Przetwarzanie mowy w celu sterowania urządzeniami mechatronicznymi

Speech processing work for mechatronics control

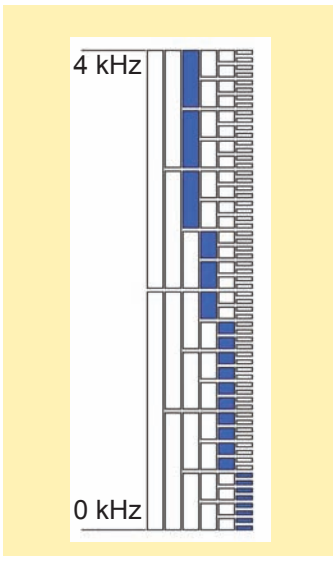
MIROSŁAW TARASIUK
ZDZISŁAW GOSIEWSKI*

DOI: 10.17814/mechanik.2015.7.329

Przedstawiono etapy opracowania metody parametryzacji sygnałów mowy. **Adaptowano dekompozycję paczkowej transformacji falkowej oraz zastosowano rozplot homomorficzny. Dzięki wykorzystaniu niejawnych modeli Markowa do rozpoznawania zweryfikowano działanie opracowanej metody. Badania stanowią punkt wyjścia do wdrożenia automatycznego systemu rozpoznawania mowy do sterowania urządzeniami mechatronicznymi.**
SŁOWA KLUCZOWE: transformacja falkowa, analiza cepstralna, automatyczne rozpoznawanie mowy

Illustrated are the steps to develop a method of speech parameterization. Adapted for the purpose was packet decomposition of the wavelet transformation with homomorphic deconvolution also applied. The hidden Markov Models for speech recognition ***as used were providing at the same time for verification of the*** *developed method. These studies should be considered as the starting point for further implementation of an automatic speech recognition system for control of mechatronic devices.*
KEYWORDS: wavelet transformation, cepstral analysis, automatic speech recognition

Sygnał mowy po wydzieleniu go z otaczającej ciszy [2] wymaga parametryzacji, ponieważ cechuje go duża nadmiarowość informacji. W tym celu najczęściej wykorzystywano dwie metody [3]: predykcji liniowej lub współczynników cepstralnych w częstotliwościowej skali mel (z ang. MFCC). Kolejnym krokiem w procesie rozpoznawania mowy jest jej modelowanie z zastosowaniem ukrytych modeli Markowa (z ang. HMM), nieliniowego dopasowania czasowego lub sieci neuronowych [4].



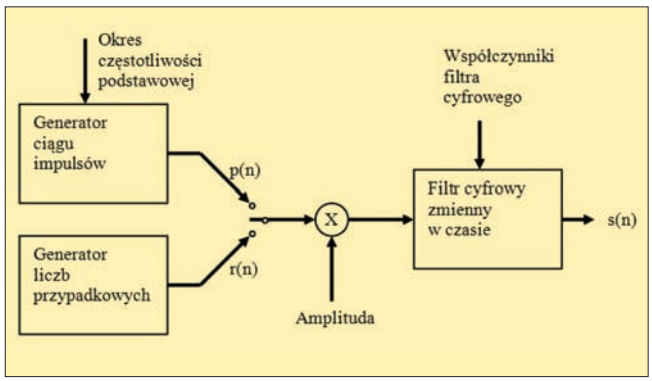
Rys. 1. Sześciopoziomowa WPT sygnału z wyróżnionym rozkładem pasm częstotliwości zbliżonym do MFCC

W celu przeprowadzenia transformacji falkowej (z ang. WT*)* sygnału używa się dwóch filtrów: górnoprzepustowego i dolnoprzepustowego. Kolejne operacje dekompozycji prowadzone są w stosunku do części dolnoprzepustowej. Każdej operacji towarzyszy zmniejszenie liczby współczynników falkowych o połowę (decymacja) w stosunku do analizowanego sygnału wejściowego z poprzedniego poziomu dekompozycji. Jeżeli proces dekompozycji prowadzony jest zarówno z częścią dolnoprzepustową, jak i górnoprzepustową, to taką transformację określa się jako paczkową transformację falkową (z ang. WPT). Otrzymuje się wówczas równomierny podział pasm częstotliwości (rys. 1). Żadna z dwóch dekompozycji nie odpowiada charakterystyce traktu głosowego człowieka.

Parametryzacja sygnału mowy

Parametryzację sygnału przeprowadza się przez podzielenie go na krótkie fragmenty, wykorzystując ich quasi-stacjonarność [7]. W ten sposób odwzorowuje się ludzkie ucho, które nie reaguje na krótsze zmiany w sygnale akustycznym. Otrzymane stacjonarne ramki obserwacji można zastąpić wyznaczonym z każdej z nich wektorem cech. Do podziału sygnału stosuje się okno Hamminga z zakładkowaniem, aby nie wprowadzać zakłóceń w charakterystykach częstotliwościowych.

W trakcie mówienia człowiek pobudza kanał głosowy (nazywany też torem akustycznym) – od głośni do warg. Sygnał akustyczny wytworzony przez struny głosowe podlega zmianom w trakcie przechodzenia przez tor akustyczny. Może on być określony (gdy rozpatruje się odcinki quasi-stacjonarne) modelem liniowym [9], w którym sygnał wyjściowy jest opisany jako splot pobudzenia i odpowiedzi impulsowej toru (rys. 2).



Rys. 2. Model generacji sygnału mowy

Przyjęty model pozwala wykorzystać rozplatanie homomorficzne krótkich segmentów mowy do rozdzielenia charakterystyki toru od pobudzenia (analiza cepstralna) i wykorzystać ją do rozpoznawania mowy. W celu rozdzielenia składników należy dokonać przejścia od ich multiplikatywności do addytywności.

MFCC [1] jest najdłużej rozwijaną metodą parametryzacji mowy człowieka. Widmo sygnału uzyskane przez szybką transformatę Fouriera jest filtrowane przez bank filtrów o szerokości 300 meli, przesuniętych względem siebie o 150 meli, obejmujących całe pasmo częstotliwości sygnału. Zestaw filtrów naśladuje charakterystykę układu słuchu człowieka.

Sarikaya i in. [5] do rozpoznawania mówców zastosowali sześciopoziomową WPT – przez odpowiedni dobór drzewa dekompozycji uzyskali zakres częstotliwości odpowiadający bankowi filtrów z metody MFCC. Zaciemnione prostokąty (rys. 1) obrazują użytą część drzewa dekompozycji. Tak otrzymuje się wektor opisujący chwilowy stan sygnału mowy, określony w przestrzeni 24 cech.

* Mgr inż. Mirosław Tarasiuk (miroslaw.tarasiuk@policja.gov.pl) – CBŚP; prof. dr hab. inż. Zdzisław Gosiewski (z.gosiewski@pb.edu.pl) – Katedra Automatyki i Robotyki Politechniki Białostockiej

Dalej wykorzystano ten sposób dekompozycji do rozpoznawania mowy człowieka. Posłużono się czterema poziomami dekompozycji o różnej liczbie współczynników falkowych. Aby wyznaczyć energię sygnału w każdym z wydzielanych pasm, można obliczyć uśrednioną energię pasma – za pomocą wzoru (1) – lub całkowitą energię pasma – za pomocą wzoru (2):

$$\overline{D}_k = \frac{1}{N_k}\sum_{n=1}^{N_k} d_{n,k}^2, \quad \text{dla } k = 1, \ldots 24 \tag{1}$$

$$D_k = \sum_{n=1}^{N_k} d_{n,k}^2, \quad \text{dla } k = 1, \ldots 24 \tag{2}$$

gdzie: N_k – liczba współczynników falkowych w k-filtrze, $d_{n,k}$ – współczynniki falkowe w paśmie przepuszczania k-filtra.

Jeżeli całkowitą energię wektora zdefiniuje się następująco:

$$D = \sum_{k=1}^{24} D_k \tag{3}$$

wtedy znormalizowany wektor cech sygnału, uwzględniający energię całkowitą, jest opisany wzorem:

$$F = \left[\frac{D_1}{D}, \quad \frac{D_2}{D}, \quad \ldots, \quad \frac{D_{24}}{D}\right] \tag{4}$$

Porównano, o ile wartość energii wyznaczonej w zaznaczonych pasmach (rys. 1) różni się od energii wyznaczonej przez sumowanie energii z odpowiadających im pasm szóstego poziomu dekompozycji. Różnica wynosi 10^{-16}, a więc jest pomijalna – tym samym można wykorzystać komendę *wenergy* programu Matlab.

Badania

Głosy kobiety i mężczyzny różnią się pod względem charakterystyk częstotliwościowych. Badano więc słowa wypowiadane przez kobietę i mężczyznę, rejestrowane dwoma urządzeniami: dyktafonem cyfrowym Sony (model ICD-P28 o częstotliwości próbkowania 8 kHz) i wielokanałowym modułem akwizycji danych VibDAQ 4+ firmy EC Electronics (o częstotliwości próbkowania 105 kHz) z dołączonym mikrofonem pomiarowym Bruel & Kjaer, przy czym używając pakietu Adobe Audition, zmieniono częstotliwości próbkowania na 8 kHz.

Rozpoznawano słowa pochodzące z ograniczonego słownika 20 komend, a każdą rejestrowano – wykonano 20 powtórzeń w grupie uczącej i pięć powtórzeń w grupie badawczej. W słowniku znajdowały się polskie liczebniki (od 1 do 10) i nazwy kolorów (amarant, biały, brąz, czarny, czerwony, fiolet, niebieski, turkus, zielony i żółty).

Do rozpoznawania pojedynczych słów z ograniczonego słownika z wykorzystaniem niejawnych modeli Markowa zastosowano model systemu działający w oparciu o schemat blokowy (rys. 3) [11].

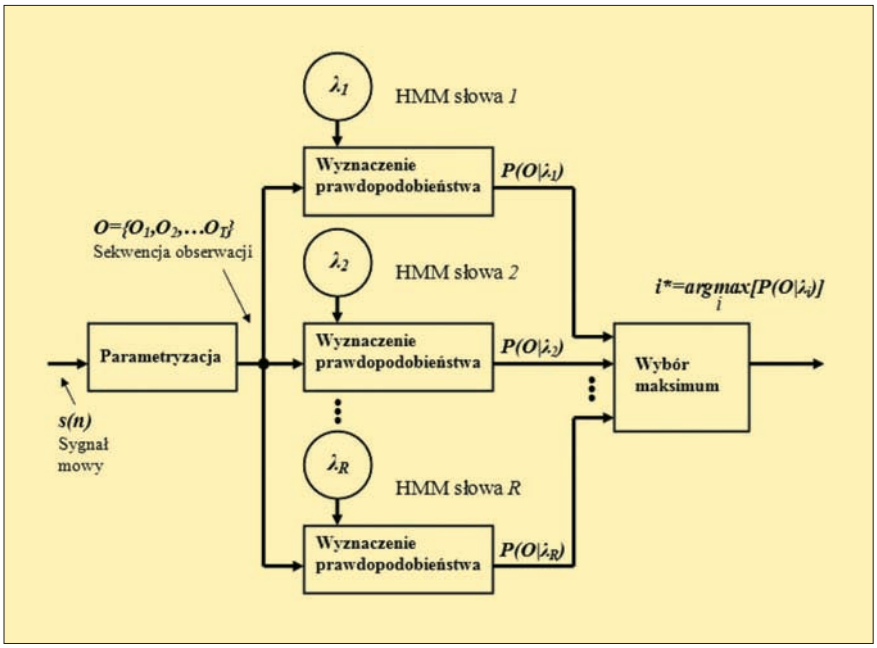
Do porównania własnych rezultatów z wynikami innych badaczy wykorzystano oprogramowanie udostępnione przez Uniwersytet w Kopenhadze [6]. W programie zmodyfikowano długość okna, obszar nakładkowania i częstotliwość próbkowania. Za pomocą oprogramowania z zaimplementowanymi algorytmami HMM (model pięciostanowy), MFCC i k-średnich (z maksymalną liczbą centroidów wynoszącą 16) osiągnięto wyniki rozpoznawania zaprezentowane w tablicy.

TABLICA. Wyniki rozpoznawania słów rejestrowanych dwoma urządzeniami (dyktafonem i rejestratorem) dla mężczyzn (m) i kobiet (k)

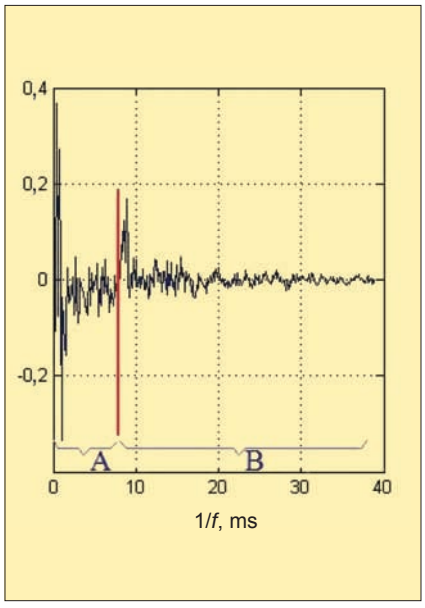
Płeć_urządzenie	Wynik, %
m_dyktafon	93
k_dyktafon	93
m_rejestrator	92
k_rejestrator	96

Zastosowano segmentację równomierną, korzystając z ramki o długości 32 ms, z zachodzeniem ramek na 10 ms (co spełnia warunek, że wielkość zachodzenia wynosi przynajmniej 1/3 długości ramki [9]). Po wyznaczeniu cepstrum rzeczywistego i zastosowaniu filtra dolnoprzepustowego wydzielono charakterystykę traktu głosowego (oznaczonego jako *A*) od wymuszenia (*B*) (rys. 4).

Z przeprowadzonych wcześniej badań [8] wynikało, że w przypadku wykorzystania WPT w parametryzacji sygnału mowy właściwe jest użycie sześciopoziomowej dekompozycji falek db12. Adaptowano oprogramowanie wykorzystane do wyznaczenia wyników porównawczych w celu łącznego zastosowania analizy cepstralnej i WPT z zaproponowanym rozkładem pasm (rys. 1).

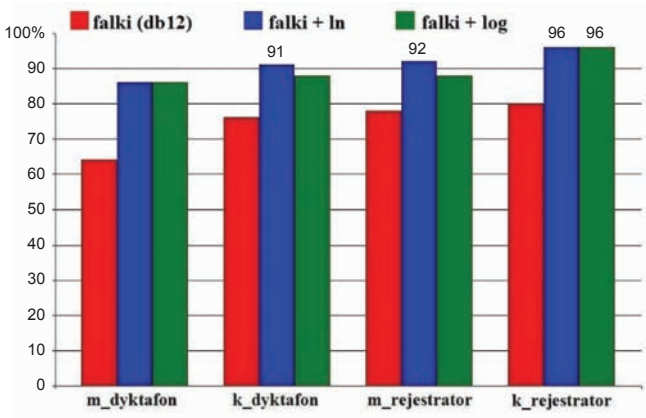


Rys. 3. Schemat blokowy wykorzystania HMM do rozpoznawania słów z ograniczonego słownika



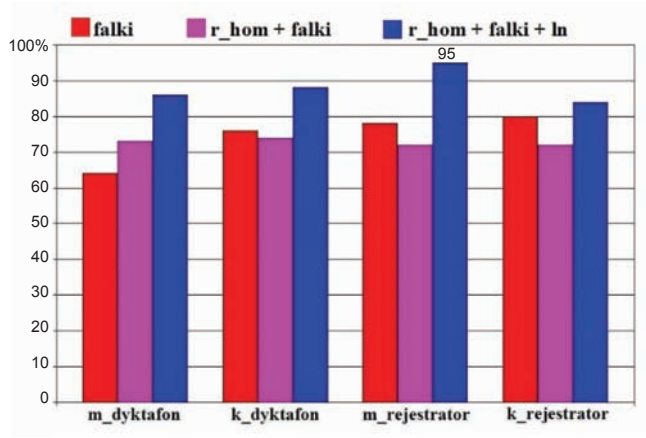
Rys. 4. Cepstrum rzeczywiste sygnału mowy z zaznaczoną granicą między wymuszeniem (*B*) a charakterystyką traktu głosowego (*A*)

Dzięki wykorzystaniu WPT oraz parametryzacji według wzoru (1) otrzymano wyniki rozpoznawania sygnałów w czterech badanych grupach (rys. 5). Osiągnięte wartości podawano wtedy, gdy wynik przekraczał 90%. Jedynie w przypadku głosu kobiety, który nagrywano za pomocą rejestratora, maksymalny uzyskany wynik osiągnął poziom 80%. Charakterystka ucha człowieka nie jest liniowa, więc przeprowadzono dalsze rozpoznawanie sygnałów po zlogarytmowaniu wartości cech wektorów obserwacji. Wyniki uległy znaczącej poprawie, zbliżając się praktycznie do wartości otrzymanych w programie porównawczym. Wyjątek stanowił nagrywany dyktafonem głos męski, którego wynik rozpoznawania nie osiągnął granicy 90%.



Rys. 5. Wyniki rozpoznawania słów parametryzowanych w różnych skalach

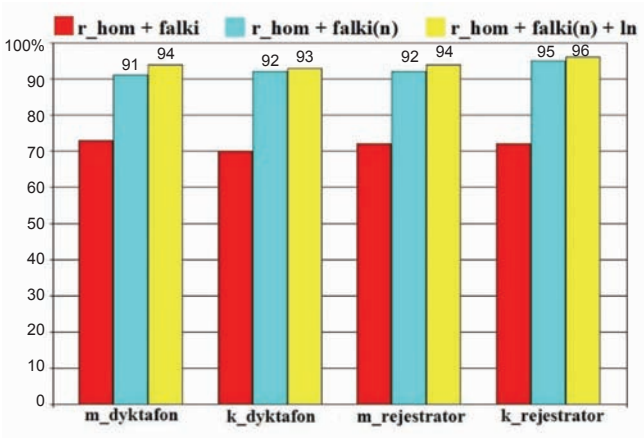
Następnie do zarejestrowanych sygnałów zastosowano rozplot homomorficzny i parametryzację opartą na WPT i wzorze (1). Wyniki pokazano na rys. 6 (pierwsze słupki w poszczególnych grupach reprezentują otrzymane wcześniej wyniki bez stosowania rozplotu).



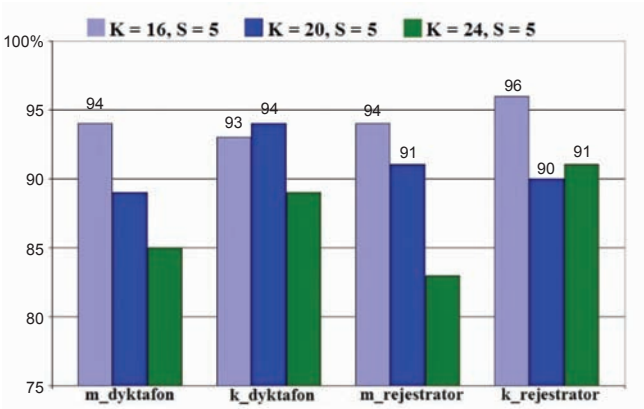
Rys. 6. Wyniki rozpoznawania słów po zastosowaniu rozplotu homomorficznego i skali logarytmicznej

W odniesieniu do zarejestrowanych sygnałów zastosowano z kolei rozplot homomorficzny oraz parametryzację opartą na WPT i wzorze (4), a więc opracowaną własną metodę. Uzyskane wyniki zaprezentowano na rys. 7 (pierwsze słupki w poszczególnych grupach reprezentują wyniki otrzymane wcześniej z zastosowaniem wzoru (1)). W przypadku kobiet osiągnięto wyniki identyczne z otrzymanymi w programie porównawczym. W przypadku mężczyzn wyniki okazały się lepsze o 1%.

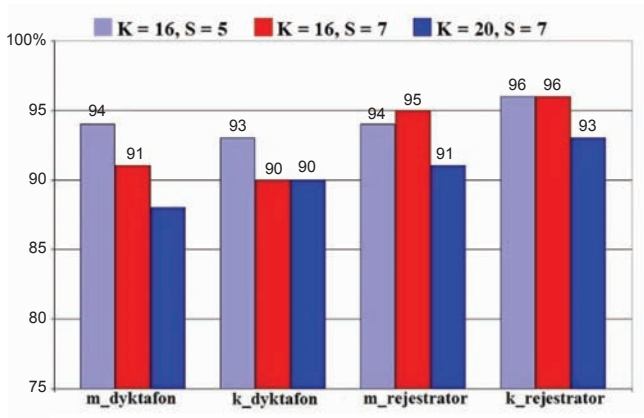
Sprawdzono, jaki wpływ na poziom rozpoznawania mają:
- wzrost liczby centroidów (rys. 8),
- zwiększenie liczby stanów modelu HMM (rys. 9).



Rys. 7. Wyniki rozpoznawania słów po zastosowaniu rozplotu homomorficznego i normalizacji cech WPT



Rys. 8. Wyniki rozpoznawania słów w takcie wzrostu wartości maksymalnej liczby centroidów



Rys. 9. Wyniki rozpoznawania słów w takcie wzrostu liczby stanów HMM i ilości centroidów

Wnioski

W trakcie badań z wykorzystaniem własnej metody rozpoznawania słów z ograniczonego słownika osiągnięto wyniki na poziomie identycznym lub przekraczającym poziom rozpoznawania tych samych słów w programie porównawczym. Sprawdzono wymaganą liczbę centroidów i stanów HMM umożliwiających maksymalizację poziomu rozpoznawania mowy polskiej. Wykazano, że właściwe jest stosowanie

liczby stanów HMM na poziomie 5 i maksymalnej liczby centroidów na poziomie 16.

W przypadku wdrażania opracowanej własnej metody rozpoznawania słów z ograniczonego słownika do sterowania urządzeniami mechatronicznymi należy zbadać jej działanie w rzeczywistych warunkach i eksperymentalnie potwierdzić, że zastosowane algorytmy sprawdzają się w warunkach czasu rzeczywistego.

LITERATURA

1. Furui S. "*Selected topics from 40 years of research in speech and speaker recognition*". Brighton (UK): Interspeech, 2009.
2. Gosiewski Z., Tarasiuk M. "Preliminary study of the automatic speech recognition for devices supporting the ill and disabled". *Journal of Vibroengineering*. Vol. 11 (2009), No. 3: pp. 497÷503.
3. Kasprzak W. „*Rozpoznawanie obrazów i sygnałów mowy*". Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej, 2009.
4. Rabiner L., Huang B.H. "Historical Perspective of the Filed of ASR/ /NLU". "*Springer Handbook of Speech Processing*". Springer-Verlag, 2008.
5. Sarikaya R., Pellom L.B., Hansen J.H.L. "Wavelet packet transform features with application to speaker identification". NORSIG'98 (1998): pp. 81÷84.
6. Sorensen J.A. "Speech Coding and Recognition Course". IT University of Copenhagen, TKG, 2005.
7. Tarasiuk M., Gosiewski Z. „Segmentacja mowy polskiej z wykorzystaniem transformacji falkowej". *Acta Mechanica et Automatica.* Vol. 4 (2010), No. 1: pp. 92÷95.
8. Tarasiuk M., Gosiewski Z. "The Application of Wavelets Vector Quantization of Polish Speech". *Journal of Vibroengineering.* Vol. 14 (2012), No. 1: pp. 87÷94.
9. Zieliński T.P. „*Cyfrowe przetwarzanie sygnałów, od teorii do zastosowań*". Warszawa: WKŁ, 2005.

Publikacja została przygotowana przez stypendystę programu „Stypendia dla doktorantów województwa podlaskiego", realizowanego w ramach Działania 8.2 – Transfer wiedzy, Poddziałania 8.2.2. – Regionalne Strategie Innowacji i Priorytetu VIII Programu Operacyjnego Kapitał Ludzki oraz współfinansowanego ze środków Europejskiego Funduszu Społecznego, budżetu państwa i środków budżetu województwa podlaskiego.