej cechą charakterystyczną jest to, że uwzględnia w procesie podejmowania decyzji rozproszenie wzorców.

1. Umiejscowienie zadania weryfikacji mówcy

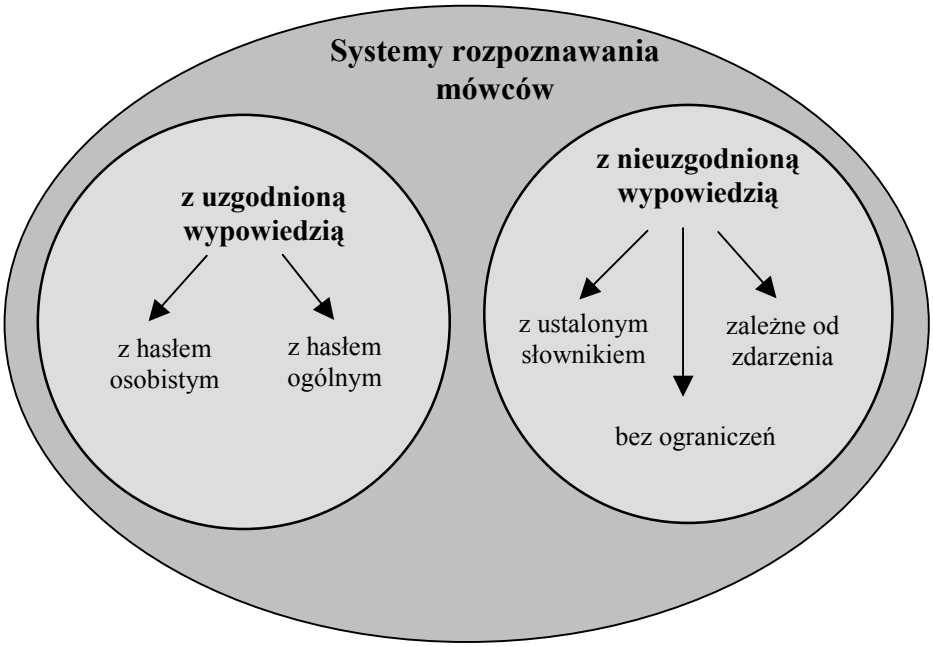
Zadania rozpoznawania w dziedzinie sygnału mowy można podzielić na rozpoznawanie mowy (treści wypowiedzi) oraz rozpoznawanie mówców (osób na podstawie próbki zarejestrowanego głosu) (rys. 1). Rozpoznawanie mówcy może polegać na jego identyfikacji (w zamkniętym lub otwartym zbiorze mówców) lub weryfikacji. W przypadku identyfikacji mówcy w zbiorze zamkniętym, system określa, do której spośród osób zarejestrowanych w systemie należy badana próbka głosu. W przypadku identyfikacji w otwartym zbiorze mówców, możliwa jest decyzja o odrzuceniu próbki. Weryfikacja mówcy jest zadaniem potwierdzenia deklarowanej tożsamości. Weryfikacja oraz identyfikacja w otwartym zbiorze są zadaniami blisko ze sobą związanymi ze względu na metodę.



Rys. 1. Klasyfikacja zadań rozpoznawania w dziedzinie sygnału mowy

Systemy rozpoznawania mówców można również podzielić, stosując jako kryterium wymagania systemu na wypowiedź. Podstawowy podział został pokazany na rysunku 2. Wyróżniamy systemy, które wymagają podania konkretnej wypowiedzi (hasła), oraz takie, których wymagania na wypowiedź są mniejsze (z ustalonym słownikiem, oczekujące na zdarzenie fonetyczne) lub nie

ma ich praktycznie wcale. Pierwsze z nich można nazwać systemami z uzgodnioną wypowiedzią (ang. *text dependent*), a drugie – systemami z nieuzgodnioną wypowiedzią (ang. *text independent*) [5].



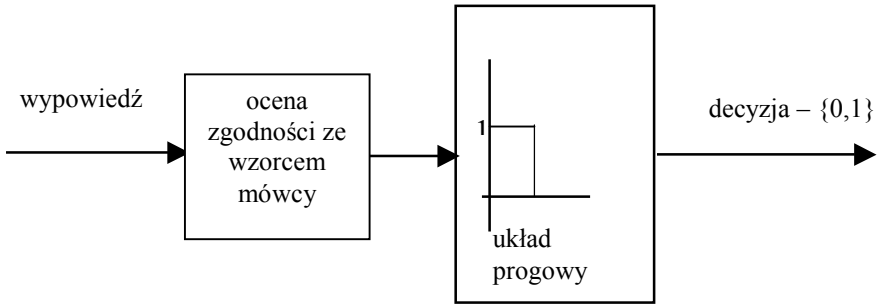
Rys. 2. Podział systemów rozpoznawania mówców ze względu na wymagania stawiane wypowiedzi

Prezentowana w artykule metoda weryfikacji mówcy nie stawia wymagań co do treści wypowiedzi, wymagane jest jedynie, aby wypowiedź zawierała okresy dźwięczne.

2. Ogólny schemat procesu weryfikacji mówcy

Systemy weryfikacji mówcy w sposób jawny lub ukryty badają zgodność wypowiedzi ze wzorcem osoby, której tożsamość jest deklarowana. W przypadku jawnego badania zgodności wyróżnia się dwa etapy. W etapie pierwszym obliczana jest odległość od wzorca. Etap drugi polega na sprawdzeniu, czy odległość nie przekracza odległości granicznej, zwanej progiem. Jeśli tak, to użytkownik przechodzi weryfikację pozytywnie. Ogólny schemat weryfikacji mówcy przedstawiony został na rysunku 3.

Zarówno wzorce mówców, jak i wartości progowe, są wyznaczane w procesie treningu (uczenia) poprzedzającym właściwą pracę systemu.



Rys. 3. Ogólny schemat procesu weryfikacji mówcy

3. Wskaźniki oceny jakości systemu weryfikacji mówców

W systemie weryfikacji mówcy wyróżnia się następujące zdarzenia:

- prawidłową akceptację – akceptację mówcy autentycznego (ma miejsce w przypadku gdy tożsamość: deklarowana i rzeczywista są zgodne i system zaakceptuje mówcę),

- fałszywą akceptację – akceptacją oszusta (ma miejsce w przypadku gdy tożsamość deklarowana nie jest zgodna z rzeczywistą, a system zaakceptuje mówcę),

- fałszywe odrzucenie – odrzucenie mówcy autentycznego (ma miejsce w przypadku gdy tożsamość: deklarowana i rzeczywista są zgodne, a system odrzuci mówcę),

- prawidłowe odrzucenie – odrzucenie oszusta (ma miejsce w przypadku gdy tożsamość deklarowana nie jest zgodna z rzeczywistą i system odrzuci mówcę).

Zdarzeniami niepożądanymi w systemie weryfikacji mówców są: fałszywe odrzucenie oraz fałszywa akceptacja[1]. Prawdopodobieństwa tych zdarzeń ocenia się na podstawie: stopy fałszywego odrzucenia oraz stopy fałszywej akceptacji. Sposób ich obliczania przedstawiono poniżej.

[1] Szczegółowy opis wskaźników oceny stosowanych w systemach rozpoznawania przedstawiono w [7].

Stopa fałszywego odrzucenia:

$$\alpha = 1 - \frac{1}{N}\sum_{i=1}^{N} W(u_i),\ N \neq 0$$

gdzie: u_i – wypowiedź należąca do weryfikowanego mówcy, N – liczba wypowiedzi mówcy.

$$W(u_i) = \begin{cases} 1 \text{ gdy } d(u_i) \leq \Theta \\ 0 \text{ gdy } d(u_i) > \Theta \end{cases},$$

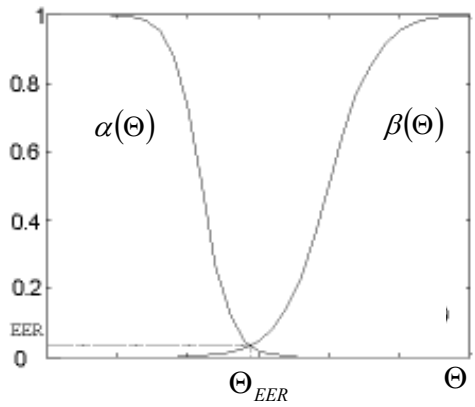
gdzie: u_i – wypowiedź mówcy, $d(u_i)$ – funkcjonał zgodności wypowiedzi u_i ze wzorcem mówcy, Θ – wartość progowa.

Stopa fałszywej akceptacji:

$$\beta = \frac{1}{M}\sum_{j=1}^{M} W(u_j), \quad M \neq 0,$$

gdzie: u_j – wypowiedź nienależąca do weryfikowanego mówcy (fałszywa), M – liczba wypowiedzi fałszywych.

Uniwersalnym miernikiem jakości działania sytemu weryfikacji mówcy, pozwalającym na porównywanie jakości działania różnych systemów weryfikacji, jest wskaźnik EER (ang. *Equal Error Rate*) wyznaczany jako wartość $\alpha(\Theta_{EER}) = \beta(\Theta_{EER})$ uzyskiwana w punkcie przecięcia przebiegów: stopy fałszywego odrzucenia i stopy fałszywej akceptacji w funkcji progu Θ (rys. 4). Fakt przyjęcia takiego wskaźnika oceny nie oznacza, że wartość progowa Θ powinna być ustalana w punktach przecięcia wykresów stóp. Zwykle dąży się do zapewnienia większego bezpieczeństwa działania systemu i rzeczywisty próg jest przesunięty, tak aby prawdopodobieństwo fałszywej akceptacji było mniejsze.



Rys. 4. Sposób określania wskaźnika EER

4. Ekstrakcja cech sygnału mowy

Proponowana w artykule metoda wymaga dokonania parametrycznego opisu ramek sygnału mowy. Wcześniejsze badania autora [3] wskazują na to, że opis sygnału mowy wektorem współczynników liniowego kodowania predykcyjnego (LPC ang. *Linear Predictive Coding*) daje dobre rezultaty w zadaniach rozpoznawania mówców [3]. Przesłankami przemawiającymi za wyborem tego sposobu opisu są także:

- współczynniki LPC – są wykorzystywane do modelowania toru głosowego człowieka; należy oczekiwać, że modele te będą różne dla różnych osób;
- metoda LPC – zapewnia dużą dokładność aproksymacji sygnału mowy; współczynniki LPC stosowane są do kodowania głosu (standardy LPC-10, CELP); metoda LPC zapewnia duży stopień kompresji przy jednoczesnym zachowaniu dobrej wierności odtworzenia, potwierdziły to także wyniki porównania z nieliniowym modelem predykcji [2];
- istnieją ścisłe zależności pomiędzy współczynnikami LPC a parametrami uzyskiwanymi z widma sygnału.

Opis sposobów wyznaczania współczynników LPC można znaleźć w [1].

5. Metoda niezależnej detekcji klas w zadaniu weryfikacji mówcy

Do rozwiązania zadania weryfikacji mówcy wykorzystano metodę niezależnej detekcji klas, zaliczaną do metod klasyfikacji minimalno-odległościowej [6]. Charakteryzuje ją to, że odległość do każdej klasy (zgodność ze wzorcem klasy) mierzona jest w sposób uwzględniający korelację i rozproszenie elementów z klasy.

Do oceny zgodności wektora cech $\mathbf{u}$ z klasą wykorzystuje się liniowe, ortogonalne przekształcenie Karhunena-Loeve'a (K-L) o macierzy przekształcenia $\mathbf{T}$, utworzonej z wektorów własnych $\mathbf{t}_i$ macierzy kowariancji $\mathbf{R} = E\left(\mathbf{u}_j - \overline{\mathbf{u}}\right)\left(\mathbf{u}_j - \overline{\mathbf{u}}\right)^T$ w następujący sposób:

$$\mathbf{T} = \left[\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_p\right]^T,$$ p - długość wektora cech $\mathbf{u}_j$.

Wektory $\mathbf{u}_j$ należą do jednej klasy w zbiorze uczącym, a $\overline{\mathbf{u}}$ jest wartością oczekiwaną elementów tej klasy.

Ocena zgodności badanej cechy z zadaną klasą polega na obliczeniu różnicy badanej cechy $\mathbf{u}$ i wartości oczekiwanej $\overline{\mathbf{u}}$ cech w klasie, tzn. wartości:

$$\widetilde{\mathbf{u}} = \mathbf{u} - \overline{\mathbf{u}},$$

a następnie na zastosowaniu do otrzymanej różnicy przekształcenia K-L:

$$\mathbf{y} = \mathbf{T}\widetilde{\mathbf{u}}.$$

Unormowanie oceny zgodności polega na unormowaniu składowych transformaty K-L w następujący sposób:

$$\mathbf{z} = \left[z_1, z_2, ..., z_p\right]^T = \left[\frac{y_1}{\sqrt{\lambda_1}}, \frac{y_2}{\sqrt{\lambda_2}}, ..., \frac{y_p}{\sqrt{\lambda_p}}\right]^T$$

Jako funkcjonał oceny zgodności wektora cech $\mathbf{u}$ z klasą przyjmuje się kwadrat normy wektora $\mathbf{z}$:

$$d\left(\mathbf{u}\right) = \mathbf{z}^T\mathbf{z} = \sum_{k=1}^{p} z_k^2 = \sum_{k=1}^{p} \frac{y_k^2}{\lambda_k}.$$

Proces weryfikacji z wykorzystaniem niezależnej detekcji klas musi zostać poprzedzony procesem uczenia. Uczenie polega na wyznaczeniu dla każdej klasy, na podstawie zbioru uczącego, trójki $(\bar{\mathbf{u}}, \mathbf{T}, \boldsymbol{\lambda})$.

Określanie wzorca mówcy – uczenie

Poniżej przedstawiony został algorytm określania wzorców, zwany też procesem uczenia lub treningu.

1. Pozyskanie kilkunastosekundowego nagrania wypowiedzi mówcy, nazwanego *wypowiedzią uczącą*.

2. Podział wypowiedzi uczącej na segmenty czasowe o jednakowej długości. Długość segmentu powinna wynosić od kilkunastu do kilkudziesięciu milisekund.

3. Wyznaczenie dla każdego segmentu wektora cech w postaci wektora $\mathbf{a}$ współczynników LPC. Informacja o położeniu wektorów cech w czasie nie jest istotna z punktu widzenia proponowanej metody. Dlatego też otrzymane wektory traktowane są jako zbiór: $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, .., \mathbf{a}_N\}$, gdzie N jest liczbą segmentów.

4. Grupowanie wektorów cech.

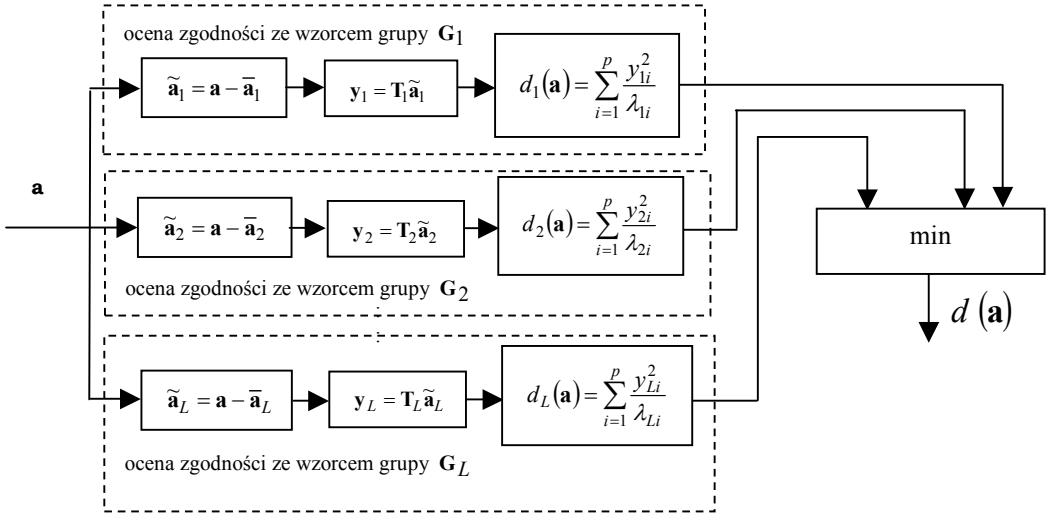
   W procesie tym zbiór wektorów cech $\mathbf{A}$, otrzymany z wypowiedzi uczącej, poddawany jest procesowi grupowania. W tym celu można wykorzystać jedną ze znanych metod grupowania. W wyniku procesu grupowania, zbiór $\mathbf{A}$ zostaje podzielony na L grup $\mathbf{G}_i, i = 1,2,..., L$. Oznacza to, że z wypowiedzi uczącej każdego mówcy zostaje wyłonionych L grup cech.

5. Utworzenie wzorca mówcy.

   Dla każdej grupy $\mathbf{G}_i$, w procesie uczenia wyznaczyć należy trójkę $(\bar{\mathbf{a}}_i, \mathbf{T}_i, \boldsymbol{\lambda}_i)$.
   Wzorzec mówcy jest rozumiany jako zbiór L trójek $(\bar{\mathbf{a}}_i, \mathbf{T}_i, \boldsymbol{\lambda}_i)$, $i = 1,2,..., L$, z których każda stanowi wzorzec grupy wektorów cech.

Przebieg procesu weryfikacji

Wskaźnik zgodności $d(\mathbf{a})$ wektora cech ze wzorcem mówcy jest wyznaczany w układzie niezależnej detekcji klas, zgodnie ze schematem przedstawionym na rysunku 5.



Rys. 5. Schemat oceny zgodności wektora cech ze wzorcem mówcy

Ocena zgodności wypowiedzi A ze wzorcem mówcy jest dana wzorem:

$$d(\mathbf{A}) = \frac{1}{M}\sum_{i=1}^{M} d(\mathbf{a}^i), \qquad \mathbf{A} = \{\mathbf{a}^1, \mathbf{a}^2, ..., \mathbf{a}^M\},$$

gdzie M jest liczbą segmentów wydzielonych z wypowiedzi A.

Odpowiedź systemu weryfikacji mówcy jest wyznaczana zgodnie z zależnością:

$$W(\mathbf{A}) = \begin{cases} 1, & \text{gdy } \mathrm{d}(\mathbf{A}) \le \Theta \\ 0, & \text{gdy } \mathrm{d}(\mathbf{A}) > \Theta \end{cases}.$$

6. Wyniki badań

Badania weryfikacji mówców przeprowadzono z wykorzystaniem bazy nagrań Student. Baza ta zawiera 2800 elementów. Słownik bazy składa się z 20 słów (tab. 1.), 20 powtórzeń każdego słowa przez każdego z 7 mówców, czyli: 7 mówców x 20 słów x 20 powtórzeń = 2800 elementów. Próbki sygnału mowy zostały nagrane przy pomocy komputera klasy PC z systemem operacyjnym Windows, wyposażonego w kartę muzyczną klasy Sound Blaster. Parametry akwizycji sygnału były następujące: częstotliwość próbkowania 11kHz, kwantyzacja 16-bitowa.

Tab. 1. Wykaz słów słownika bazy Student

Lp.	Słowo
1	zero
2	jeden
3	dwa
4	trzy
5	cztery
6	pięć
7	sześć
8	siedem
9	osiem
10	dziewięć
11	broda
12	Danuta
13	kałamarz
14	kapitan
15	mama
16	metoda
17	moda
18	sezam
19	szczęście
20	zdrowie

Badania weryfikacji mówcy przeprowadzono przy następujących wartościach parametrów segmentacji czasowej i ekstrakcji parametrów:

- szerokość okna czasowego 30 ms,
- skok okna czasowego 10 ms,
- liczba współczynników LPC p=10.

Tab. 2. Podział słów ze słownika na zbiór uczący i testowy

Zbiór uczący		Zbiór testowy
jeden	Danuta	siedem
dwa	kałamarz	osiem
trzy	kapitan	dziewięć
cztery	mama	moda
pięć	szczęście	sezam
sześć	metoda	zdrowie
broda		

Liczbę grup, na którą dzielono wektory cech w trakcie procesu uczenia, przyjęto równą 11 [5]. Wypowiedzi uczące zostały utworzone ze słów słownika bazy Student zamieszczonych w lewej kolumnie tabeli 2. Pozostałe (zamieszczone w prawej kolumnie) zostały wykorzystane do testów.

Ze słów słownika bazy utworzono wypowiedzi testowe (tab. 3.), które następnie podano na wejście systemu. Dla testu wygenerowanych zostało 2450 wypowiedzi. Otrzymane wyniki przedstawione zostały w tabeli 4. Wyznaczone zostały indywidualne dla każdego mówcy progi układu decyzyjnego. Wskaźniki EER (stopa fałszywego odrzucenia równa stopie fałszywej akceptacji) wahały się dla poszczególnych mówców od 0,01 do 0,25. Wartość średnia dla siedmiu mówców wyniosła EER=0,07. Wynik należy uznać za zadowalający.

Tab. 3. Wypowiedzi testowe

Wypowiedzi testowe
siedem_osiem_dziewięć
osiem_dziewięć_moda
dziewięć_moda_sezam
moda_sezam_siedem
sezam_siedem_zdrowie

Tab. 4. Wyniki weryfikacji mówców z bazy Student

	Mówca						
	A	B	C	D	E	F	G
EER	0,135	0,030	0,015	0,250	0,010	0,045	0,010

7. Podsumowanie

Zaproponowana w pracy metoda weryfikacji realizuje rozpoznawanie w przestrzeni współczynników LPC sygnału mowy. Wykorzystuje technikę grupowania oraz metodę niezależnej detekcji klas, opartą na transformacie Karhunena-Loeve`a. Cechuje się: dobrą skutecznością (wartość średnia EER=0,07), niską uciążliwością pozyskiwania danych do procesu uczenia (kilkunastosekundowa wypowiedź ucząca) oraz niskimi kosztami obliczeniowymi procesu uczenia i rozpoznawania.

Literatura

[1] Basztura Cz. i inni: *Metody parametryzacji sygnału mowy do automatycznego rozpoznawania głosów*, Prace Naukowe ITiA Politechniki Wrocławskiej, nr 31, 1990.

[2] Grad L.: *Badania porównawcze zastosowania liniowego i nieliniowego modelu predykcji w analizie sygnału mowy*, Biuletyn IAiR, nr 10, WAT, Warszawa, 1999.

[3] Grad L.: *Badanie możliwości rozpoznawania mówcy na podstawie reprezentacji LPC sygnału mowy*, Biuletyn IAiR, nr 13, WAT, Warszawa, 2000.

[4] Grad L.: *Zastosowanie transformaty Karhunena-Loeve'a do rozpoznawania mówcy*, Biuletyn IAiR, nr 13, WAT, Warszawa, 2000.

[5] Grad L.: *Rozpoznawanie mówcy metodą niezależnej detekcji klas, Rozprawa doktorska*, WAT, 2002.

[6] Kwiatkowski W.: *Metody automatycznego rozpoznawania wzorcó*w, WAT, Warszawa, 2001.

[7] Wiśniewski A. M.: *Metody oceny systemów rozpoznawania mówców*, Biuletyn IAiR, nr 13, WAT, Warszawa, 2000.

[8] Wiśniewski A. M.: *Niejawne modele Markowa w rozpoznawaniu mowy*, Biuletyn IAiR, nr 7, WAT, Warszawa, 1997.

Recenzent: prof. dr hab. inż. Włodzimierz Kwiatkowski

Praca wpłynęła do redakcji: 24.12.2005