Analiza sygnału mowy sterowana danymi dla rozpoznawania komend głosowych*

Włodzimierz Kasprzak[1], Adam B. Kowalski[1]

Streszczenie

W artykule omówiono podstawowe etapy analizy sygnału mowy "sterowanej danymi": detekcji sygnału użytecznego w sygnale dźwiękowym, segmentacji sygnału mowy, detekcji cech segmentów i kodowania cech w kategoriach fonetycznych. W fazie uczenia systemu rozpoznającego te etapy prowadzą do stworzenia słownika kodowego dla ramek sygnału a w fazie aktywnej pracy dostarczają one wektory cech dla sekwencji ramek sygnału. Dla porównania poprawności różnych wersji bazowego wektora cech zrealizowano klasyfikator sekwencji wektorów cech.

1. WSTĘP

Jedną z dziedzin zastosowań analizy sygnału mowy stanowi komunikacja głosowa operatora ludzkiego z usługowym robotem mobilnym. W systemie zbudowanym na bazie programowej struktury ramowej MRROC++ [9] za efektor uważa się każde urządzenie mogące oddziaływać na otoczenie, a więc zarówno manipulator jak i głośnik. Natomiast receptorem jest dowolne urządzenie zbierające dane o stanie środowiska, a wiec także mikrofon. Układ programowy - sterownik - takiego receptora realizuje automatyczne wykrywanie mowy i analizę mowy w terminach zadanego zestawu komend. Całościowa struktura programu rozpoznawania komend głosowych i niezbędne modele fonetyczne, służące do analizy "sterowanej modelem", opisane zostały w innym artykule [6].

Niniejsza praca koncentruje się na zagadnieniach przetwarzania wstępnego i analizy sygnału "sterowanej danymi". W ramach potoku analizy sygnału wyróżnimy: detekcję sygnału mowy w sygnale dźwiękowym, optymalne rozmieszczenie ramek (problem segmentacji sygnału), detekcję cech ramki, kodowanie cech w terminach jednostek fonetycznych [5].

Typowe podejście do segmentacji sygnału mowy zakłada podział sygnału na ramki o krótkim okresie trwania [3], [8]. Ważną sprawą w tym zakresie jest sposób rozmieszczenia ramek, tak aby pokrywały one w pełni istotne fragmenty sygnału.

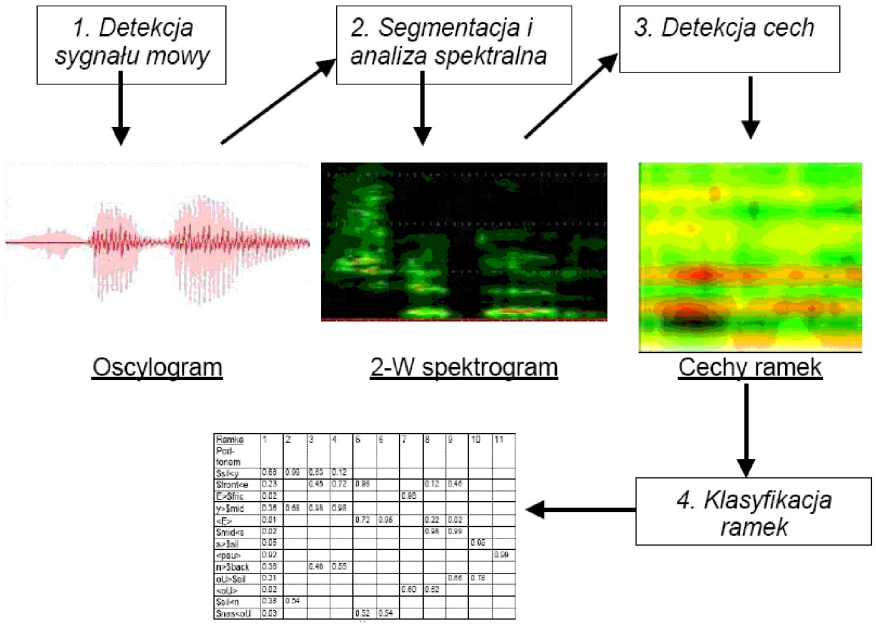
Następnym istotnym zagadnieniem jest właściwy wybór wektora cech ramki. Chociaż w literaturze przyjęło się alternatywne stosowanie schematu LPC (kodowania w liniowej predykcji) lub cech MFCC (tzw. cech mel cepstralnych) ([2], [8]), to jednak sprawą otwartą pozostaje dobór właściwego zestawu cech. Dla zapewnienia

*Praca jest finansowana przez grant MiNI: 4T11A 003 25.

[1]Instytut Automatyki i Informatyki Stosowanej, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska, ul. Nowowiejska 15/19, 00–665 Warszawa, W.Kasprzak@ia.pw.edu.pl

skuteczności rozpoznawania niezbędna jest drobiazgowa analiza rozkładów klas fonetycznych w różnych możliwych przestrzeniach cech.

2. PROCES ANALIZY MOWY STEROWANEJ DANYMI



Rys. 1. Potok analizy sterowanej danymi w systemie rozpoznawania mowy

W każdym systemie analizy mowy (rys. 1) możemy wyróżnić początkowe (tzn. bliskie poziomowi sygnału) etapy analizy - zasadniczo jest to analiza sterowana danymi (ang. data-driven, "bottom-up") - i etapy "wyższego poziomu" (tzn. posługujące się opisami symbolicznymi i jawnymi modelami rozpoznawanych słów, zdań lub mowy ciągłej) - składające się na analizę sterowaną modelem.

Niniejsza praca ogranicza się do proponowania rozwiązań dla etapów analizy sterowanej danymi, realizowanej na potrzeby systemu rozpoznawania komend głosowych.

3. DETEKCJA SYGNAŁU MOWY

Zasadnicze kryteria detekcji sygnału użytecznego mowy w sygnale dźwiękowym to: właściwy poziom energii sygnału i występowanie na przemian dźwięcznych i bezdźwięcznych fragmentów głosek.

Wyznaczenie impulsów energetycznych realizowane jest metodą automatu skończonego opartą na pracach [4], [7]. Początek impulsu występuje wtedy, gdy ener-

gia widma wzrośnie powyżej progu K1 oraz nie spadając poniżej tej wartości wzrośnie powyżej parametru K2. Koniec impulsu oznaczany jest wtedy, gdy energia spadnie poniżej progu K3 i w okresie czasu L od momentu spadku nie wzrośnie powyżej progu K4. Parametry K1, K2, K3, K4 oraz L wyznaczone zostały na podstawie eksperymentów i wynoszą odpowiednio: K1=0.1, K2=0.15, K3=0.12, K4=2 i L=200 msec.

Obliczamy szereg współczynników auto-korelacji okna sygnału, rozpoczynającego się od próbki o indeksie czasowym m, z sygnałem przesuniętym o k chwil czasowych, dla okna sygnału o szerokości N:

$$r_m^{(k)} = \frac{\sum_{n=m}^{m+N-k-1} s_n \cdot s_{n+k}}{|[s_n, \ldots, s_{n+N-1}]| \cdot |[s_{n+k}, \ldots, s_{n+k+N-1}]|} \tag{1}$$

Oczywiście maksymalna wartość auto-korelacji przypada dla $k = 0$ i wynosi:

$$r_m^{(0)} = 1, \forall m. \tag{2}$$

Dla dźwięcznego fragmentu sygnału mowy obserwujemy charakterystyczne powtarzanie się drgań harmonicznych, dlatego co pewne $k0$ jednostek czasu zachodzi:

$$r_k \approx r_{k+j \cdot k0}. \tag{3}$$

Występowanie wielokrotnego maksimum funkcji auto-korelacji pozwala na odróżnienie głosek dźwięcznych od bez-dźwięcznych.

4. SEGMENTACJA SYGNAŁU MOWY

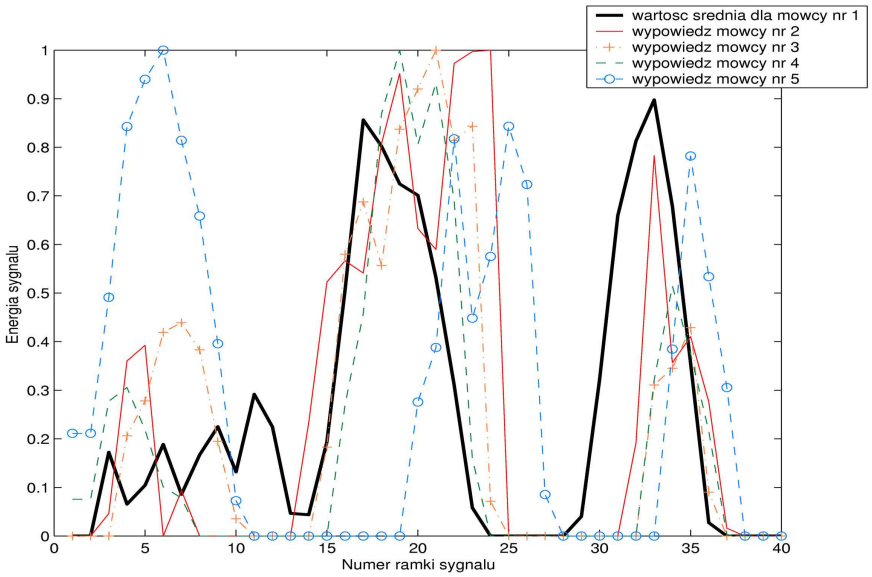
Zadaniem tego etapu analizy mowy jest optymalne rozmieszczenie okien sygnału. Analizując przebieg energii sygnału wyznaczamy ich lokalne maksima. Staramy się rozmieścić sekwencję ramek o tych samych rozmiarach i o tych samych odstępach w taki sposób, aby lokalne maksima energii przypadały na środki ramek.

Opcjonalnym krokiem jest *przepróbkowanie* wybranego fragmentu sygnału (*resampling*). Fragment sygnału zawierający wyselekcjonowane polecenie głosowe należy znormalizować do określonej ilości ramek, tak aby możliwa była późniejsza klasyfikacja. W programie ilość ramek zadana jest parametrem, a domyślna wartość wynosi 40 (40 ramek x ok. 12ms $\approx$ 480ms), co dla danych testowych zawierających komendy średnio o długości 0.5 sekundy oznacza, że marszczenie czasu nie będzie wprowadzało znaczących błędów. W pierwszym podejściu *resampling* zaimplementowany został wykorzystując algorytm najbliższego sąsiada. W celu poprawienia działania wykorzystano ostatecznie interpolację liniową. Jeśli x określa wartość z przedziału: $x_0 < x < x_1$; $y_0 = f(x_0)$ i $y_1 = f(x_1)$ są wartościami danej funkcji a $h = x_1 - x_0$ jest odstępem pomiędzy argumentami, to liniowa interpolacja $L(x)$ funkcji f wynosi:

$$L(x) = y_0 + \frac{y_1 - y_0}{h}(x - x_0). \tag{4}$$

5. DETEKCJA CECH RAMEK

Dalsza analiza sygnału mowy w dziedzinie czasu nie prowadzi do uzyskania dobrych cech ramek. Główną przeszkodą jest duża zmienność przebiegu energii sygnału w różnych wypowiedziach słowa przez różnych mówców (rys. 2).



Rys. 2. Rozkład energii dla wypowiedzi słowa "puść" - średnie dla 5 różnych mówców

Dlatego też powszechnie stosowane w systemach rozpoznawania mowy są cechy *cepstralne* sygnału mowy będące wynikiem charakterystycznego (homomorficznego) przekształcenia:

$$MFCC(h) = FT^{-1}\{MFC\{FT\{h\}\}\}, \text{ dla } \mathbf{h} = \mathbf{x} \otimes \mathbf{w}, \tag{5}$$

gdzie FT oznacza transformatę Fouriera, a $\otimes$ oznacza splot $\mathbf{x}$ z $\mathbf{w}$.

W ogólności "cepstrum" ma wartości zespolone. Jednak dla MFC(.) wynik powyższego przekształcenia ma wartości rzeczywiste. Dlatego też w FT^{-1} wykorzystane będzie tylko przekształcenie "kosinus". W praktyce stosujemy okienkową szybką transformatę Fouriera (FFT) dla dyskretnego sygnału:

$$F(\tau,k) = \frac{1}{\sqrt{M}} \sum_{t=0}^{M-1} x(t) \cdot w_\tau(t) \cdot m^{kt}, \text{ dla } k = 0,1,...,M-1, \tag{6}$$

gdzie $m = \exp(-i \cdot 2\pi/M)$, dla ramki sygnału o szerokości M i indeksie $t = \tau$. Skończona i stosunkowo krótka długość okna sygnału sprawia, że nie jest spełniony warunek okresowości sygnału w tym oknie. Zaproponowano wiele funkcji okna o łagodnie opadających "zboczach" po obu końcach okna, co zmniejsza szkodliwy efekt wzmacniania wyższych częstotliwości w spektrum. Okno Hamminga jest szeroko stosowane

w tym kontekście. Ramka sygnału jest mnożona przez funkcję w postaci dzwonu:

$$w_\tau(t) = 0.54 - 0.46\cos\left(\frac{2\pi t}{M-1}\right) \tag{7}$$

gdzie M - rozmiar okna (liczba próbek sygnału w ramce).

Ucho człowieka reaguje nieliniowo na częstotliwości sygnału dźwięku - różnice w zakresie niskich częstotliwości (< 1 kHz) są łatwiej wykrywane aniżeli podobne różnice w zakresie wysokich częstotliwości słyszanego spektrum. Innymi słowy, im wyższa częstotliwość tym gorsza dokładność i tym większe odstępy między kolejnymi pasmami są potrzebne dla zrekompensowania nieliniowości. Tworzony jest zbiór filtrów dla kolejnych pasm częstotliwości, rozmieszczonych w nieliniowy sposób wyznaczony przez tzw. *Mel-skalę*. Filtry są zdefiniowane w dziedzinie częstotliwości co umożliwia proste wymnożenie ich wartości przez przekształcony sygnał.

Korzystamy ze zbioru L trójkątnych filtrów $\{D(l,k)\}$ dla obliczenia (np. L = 32) tzw. współczynników Mel-spektralnych $MFC(\tau,l)$ dla każdej ramki sygnału τ:

$$FC(\tau,k) = |F(\tau,k)|^2 \tag{8}$$

$$MFC(\tau,l) = \sum_{l=1}^{L} D(l,k)\cdot FC(\tau,k), \text{ dla } l = 1,...,L; k = 0,1,...,M-1. \tag{9}$$

Wartość pojedynczego współczynnika $MFC(\tau,l)$ odpowiada ważonej sumie wartości $FC(\tau,.)$ należących do zakresu trójkątnego filtra pasmowego odpowiadającego danemu współczynnikowi MFC.

Ponieważ układ głosu ma charakter ciągły, poziomy energii w sąsiednich pasmach są skorelowane. Dlatego też niezbędna transformata odwrotna Fouriera (tu wystarczy przekształcenie kosinusowe, gdyż współczynniki MFC posiadają wartości rzeczywiste) zamienia zbiór logarytmów energii na nie-skorelowane ze sobą współczynniki "cepstralne". K współczynników "mel-cepstralnych" MFCC (np. K=12) oblicza się według wzoru:

$$MFCC(\tau,k) = \sum_{l=1}^{L} log[MFC(\tau,l)]\cos\left[\frac{k(2l-1)\pi}{2M}\right],\ k = 1,...,K. \tag{10}$$

Nadal jednak nieużyteczna dla rozpoznawania głosek energia podstawowej częstotliwości drgań krtaniowych i jej harmoniczne nakładają się na amplitudy mierzonych częstotliwości. Celem następnego kroku "liftrowania" jest usunięcie tego szkodliwego wpływu z zestawu cech.

Cechy MFCC o indeksie równym lub wyższym niż indeks cechy o maksymalnej wartości (który odpowiada częstotliwości podstawowej mówcy) nie są reprezentatywne dla wymawianego dźwięku a jedynie dla samego mówcy i nie są brane pod uwagę w systemie rozpoznawania mowy.

Jednak pozostałe cechy MFCC (o niższych indeksach) zawierają energie składowych harmonicznych od częstotliwości podstawowej - efekt drgań krtaniowych. Aby usunąć wpływ częstotliwości podstawowej na te pozostałe cechy MFCC, wektor

MFCC poddajemy tzw. liftrowaniu (ang. fil | ter –> lif | ter) [8], [2]. Ta operacja odpowiada dolnoprzepustowej filtracji sygnału czasowego i przenosi wektor MFCC w dziedzinę częstotliwości). Niech c_n oznacza n-ty współczynnik MFCC. Wtedy:

$$c_n^{lift} = \left[1 + \frac{L}{2}\sin\left(\frac{\pi n}{L}\right)\right] c_n,\ n = 1,2,...,K < L, \tag{11}$$

gdzie stała K odpowiada indeksowi cechy "związanej z częstotliwością podstawową". W praktyce za wartość stałej K przyjmuje się liczbę filtrów pasmowych L.

Na koniec wektor cech MFCC uzupełniamy o wartość energii ramki sygnału. Sumaryczną energię w ramce sygnału τ obliczymy w dziedzinie czasu.

6. WYZNACZANIE SŁOWNIKA KODOWEGO DLA CECH

Domyślne klasy jednostek fonetycznych mogą być wyznaczone ze zbioru wektorów cech metodą klasteryzacji w połączeniu z kwantyzacją wektorową (np. wykonywaną algorytmem LBG).

Algorytm klasteryzacji

DANE wejściowe: zbiór wektorów cech $\omega = \{\mathbf{c}_i | i = 1,...,N\}$, próg dla błędu Γ.
KROKI:
1) Oblicz wartość średnią próbek ω: $\mu^0 = \frac{1}{N}\sum_i \mathbf{c}_i$.
2) Ustaw początkowy słownik $Z^{(0)}$ dla $K_0 = 2$ klas o prototypach:
$z_{1,2} = (1 \pm \delta)\mu^0$, gdzie $\delta << 1$.
3) Iteruj kroki (4)-(6) dla $K = K_0, 2 \cdot K_0, 4 \cdot K_0, ...$
4) Wywołaj $LBG(\omega, K, Z^{(0)})$. Wynikiem jest nowy słownik $Z^{(I)}$ i jego błąd $e^{(I)}$.
5) IF ($e^{(I)} < \Gamma$) THEN STOP.
6) Ustaw nowy początkowy słownik $Z^{(0)}$ dla $2 \cdot K$ klas o prototypach:
$z_{k,K+k} = (1 \pm \delta) z_k$, gdzie $k = 1,...,K$ i $\delta << 1$.
7) Przejdź do (3) i wykonuj kolejną iterację (3-6).

Algorytm LBG ma charakter adaptacyjny i kolejne iteracje wprowadzają coraz większy porządek w próbkach. Z czasem ruchy prototypów i obszarów klas przestają się zmieniać. Wtedy zbiór próbek można uznać za ostatecznie sklasyfikowany.

7. IMPLEMENTACJA I WYNIKI TESTOWE

Opisane etapy analizy sygnału mowy sterowanej danymi zostały zaimplementowane w języku C++ i były testowane w stworzonym do tego celu programie testowym. Nie uwzględniamy tu jawnie klas fonetycznych, wynikających z modelu języka narodowego, a jedynie mamy dostęp do wyników klasteryzacji wektorów cech stanowiących próbki uczące. Dlatego też zagadnienie rozpoznawania komend sprowadza się do optymalnego dopasowania sekwencji reprezentantów klas, stanowiącej model słowa, do aktualnie wyznaczonej sekwencji wektorów cech.

Klasyfikator geometryczny minimalnej odległości jest prosty w realizacji i został zastosowany do testowania skuteczności rozpoznawania komend. Jako miarę

odległości wyznaczonego wektora $\mathbf{c}$, w n-wymiarowej przestrzeni cech, od reprezentanta $\mathbf{z}^{\kappa}$ zadanej klasy Ω_{κ} przyjęto ważoną odległość Euklidesa:

$$D(\mathbf{z}^{(\kappa)}, \mathbf{c}) = \sqrt{\sum_{i=1}^{n} \frac{(c_i - z_i^{\kappa})^2}{\sigma_i^2}}, \tag{12}$$

gdzie waga σ_i jest wariancją i-tej składowej próbek uczących, obliczoną jako:

$$\sigma_i^2 = \frac{1}{N} \sum_{k=1}^{N} (c_i^{(k)} - z_i^{(k)})^2, \tag{13}$$

i wektor $\mathbf{c}^{(k)}$ przynależy do klasy Ω_k oraz N jest liczbą próbek uczących.

Niech $\mathbf{S} = [\mathbf{z}_1, ..., \mathbf{z}_M]^T$ będzie reprezentacją słowa w postaci sekwencji M reprezentantów klas ramek sygnału. Odległość pomiędzy $\mathbf{S}$ a aktualną sekwencją cech $\mathbf{C} = [\mathbf{c}_1, ..., \mathbf{c}_M]^T$ obliczymy jako sumę odległości wszystkich par $(\mathbf{z}_i, \mathbf{c}_i)$:

$$T(\mathbf{S}, \mathbf{C}) = \frac{1}{M} \sum_{i=1}^{M} D(\mathbf{z}_i, \mathbf{c}_i). \tag{14}$$

Oczywiście funkcja decyzyjna klasyfikatora minimalno-odległościowego wybiera dla aktualnej sekwencji wektorów $\mathbf{C}$ to słowo $\mathbf{S}$, dla którego wartość $T(\mathbf{S}, \mathbf{c})$ jest najmniejsza. Jednak pozostaje pytanie, czy taka decyzja zawsze jest wiarygodna? Narzędziem porównania jakości różnych wektorów cech będzie dla nas nie tylko poprawność wyboru słowa lecz także wiarygodność takiego wyboru.

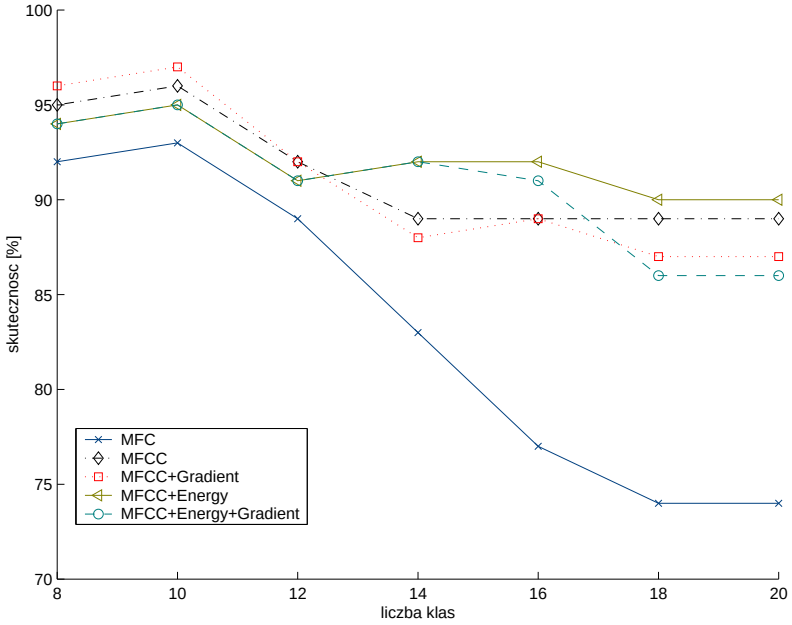
Wprowadzimy funkcję *wiarygodności klasyfikacji* wyrażoną wzorem:

$$Q = \begin{cases} (1 - T) \cdot 100\%, & \text{dla } T \leq 1 \\ 0, & \text{dla } T > 1 \end{cases} \tag{15}$$

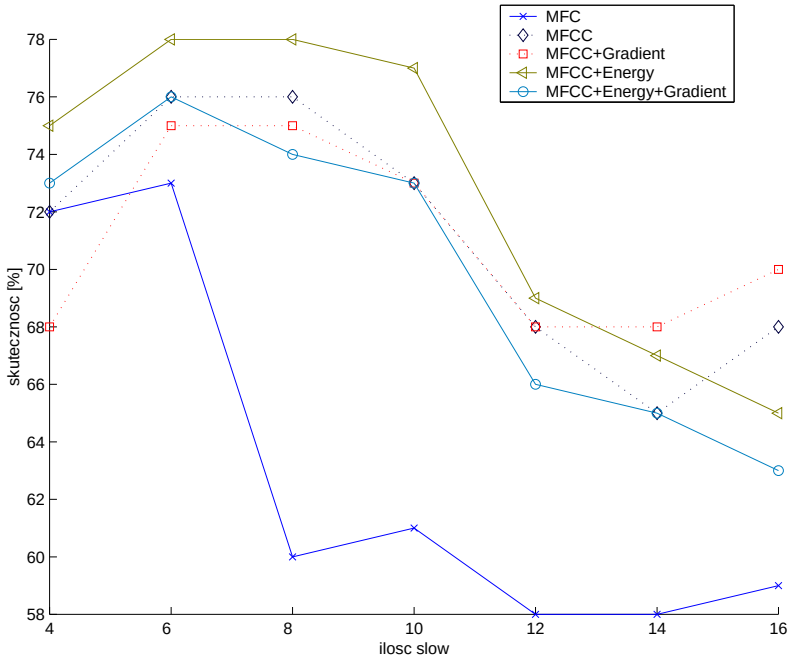
Wyniki klasyfikacji dla zbiorów zawierających wypowiedzi jednego mówcy zebrano w tabeli 1. Zbiór testowy zawierał 20 słów, gdzie dla każdego słowa podano 18 próbek uczących i 12 próbek testowych. W tej serii testów przyjęto, że próbka była poprawnie identyfikowana, jeśli wartość współczynnika wiarygodności $>= 50\%$ i gdy wybrano właściwe słowo. Wyniki klasyfikacji dla zbiorów zawierających wypowiedzi sześciu mówców zawarto w tabeli 2. Zbiór testowy zawierał 20 słów, gdzie dla każdego słowa i każdego mówcy podanych było 18 próbek uczących i 12 testowych.

Analizując wyniki z tabel 1 i 2 zauważymy ciekawe zależności poprawności rozpoznawania od zestawów cech. Wydawało się, że im bardziej zróżnicowany zestaw, tym większy procent poprawnej identyfikacji. Tymczasem zbiór zawierający cechy MFCC, energię i gradienty (jeden mówca - 83.45% i wielu mówców - 62%) dał gorszy wynik od zbioru zawierającego cechy MFCC i energię (87.60% i 63.30%) oraz od zbioru z cechami MFCC i ich gradientami (85.65% i 72.20%).

Na rys. 3 i 4 przedstawiono charakterystyki poprawnego rozpoznawania (przy co najmniej 50% wiarygodności) w zależności od liczby rozpoznawanych słów. Na podstawie tych wyników stwierdzamy, że poprawność procesu rozpoznawania jest zdecydowanie większa dla zbioru wypowiedzi jednego mówcy (82.55%) od zbioru z wieloma mówcami (65.36%). Wydaje się to być oczywiste, tym bardziej, że dla przypadku z wieloma mówcami komendy wypowiadane były przez lektorów zróżnicowanych przede wszystkim pod względem płci i wieku.



Rys. 3. Poprawność rozpoznawania w zależności od ilości słów - jeden mówca



Rys. 4. Poprawność rozpoznawania w zależności od ilości słów - wielu mówców

Tab. 1. Stopień poprawnie rozpoznanych słów (pod warunkiem, że współczynnik wiarygodności $>= 50\%$) dla jednego mówcy

Słowo	MFC	MFCC	MFCC + gradient	MFCC + energia	MFCC+en.+grad.
"zero"	100	100	100	100	100
"jeden"	91	100	100	100	100
"dwa"	66	66	66	66	66
"trzy"	83	100	100	91	83
"cztery"	100	100	100	100	100
"pięć"	100	100	100	100	100
"sześć"	91	91	83	83	83
"siedem"	91	91	100	91	91
"osiem"	100	100	100	100	100
"dziewięć"	100	100	100	100	100
"start"	50	50	50	50	50
"stop"	83	83	83	91	91
"lewo"	83	91	91	91	91
"prawo"	16	50	25	100	100
"góra"	50	100	100	91	75
"dół"	8	83	83	83	83
"puść"	50	66	75	83	66
"złap"	25	83	66	66	41
"oś"	83	91	91	83	58
"chwytak"	83	83	100	83	91
Średnia wartość	72.65	81.40	85.65	87.60	83.45

Tab. 2. Stopień poprawnie rozpoznanych słów (pod warunkiem, że współczynnik wiarygodności $>= 50\%$) dla sześciu mówców

Słowo	MFC	MFCC	MFCC + gradient	MFCC + energia	MFCC+en.+grad.
"zero"	16	66	66	41	33
"jeden"	25	41	41	41	58
"dwa"	27	72	63	36	36
"trzy"	41	58	58	66	66
"cztery"	50	58	58	50	41
"pięć"	66	66	66	75	83
"sześć"	91	75	66	66	58
"siedem"	91	91	83	83	83
"osiem"	50	66	66	58	50
"dziewięć"	91	75	83	100	91
"start"	66	58	66	58	58
"stop"	46	76	92	92	84
"lewo"	75	50	66	50	66
"prawo"	75	66	66	83	83
"góra"	100	89	89	69	60
"dół"	33	75	75	66	58
"puść"	58	91	100	66	66
"złap"	41	100	91	58	58
"oś"	75	83	83	75	75
"chwytak"	33	50	66	33	33
Średnia wartość	57.50	71.80	72.20	63.30	62

8. PODSUMOWANIE

Przedstawiono analiza sygnału mowy jest sterowana danymi, gdyż nie uwzględnia się w niej żadnej wiedzy o możliwych klasach fonetycznych. Porównano kilka typowych zestawów cech realizując klasyfikację sekwencji ramek o znormalizowanej liczbie ramek dla każdego słowa. Warto zauważyć, że niektóre słowa były rozpoznawane zdecydowanie gorzej od innych (np. "chwytak", i "złap" często identyfikowane były zamiennie). Takie zachowanie klasyfikatora spowodowane jest podobnymi sekwencjami cech, różniącymi się tylko początkiem lub końcem słowa, co w połączeniu z nie dość precyzyjną segmentacją, prowadzi do uzyskiwania niepoprawnych wyników. Te wyniki pracy potwierdzają konieczność stosowania w procesie rozpoznawania mowy jawnych modeli fonetycznych mowy i algorytmów dopasowywania par sekwencji o różnych liczbach ramek [2], [6], [8].

LITERATURA

[1] *The CSLU Speech Toolkit*. CSLU centre, Oregon Graduate Institute, USA, 2000.

[2] S.Grocholewski. *Statystyczne podstawy systemu ARM dla języka polskiego*. Rozprawy Nr. 362, Poznań, Wyd. Politechniki Poznańskiej, 2001.

[3] J.C. Junqua, J.P. Haton. *Robustness in automatic speech recognition*. Kluwer Academic Publications, Boston etc., 1996.

[4] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, J.G. Wilpon. An Improved Endpoint Detector for Isolated Word Recognition. *IEEE Trans. Acoustics, Speech and Signal Proccessing*, Vol. ASSP–29, No. 4, August 1981.

[5] W.Kasprzak (red.). *Detekcja sygnału mowy i cech osobniczych sygnału mowy na potrzeby rozpoznawania komend głosowych*. Raport IAiIS 04–15, Politechnika Warszawska, IAiIS, Warszawa, grudzień 2004.

[6] W.Kasprzak i inni. Zastosowanie MRROC++ do budowy układu sterowania robotem zdolnym do werbalnej komunikacji z człowiekiem. In: *9-ta Krajowa Konferencja Robotyki*, Politechnika Wrocławska, 2006, *w druku*.

[7] M. Piasecki, Sz. Zyśko. *Rozpoznawanie granic słowa w systemie automatycznego rozpoznawania izolowanych słów*. Raport, Wydziałowy Zakład Informatyki Politechniki Wrocławskiej, Wrocław, 1999.

[8] L. Rabiner, B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, New York 1993.

[9] C. Zieliński, W. Szynkiewicz, K. Mianowski, K. Nazarczuk. Mechatronic Design of Open-Structure Multi-Robot Controllers. *Mechatronics*, Vol. 11, No. 8, November 2001. s. 987–1000.

DATA-DRIVEN SPEECH ANALYSIS FOR SPOKEN WORD RECOGNITION

A spoken word recognition system contains both data-driven and model-driven analysis stages. This paper proposes solutions for the data-driven part of it. The following analysis steps are considered: useful speech detection in sound signals, frame-like signal segmentation, feature extraction and phonetic coding of frames. In order to test various feature vectors a feature sequence classifier was implemented and tested.