ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2000
Seria: AUTOMATYKA z. 132 Nr kol. 1477

Marek BINKOWSKI

ROZPOZNAWANIE POLECEŃ GŁOSOWYCH Z WYKORZYSTANIEM TECHNIK TRANSFORMACJI FALKOWEJ

Streszczenie. W pracy opisano przykładowy system rozpoznawania poleceń głosowych, wyposażony w bazę wiedzy zawierającą 21 słów. W systemie sygnał mowy jest dekomponowany za pomocą transformacji falkowej. Poszczególne pasma zdekomponowanego sygnału są poddawane analizie cepstralnej, w wyniku czego ekstrahowane są cechy związane z informacją, niesioną w sygnale mowy. Cechy te są następnie poddawane dwupoziomowej klasyfikacji za pomocą sieci neuronowej typu sieć samoorganizująca się. Skuteczność rozpoznawania w systemie, omówiona na końcu pracy, plasuje się na poziomie 39% (rozpoznawanie pewne) plus 43,5% (rozpoznawanie niepewne). Na końcu pracy zasugerujemy również metody potencjalnego podniesienia skuteczności rozpoznawania w proponowanym systemie.

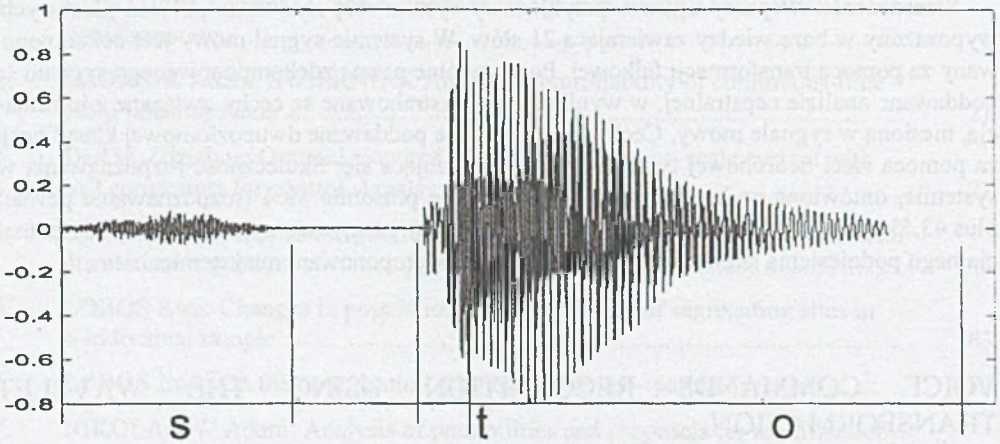
VOICE COMMANDS RECOGNITION USING THE WAVELET TRANSFORMATION

Summary. This work describes an example of voice commands recognition system, equipped with a database with 21 words included. The speech signal in the system is decomposed using Wavelet Transformation. Individual sub-bands of the decomposed signal are then analysed using cepsrtal analysis, and the features related to spoken information are extracted. This features are then classified with a self organizing map neural network. The effectiveness of recognition is about 39% (sure recognition), plus about 43,5% (unsure recognition). Additionally, some potential improvements of the recognition effectiveness are proposed.

1. Wprowadzenie

Transformacja falkowa, stosowana w dziedzinie przetwarzania sygnału, posiada pewną przewagę nad klasyczną transformacją Fouriera oraz krótkoterminową (okienkową) transformacją Fouriera. W odróżnieniu od wymienionych transformacji Fouriera transformacja fal-

kowa pozwala na dokładne pozycjonowanie zdarzeń w sygnale, które mogą mieć różny czas trwania (różną skalę). Jest to niezwykle istotne w przypadku przetwarzania sygnału mowy, który jest sygnałem bardzo zróżnicowanym zarówno na poziomie pojedynczych fonemów, jak i grup fonemów bądź głosek. Pokazano to na przykładzie słowa „sto", którego przebieg jest widoczny na rysunku 1. Pierwsza część przebiegu to w miarę długa, lecz mało energetyczna głoska „s". Następnie, po chwili ciszy widzimy bardzo krótką, lecz bardziej energetyczną głoskę „t". Głoska ta płynnie przechodzi w długą, bardzo energetyczną głoskę „o", która stopniowo wygasa.



Rys.1. Słowo „sto" z zaznaczonymi poszczególnymi głoskami: „s", „t" i „o"
Fig.1. The word „sto" with individual letters „s", „t", and „o" marked

Jak widać, nawet na przykładzie krótkiego, pojedynczego słowa możemy wyróżnić kilka zdarzeń, zróżnicowanych pod względem czasu trwania i niesionej energii. Tak zróżnicowany sygnał możemy z powodzeniem analizować za pomocą transformacji falkowej, która jest obiecującym i coraz chętniej wykorzystywanym narzędziem analizy sygnałów, również w systemach rozpoznawania mowy.

Problem rozpoznawania mowy jest bardzo złożony. Rejestrowany sygnał mowy zmienia się w zależności od wielu czynników, które utrudniają wyłowienie wypowiedzianej informacji z całościowego sygnału bądź też ustalenie jednolitych wzorców poszczególnych fonemów, głosek, czy wyrazów. W rejestrowanym sygnale głosowym najczęściej możemy wyróżnić następujące składowe:

- Rzeczywista wypowiadana informacja. Możemy ją nazwać informacją „tekstową", ponieważ pokrywa się ona z tekstem pisanym, jaki moglibyśmy stworzyć słuchając rejestrowanej wypowiedzi. W przypadku rozpoznawania poleceń głosowych głównym celem jest możliwie dokładne wyekstrahowanie tej informacji z zarejestrowanego sygnału.
- Informacja o mówcy. Słuchając zarejestrowanej wypowiedzi, na podstawie samej barwy głosu (bez jednoczesnego oglądania obrazu mówcy), najczęściej możemy zidentyfikować danego mówcę. Dowodzi to faktu, że sygnał mowy zawiera pewne cechy osobnicze, charakterystyczne dla poszczególnych mówców. Owe cechy osobnicze są wykorzystywane w systemach automatycznej identyfikacji mówców. Jednakże, w przypadku systemów rozpoznawania mowy, informacja ta jest zbędna, ponieważ wprowadza ona do sygnału pewne zróżnicowanie utrudniające rozpoznanie informacji tekstowej.
- Cechy osobnicze są jednym z czynników, utrudniających budowę jednolitego zestawu wzorców, które można by było wykorzystać w systemach rozpoznawania mowy niezależnych od mówcy. Oczywiście, cechy osobnicze nie są przeszkodą w budowaniu systemów rozpoznawania mowy zależnych od mówcy. (W pracy staramy się jednak zbudować system niezależny od mówcy).

  Dodatkowym elementem niepożądanym zarówno w przypadku systemów rozpoznawania mowy, jak i mówcy jest zależność barwy głosu od stanu zdrowia, oraz pokrewnych czynników, takich jak zmęczenie mówcy czy nawet warunki pogodowe.
- Stan emocjonalny. W mowie potocznej sygnał mowy jest silnie zróżnicowany w związku z różnymi stanami emocjonalnymi mówców. Zupełnie inny sygnał rejestrujemy w przypadku, gdy mówca jest spokojny i opanowany, a zupełnie inny w przypadku, gdy jest on na przykład zdenerwowany. Różnice wynikają również z efektu, jaki chce osiągnąć mówca za pomocą danej wypowiedzi [1] (może to być na przykład swobodna wypowiedź, szept, krzyk, powolne bądź szybkie tłumaczenie, recytacja wiersza, czy też gra aktorska).

  Informacja o stanie emocjonalnym mówcy zawarta w sygnale mowy jest ogromnym utrudnieniem w budowaniu uniwersalnych systemów rozpoznawania mowy, ponieważ za jej sprawą rejestrowane sygnały głosowe mogą być w znacznym stopniu zróżnicowane. Jednak w tej pracy nie będziemy się zajmowali problemem wpływu stanu emocjonalnego mówcy na sygnał - zakładamy, że rejestrowane polecenia głosowe, które będą analizowane, są wypowiadane spokojnym tonem, w średnim tempie.

- Warunki otoczenia, zakłócenia sygnału. Rejestrując wypowiedź mówcy nie sposób całkowicie wyeliminować pewnych zakłóceń, zarówno pochodzących z otoczenia, jak również wynikających bezpośrednio z niedoskonałości sprzętu rejestrującego bądź samego formatu zapisu.

  Zakłócenia otoczenia to przede wszystkim towarzyszące dźwięki, takie jak hałas ulicy, rozmowy innych osób, czy ogólny szum w pomieszczeniach. Z kolei, zakłócenia wynikające z niedoskonałości sprzętu to ograniczone pasmo przenoszenia mikrofonu, niedoskonałość przetworników i szum cieplny elementów elektronicznych. W przypadku zapisu sygnału w postaci cyfrowej dodatkowym zakłóceniem jest szum kwantyzacji. W pracy nie zbadano wpływu poszczególnych typów zakłóceń na uzyskiwane rezultaty - wszystkie polecenia głosowe były rejestrowane w podobnych warunkach pokojowych.

2. Założenia

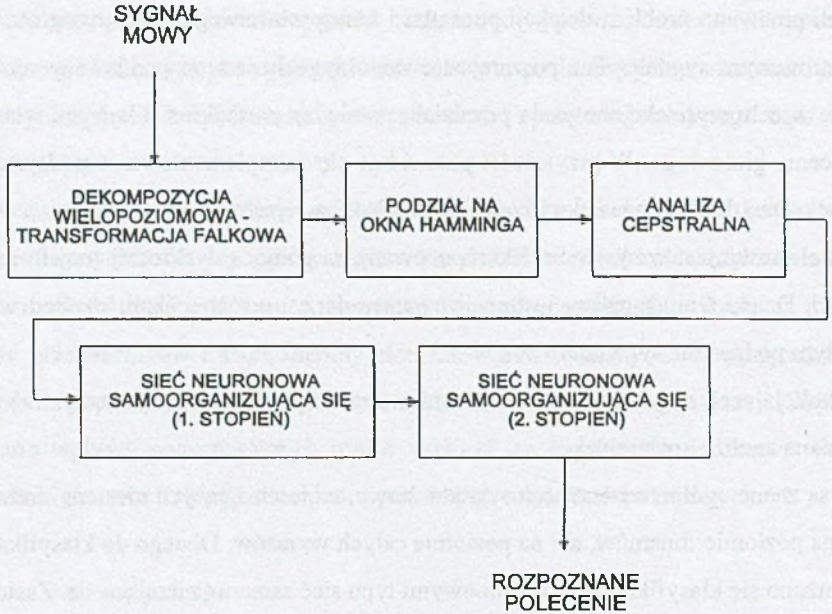
Przygotowując się do budowy systemu rozpoznawania poleceń głosowych, poczyniono następujące założenia:

- System jest przygotowany do rozpoznawania poleceń głosowych o ustalonej treści. Zaproponowano bazę dwudziestu jeden słów o następującej treści: „góra", „dół", „prawo", „lewo", „przód", „tył", „środek", „obrót", „przesuń", „jeden", „dwa", „pięć", „dziesięć", „piętnaście", „dwadzieścia", „trzydzieści", „czterdzieści pięć", „pięćdziesiąt", „sześćdziesiąt", „dziewięćdziesiąt", „sto". (Taki zestaw poleceń głosowych może na przykład służyć do sterowania położeniem ramienia robota za pomocą głosu).
- Baza poleceń głosowych została przygotowana z udziałem czterech mówców (dwóch kobiet i dwóch mężczyzn). Każdy z mówców wypowiedział każde polecenie jednokrotnie, spokojnym tonem w średnim tempie.
- Wszystkie polecenia głosowe były rejestrowane w podobnych warunkach (w cichym w miarę pomieszczeniu, jednak nie wyciszanym specjalnie na potrzeby rejestracji dźwięków). Do rejestracji użyto średniej klasy mikrofonu amatorskiego. Zarejestrowany dźwięk był zapisywany bezpośrednio w postaci cyfrowej. Docelowy format cyfrowy to monofoniczny sygnał 16-bitowy o częstotliwości próbkowania 16 kHz.

- Wyeliminowano problem detekcji początku i końca właściwego polecenia głosowego w rejestrowanym sygnale. Tuż po zarejestrowaniu sygnału jest on poddawany ręcznej obróbce w celu wyselekcjonowania przedziału pomiędzy początkiem i końcem właściwego polecenia głosowego. W przyszłości przewiduje się zaimplementowanie podsystemu automatycznej detekcji początku i końca wypowiedzi w rejestrowanym sygnale.
- Sygnał mowy jest w systemie dekomponowany za pomocą dyskretnej transformacji falkowej. Dzięki temu możliwe jest analizowanie zdarzeń o różnej skali (czasie trwania) w każdym podpaśmie sygnału.
- Ekstrakcja cech sygnału mowy jest dokonywana za pomocą analizy cepstralnej, będącej odmianą analizy spektralnej.
- Nie są znane ogólne wzorce cech sygnału mowy, odpowiadających niesionej informacji – ani na poziomie fonemów, ani na poziomie całych wyrazów. Dlatego do klasyfikacji cech posłużono się klasyfikatorami neuronowymi typu sieć samoorganizująca się. Zastosowano klasyfikację dwupoziomową.
- System po przeprowadzeniu procesu rozpoznawania polecenia głosowego udziela jednej z trzech odpowiedzi:
  - Tekst rozpoznanego polecenia.
  - Tekst dwóch lub większej liczby poleceń wraz z prawdopodobieństwem wystąpienia każdego z nich (w przypadku rozpoznania niepewnego).
  - Brak rozpoznania wypowiedzianego polecenia.

3. Narzędzia analizy sygnału

Na rysunku 2 pokazano ogólny schemat blokowy opracowanego systemu. (Szczegółowy schemat pokazano na rysunku 7). Omówimy teraz główne narzędzia analizy i klasyfikacji sygnału zastosowane w systemie: transformację falkową, analizę cepstralną oraz sieć neuronową typu sieć samoorganizująca się.



Rys.2. Ogólny schemat blokowy opracowanego systemu rozpoznawania poleceń głosowych
Fig.2. General block diagram of the created voice commands recognition system

3.1. Transformacja falkowa

Szczegółowe omówienie zagadnień związanych z analizą falkową wychodzi poza ramy tej pracy. Więcej informacji na ten temat można znaleźć w [2].

Ciągła transformacja falkowa polega na rzutowaniu analizowanego sygnału na rodzinę falek $\varphi(a,b)$, gdzie a jest skalą falki, zaś b jej przesunięciem względem zera

$$C(a,b) = \int_{-\infty}^{\infty} f(t)\varphi(a,b,t)dt \tag{1}$$

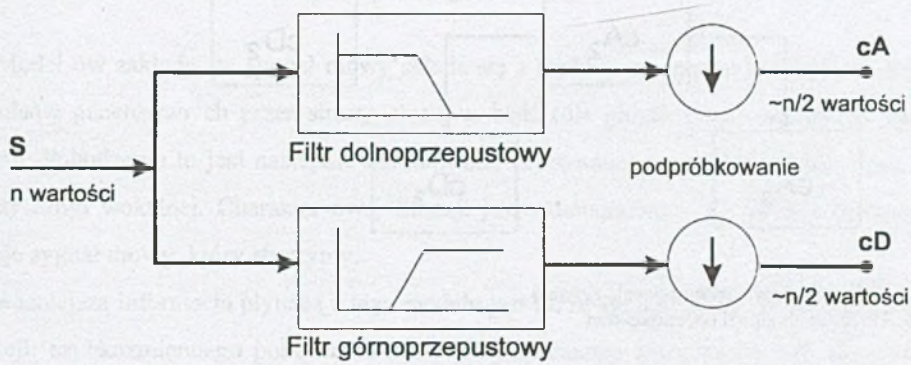
W wyniku analizy falkowej sygnału otrzymujemy rodzinę *współczynników falek* $C(a,b)$, będących funkcją skali a i przesunięcia b. Przemnożenie każdego współczynnika przez odpowiadającą mu falkę (przeskalowaną przez parametr a i przesuniętą o parametr b), a następnie sumowanie otrzymanych sygnałów pozwala na odtworzenie oryginalnego sygnału $f(t)$, występującego we wzorze (1).

Implementacja dyskretnej transformacji falkowej polega na zastosowaniu pary komplementarnych filtrów będących odpowiednikami falek w dyskretnej dziedzinie czasu. Oznaczmy analizowany sygnał cyfrowy przez $s(n)$, gdzie $n = 1..N$ to numer kolejnej próbki sygnału,

zaś N to liczba próbek. Niech symbol $g_{j,k}$ reprezentuje filtr będący dyskretnym odpowiednikiem falki. Wzór transformaty falkowej (1) możemy zmodyfikować do postaci dyskretnej:

$$C(a,b) = C(j,k) = \sum_{n \in Z} s(n) g_{j,k}(n) \tag{2}$$

Jednopoziomowa dekompozycja sygnału za pomocą transformaty falkowej polega na zdekomponowaniu go na dwa sygnały: *detale D* oraz *aproksymacje A*. Pierwotny sygnał S przechodzi przez dwa komplementarne filtry, w wyniku czego otrzymujemy dwa sygnały. Operacja filtracji została przedstawiona na rysunku 3.



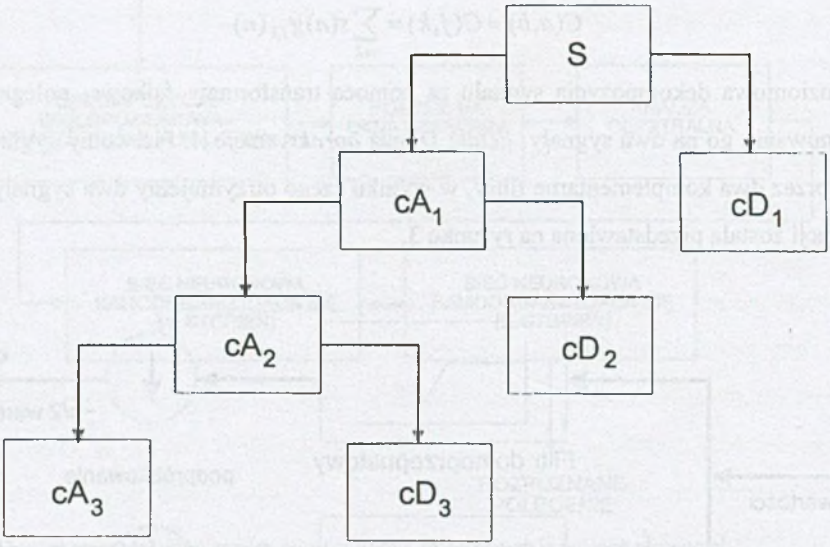
Rys.3. Dekompozycja sygnału na współczynniki aproksymacji *cA* i detali *cD* za pomocą filtrów komplementarnych

Fig.3. Decomposition of the signal with complementary filters, giving the approximation coefficients *cA* and details coefficients cD

Na rysunku widoczne są symbole strzałki skierowanej w dół - reprezentują one operację podpróbkowania, czyli redukowania liczby próbek wchodzących w skład sygnału. Na wyjściu otrzymujemy dwa sygnały oznaczone jako *cA* i *cD*. Ich elementy nazywamy odpowiednio *współczynnikami aproksymacji* i *współczynnikami detali*.

Przeprowadza się również dekompozycję sygnału na większą liczbę składowych – jest to tak zwana *dekompozycja wielopoziomowa*. Proces dekompozycji wielopoziomowej polega na powtórnym dekomponowaniu sygnału *cA* składającego się ze współczynników aproksymacji. Przebieg *cA* reprezentuje niższe częstotliwości sygnału pierwotnego. Jego dekompozycja daje w wyniku kolejne rozseparowanie składowej sygnału na składowe o wyższych i niższych częstotliwościach. Wyższe częstotliwości są określane mianem *współczynników detali drugiego poziomu* i oznaczane jako cD_2. Niższe częstotliwości, czyli *współczynniki aproksymacji drugiego poziomu* są oznaczane jako cA_2. Ich dalsza dekompozycja daje współczynniki wyż-

szych poziomów. Na rysunku (4) przedstawiono schematycznie dekompozycję trójpoziomową.

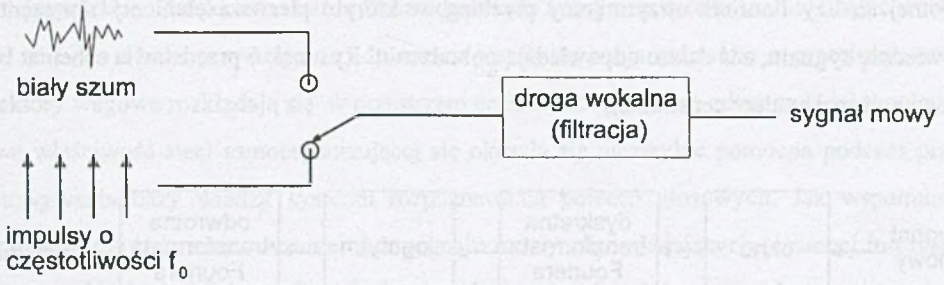


Rys.4. Trójpoziomowa dekompozycja sygnału
Fig.4. Three-levels signal decomposition

Przedstawiona metoda wielopoziomowej dekompozycji sygnału została zastosowana w opracowanym systemie rozpoznawania poleceń głosowych. Zastosowano tam pięciopoziomową dekompozycję sygnału mowy. W jej wyniku otrzymywanych jest sześć sygnałów – pięć poziomów detali oraz aproksymacje piątego poziomu.

4. Analiza cepstralna

Analiza cepstralna jest zmodyfikowaną postacią analizy spektralnej, a dokładniej krótkoterminowej analizy Fouriera [4], [5]. (Samo słowo *cepstrum* pochodzi od zmodyfikowanego słówka angielskiego *spectrum*). Pomysł zastosowania zmodyfikowanej analizy spektralnej w procesie rozpoznawania mowy wynika z uproszczonego modelu funkcjonowania narządu mowy. Model ten przedstawiono na rysunku 5.



Rys.5. Uproszczony model funkcjonowania narządu mowy
Fig.5. The simplified functional model of the speech organ

Model ów zakłada, że sygnał mowy składa się z szybko zmiennego pobudzenia, czyli z impulsów generowanych przez struny głosowe, bądź (dla głosek bezdźwięcznych) białego szumu. Pobudzenie to jest następnie modulowane (filtrowane) przez jamę ustną i inne elementy drogi wokalnej. Charakter owej filtracji jest wolnozmienny. W wyniku filtracji powstaje sygnał mowy, który słyszymy.

Najważniejszą informacją płynącą z tego modelu jest to, że sygnał mowy jest splotem dwóch funkcji: szybkozmiennego pobudzenia oraz wolnozmiennego filtra. Oznaczmy sygnał jako $S(i)$, szybkozmienne pobudzenie jako $E(i)$ oraz filtrację jako $H(i)$. Wówczas możemy zapisać:

$$S(i) = E(i) \cdot H(i) \tag{3}$$

W dziedzinie częstotliwości (po przeprowadzeniu transformaty Fouriera) równanie (3) przyjmie postać:

$$S(e^{j\omega}) = E(e^{j\omega})H(e^{j\omega}) \tag{4}$$

Znana jest nam funkcja $S(e^{j\omega})$, zaś chcielibyśmy poznać reprezentację funkcji $H(e^{j\omega})$, czyli obwiednię sygnału, oraz $E(e^{j\omega})$ – pobudzenie. Ich przybliżenie możemy osiągnąć poprzez zlogarytmowanie spektrum sygnału:

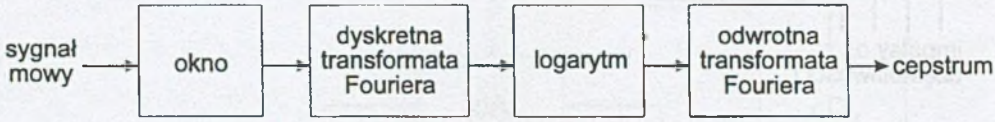
$$\log(S(e^{j\omega})) = \log(E(e^{j\omega})) + \log(H(e^{j\omega})) \tag{5}$$

Jako że w analizie sygnału mowy zwykle pomijana jest informacja o fazie, a brana pod uwagę jedynie informacja o amplitudzie spektrum, możemy również zapisać:

$$\log(|S(e^{j\omega})|) = \log(|E(e^{j\omega})|) + \log(|H(e^{j\omega})|) \tag{6}$$

Wolno zmieniające się komponenty są reprezentowane przez wyraz $\log(|H(e^{j\omega})|)$, z kolei elementom o wyższej częstotliwości odpowiada wyraz $\log(|E(e^{j\omega})|)$. Zlogarytmowanie sygnału pozwala więc na rozdzielenie go na sumę sygnałów składowych. Po przeprowadzeniu od-

wrotnej analizy Fouriera otrzymujemy przebieg, w którym pierwsze elementy reprezentują obwiednię sygnału, zaś dalsze odpowiadają pobudzeniu. Rysunek 6 przedstawia schemat blokowy opisanej analizy cepstralnej.



Rys.6. Schemat blokowy analizy cepstralnej
Fig.6. Cepstral analysis block diagram

W opracowanym systemie rozpoznawania mowy analiza cepstralna posłużyła do ekstrakcji cech związanych z informacją niesioną w sygnale mowy. Analizie tej zostało poddane każde z sześciu pasm zdekomponowanego sygnału (po uprzednim podzieleniu sygnału pasma za pomocą okna Hamminga). W wyniku przeprowadzonej analizy każde okno w każdym paśmie zdekomponowanego sygnału jest reprezentowanych przez 15 współczynników cepstrum.

5. Sieć samoorganizująca się

Szczegółowe omówienie zasad działania sieci neuronowej typu sieć samoorganizująca się wychodzi poza zakres tej pracy. Z zasadami działania tej i innych typów sieci neuronowych można się zapoznać w bogato dostępnej literaturze, między innymi w [3]. W pracy opisujemy jedynie podstawowe właściwości tej sieci, istotne dla opracowanego systemu rozpoznawania poleceń głosowych.

Nazwa sieci samoorganizującej się wynika z faktu, że sieć ta potrafi w trakcie procesu nauczania, bez pomocy nauczyciela, dobrać swoje współczynniki wagowe w ten sposób, że zbiór wektorów wejściowych zostanie podzielony na klasy. Po podziale w każdej klasie znajdują się wektory o podobnych cechach. Każdy nowy wektor, zaprezentowany sieci po procesie nauczania, zostanie przydzielony do jednej z automatycznie stworzonych klas.

Rozmieszczenie poszczególnych klas odpowiada gęstości wektorów w przestrzeni cech. Znaczy to, że jeśli w pewnym obszarze przestrzeni cech występuje duże zagęszczenie wektorów cech, wówczas obszar ten będzie reprezentowany nie przez jedną dużą, lecz przez kilka

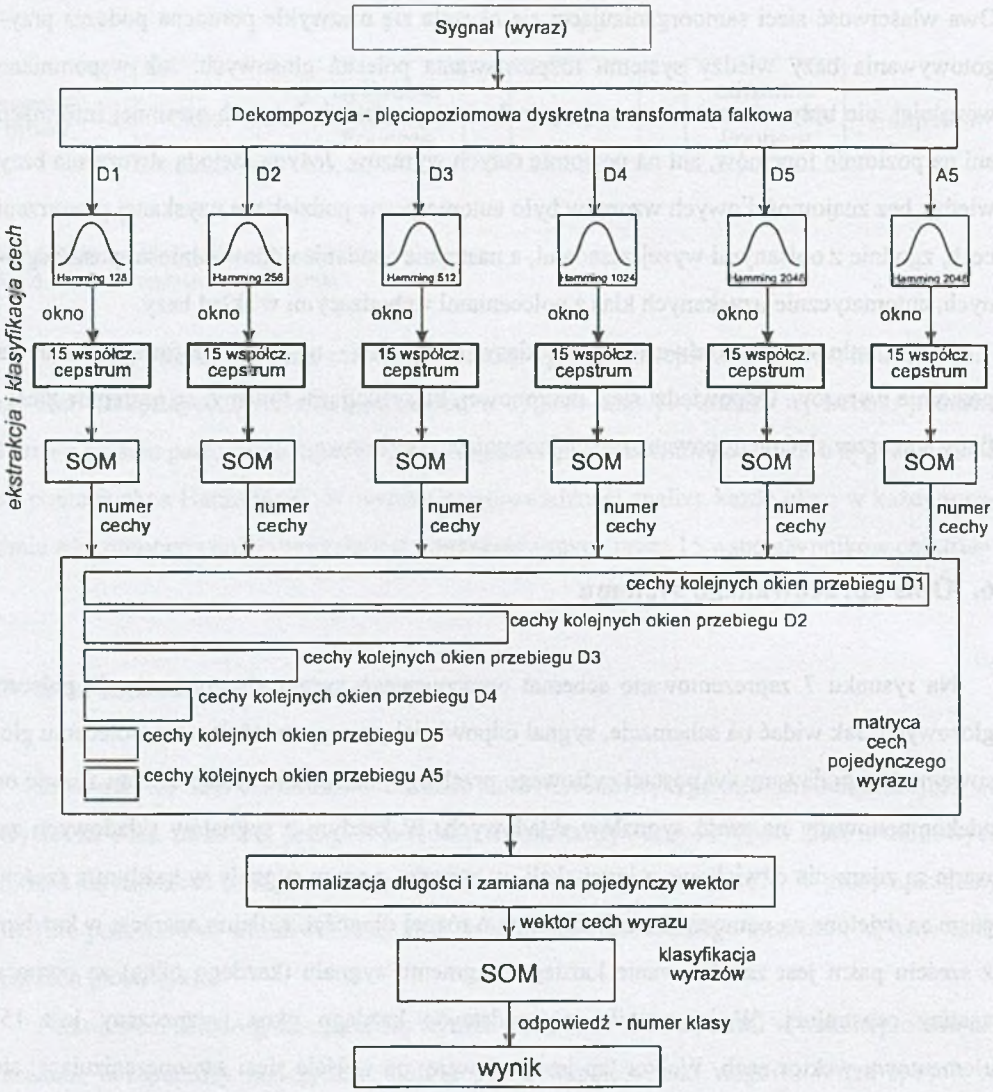
mniejszych klas. Dzięki temu każda klasa odpowiada mniej więcej takiej samej liczbie wektorów cech, co z kolei powoduje, że odtwarzana jest dystrybucja wektorów w przestrzeni cech. Wektory wagowe rozkładają się w przestrzeni cech w ten sposób, że odtwarzają jej topologię. Owa właściwość sieci samoorganizującej się okazała się niezwykle pomocna podczas przygotowywania bazy wiedzy systemu rozpoznawania poleceń głosowych. Jak wspomniano wcześniej, nie były znane wzorce cech sygnału mowy odpowiadających niesionej informacji ani na poziomie fonemów, ani na poziomie całych wyrazów. Jedyną metodą stworzenia bazy wiedzy bez znajomości owych wzorców było automatyczne podzielenie uzyskanej przestrzeni cech, zgodnie z opisanymi wyżej zasadami, a następnie zbadanie odpowiedniości poszczególnych, automatycznie uzyskanych klas z poleceniami wchodzącymi w skład bazy.

W systemie stworzono dwustopniową klasyfikację cech – na poziomie fonemów oraz na poziomie wyrazów. Odpowiedzi sieci neuronowej, klasyfikującej fonemy, są następnie klasyfikowane przez sieć neuronową rozpoznającą polecenia głosowe.

6. Opis opracowanego systemu

Na rysunku 7 zaprezentowano schemat opracowanego systemu rozpoznawania poleceń głosowych. Jak widać na schemacie, sygnał odpowiadający wypowiedzianemu poleceniu głosowemu jest podawany (w postaci cyfrowego przebiegu) na wejście systemu. Tam zostaje on zdekomponowany na sześć sygnałów składowych. W każdym z sygnałów składowych zawarte są zdarzenia dźwiękowe o innej skali, w związku z czym sygnały w każdym z sześciu pasm są dzielone za pomocą okien Hamminga o różnej długości. Kolejną operacją w każdym z sześciu pasm jest zanalizowanie każdego fragmentu sygnału (każdego okna) za pomocą analizy cepstralnej. W jej wyniku na podstawie każdego okna wyznaczany jest 15-elementowy wektor cech. Wektor ten jest podawany na wejście sieci samoorganizującej się (SOM), klasyfikującej wektory cech. Rozmiary okien i wektorów cech są dobrane w ten sposób, że na tym etapie klasyfikowane są cechy fonemów analizowanego sygnału mowy. Dla każdego z sześciu pasm jest przygotowana oddzielna sieć neuronowa SOM. Odpowiedzią każdej z tych sieci jest pojedyncza liczba będąca wynikiem zaklasyfikowania 15-elementowego wektora cech do jednej z 42 klas. W tym miejscu kończy się pierwszy poziom

dwupoziomowej klasyfikacji (oznaczonej na schemacie jako „ekstrakcja i klasyfikacja cech").



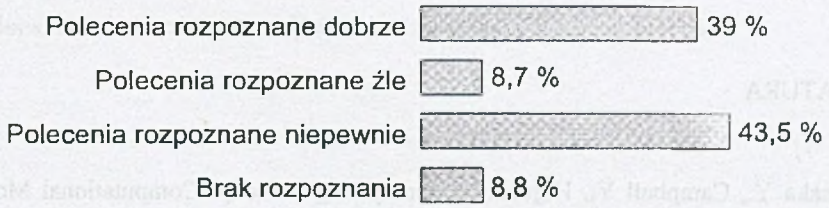
Rys.7. Schemat systemu rozpoznawania poleceń głosowych
Fig.7. The diagram of the created voice commands recognition system

Następnie w pojedynczej, dużej matrycy cech zostają zebrane wszystkie odpowiedzi udzielone przez wszystkie sieci neuronowe SOM pierwszego poziomu, dla całego aktualnego polecenia głosowego. Odpowiedzi te, to cechy drugiego poziomu klasyfikacji, odpowiadają-

cego wyrazowi (bądź w szczególnych przypadkach – kilku wyrazom). Cechy te zostają odpowiednio uporządkowane w postaci pojedynczego wektora cech dużych rozmiarów. Długość tego wektora zostaje znormalizowana. Znormalizowany wektor cech jest podawany na sieć samoorganizującą (SOM), która dokonuje ostatecznej klasyfikacji wypowiedzianego wyrazu (polecenia głosowego).

7. Wyniki i wnioski

W celu sprawdzenia skuteczności działania systemu przeprowadzono testy, polegające na wypowiadaniu kolejnych słów z bazy wiedzy i porównywaniu otrzymywanych odpowiedzi z rzeczywistością. Testy zostały przeprowadzone z tymi samymi mówcami, którzy brali udział w przygotowywaniu bazy wiedzy. Wyniki przedstawiono na rysunku 8.



Rys.8. Wyniki rozpoznawania poleceń głosowych
Fig.8. The results of speech recognition tests

Wyniki podzielono na cztery kategorie: wyrazy rozpoznane bezbłędnie, wyrazy rozpoznane błędnie (odpowiedź systemu inna niż wypowiedziane słowo), rozpoznanie niepewne (wypowiedziany wyraz znalazł się wśród zaproponowanych przez system odpowiedzi) oraz brak rozpoznania.

Otrzymane wyniki rozpoznawania wskazują na to, że zaprojektowany system w obecnej konfiguracji nie działa wystarczająco niezawodnie, aby mógł być stosowany w praktyce. Mimo zamiarów stworzenia systemu niezależnego od mówcy opracowany system nie jest w pełni niezależny – najlepiej rozpoznaje on polecenia wypowiadane przez osoby, których głosy były wykorzystywane podczas tworzenia bazy wiedzy. (Polecenia głosowe wypowiadane przez osoby spoza bazy również są rozpoznawane, lecz z mniejszą skutecznością). Stworzenie

modelu, który działałby niezależnie od mówcy, prawdopodobnie nie wymaga żadnych modyfikacji samego modelu. Wymagane jest raczej wykorzystanie znacznie większej liczby mówców w procesie przygotowywania bazy wiedzy (co wiąże się ze znacznie większym nakładem obliczeniowym i pamięciowym). Zwiększenie liczby mówców nie miałoby wpływu na wydłużenie czasu obliczeń samej procedury rozpoznawania, jedynie na czas przygotowywania bazy. Niewątpliwie zaś wpłynęłoby to na znaczne polepszenie jej skuteczności, bowiem przy większej liczbie wzorców sieć neuronowa ma możliwość lepszego podziału przestrzeni cech na klasy.

Prawdopodobnie pewną poprawę skuteczności można również uzyskać, wprowadzając modyfikacje samego algorytmu (dobór szerokości okien, zmiana liczby poziomów dekompozycji czy też liczby współczynników cepstrum, ewentualnie wybór innej metody ekstrakcji cech). Otrzymane wyniki rozpoznawania świadczą o tym, że zaproponowany system, choć nie jest doskonały, opiera się na słusznych założeniach. Wymaga jednak prowadzenia dalszych prac i modyfikacji.

LITERATURA

1. Sagisaka Y., Campbell Y., Higuchi N.: Computing Prosody: Computational Models for Processing Spontaneous Speech.
2. Misiti M., Misiti Y., Oppenheim G., Poggi J.-M.: Wavelet Toolbox User's Guide. The Math Works, Inc., 1996.
3. Demuth H., Beale M.: Neural Network Toolbox User's Guide. The Math Works, Inc., 1992.
4. Beauchamp K.G.: Przetwarzanie sygnałów metodami analogowymi i cyfrowymi. George Allen & Unwin Ltd., 1973. Tłumaczenie: Witold Włodarski, Jacek Koronacki, Wydawnictwa Naukowo-Techniczne, Warszawa 1978.
5. Izydorczyk J., Płonka G., Tyma G.: Teoria Sygnałów: Wstęp. Wydawnictwo Helion, 1999.

Recenzent: Doc.dr inż. Bohdan Wołczak

Wpłynęło do Redakcji 5 kwietnia 2000 r.

Abstract

This work describes an created example of voice commands recognition system, equipped with a database including 21 spoken words. Following topics are mentioned:

- A general review of speech signal, digital sampling of the signal, and the problems of information hidden in recorded human voice.
- Wavelet Analysis – the relative new discipline of signal analysis, with the rising fields of applications in many spheres of science and technology. In the created model, Wavelet Analysis serves as a tool for initial processing and decomposing the voice signal.
- Cepstral Analysis – a particular variety of spectral analysis, prepared especially for voice signal. In the created model, used for initial feature extraction.
- Neural Network – in the created model, the self-organizing map is used to carry out a two-stage feature classification of decomposed voice signal.

Although the results of testing of the created model are not satisfactory yet, there are a few suggestions included on how the efficiency of the system can be improved.