SEGMENTACJA MOWY POLSKIEJ Z WYKORZYSTANIEM TRANSFORMACJI FALKOWEJ

Mirosław TARASIUK[*], Zdzisław GOSIEWSKI[*]

[*]Katedra Automatyki i Robotyki, Wydział Mechaniczny, Politechnika Białostocka, ul. Wiejska 45 C, 15-351 Białystok

miroslaw.tarasiuk@bialystok.policja.gov.pl, gosiewski@pb.edu.pl

Streszczenie: W artykule przedstawiono koncepcję metody segmentacji słów wypowiadanych w języku polskim. Jako narzędzie w procesie segmentacji wykorzystano transformację falkową. Zaproponowano algorytm postępowania oraz przedstawiono wyniki prowadzonych prac badawczych. Wykorzystując opracowaną metodę dokonano podziału wypowiadanych słów i sprawdzono poprawność jego wykonania. Niniejsze badanie stanowi platformę bazową do dalszych prac zmierzających w kierunku opracowania automatycznego systemu rozpoznawania mowy. Badania i obliczenia wykonywano w oparciu o oprogramowanie Matlab.

1. WPROWADZENIE

Wiele ośrodków badawczych na świecie zajmuje się problematyką automatycznego rozpoznawania mowy (ang. Automatic Speech Recognition, ASR). Po wydzieleniu słowa z otaczającej go ciszy (Gosiewski i Tarasiuk, 2009), kolejnym etapem tworzenia ASR jest dokonanie parametryzacji otrzymanego sygnału, ponieważ w postaci czasowej wykazuje on dużą nadmiarowość informacji. Pierwszym krokiem na tym etapie obróbki sygnału mowy jest jego podział na jednorodne akustycznie fragmenty, czyli segmentacja. W rozdziale 2 przedstawiono stosowane metody segmentacji sygnałów. Rozdział 3 zajmuje się prezentacją stosowania dekompozycji falkowej w analizie sygnałów. Możliwości wykonania segmentacji sygnału mowy z wykorzystaniem transformacji falkowej przedstawiono w rozdziale 4. Rozdział 5 to przykład wykonania segmentacji rzeczywistego sygnału mowy.

2. SEGMENTACJA SYGNAŁU MOWY

Segmentacja sygnału akustycznego, w tym i sygnału mowy, może być dokonywana wieloma sposobami. Jeżeli za kryterium podziału sygnału przyjąć długość badanego fragmentu, w którym zakładamy jego quasi-stacjonarność, to możemy wyróżnić dwa typy segmentacji: równomierną (wielokrotność fragmentów stałej długości) lub nierównomierną (zmienna długość fragmentów). Segmentacja równomierna była szeroko stosowana i zakładała ona podział sygnału oknem 10÷40ms (najczęściej 2^k próbek w oknie o łącznej długości około 30ms) z zachodzeniem ramek (ang. overlapping) co najmniej na 1/3 okna (Zieliński, 2002). Niosło to jednak ze sobą ryzyko trafienia do jednej ramki obserwacji różnych fragmentów sygnału, a tym samym przekłamań. Segmentacja nierównomierna dzieli sygnał na fragmenty quasi-stacjonarne o różnej długości, dzięki czemu uwzględnia charakter sygnału. Narzędziem, wręcz predysponowanym do segmentacji nierównomiernej jest transformacja falkowa. Wynika to ze zróżnicowanej wielkości okna dla częstotliwości niskich i wysokich w trakcie badania sygnału (Mallat,1999), którego zasada szczególnie widoczna staje się podczas porównania zjawiska okienkowania transformacji Fouriera (ang. Fourier Transform – FT) i transformacji falkowej. Innym podejściem do tej segmentacji jest wykorzystanie niejawnych modeli Markowa (ang. HMM) w połączeniu z algorytmem Viterbiego (Demuynck i Laureys,2002).

Wykorzystanie funkcji zmienności widmowej sygnału, która jest zależna od czasu zastosował Fang (1994) w swoim algorytmie segmentacji dyskretnego sygnału mowy. Bazuje on na pomiarze częstotliwości chwilowej i wykryciu miejsca gdzie następuje zmiana tej częstotliwości.

Diaunys i inni (2005) zajmując się problematyką segmentacji sygnału mowy, w oparciu o dwa filtry pasmowo przepustowe stwierdzili, że do wydzielenia głosek szczelinowych można wykorzystać stosunek energii pasma wysokiej częstotliwości do energii pasma niskiej częstotliwości, liczony wg wzoru (1). Przy częstotliwości próbkowania 16kHz porównywano dwa pasma częstotliwości 50-2500Hz i 5-7kHz. Należy zauważyć, że stosowano tu okno obserwacji 10ms z 5ms overlappingiem, dla którego wyznaczano energię.

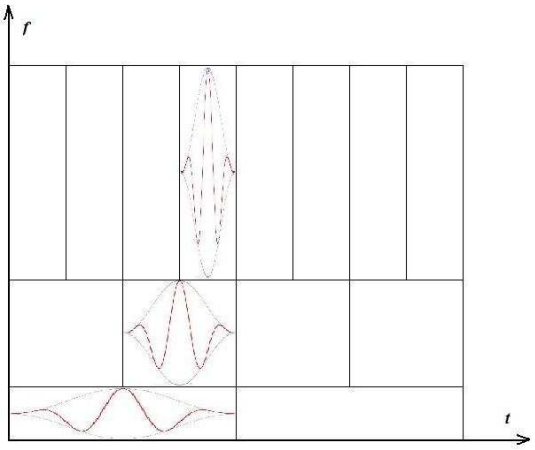
$$F[k] = \frac{E_{hig}[k]}{E_{low}[k]}, \tag{1}$$

gdzie: $E_{hig}[k]$-energia pasma wysokiej częstotliwości, $E_{low}[k]$-energia pasma niskiej częstotliwości, dla k-okna obserwacji. Wykres stworzony na podstawie tej wielkości pozwolił na graficzne wydzielenie głosek szczelinowych z sygnału mowy.

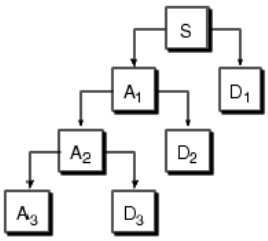
3. DEKOMPOZYCJA FALKOWA

Rozwój transformacji falkowej (ang. Wavelet Transform – WT) historycznie rozpoczął się z ciągłej transformacji falkowej (ang. Continuous Wavelet Transform – CWT).

Pozwoliła ona stosować reprezentację czas-skala do badania analogowych sygnałów, gdzie skala spełnia rolę analogiczną do częstotliwości w analizie częstotliwościowej z FT (Siafarikas i inni, 2008). W przypadku stosowania FT analizuje ona sygnał ze stałą rozdzielczością częstotliwościową zarówno w stosunku do wysokich, jak i niskich częstotliwości. Analiza częstotliwości wysokich wymaga jednak krótszych fragmentów sygnału niż w przypadku częstotliwości niskich. Stosowanie WT w sposób automatyczny usuwa tą niedogodność, ponieważ rozdzielczość częstotliwościowa zmienia się wraz ze zmianą częstotliwości samego sygnału. Schematycznie przedstawiono to na rysunku 1. W przypadku analizy sygnałów dyskretnych za pomocą falek, podstawowym narzędziem jest dyskretna transformacja falkowa (ang. Discrete Wavelet Transform - DWT)



Rys. 1. Schemat rozdzielczości czasowo-częstotliwościowej WT



Rys. 2. Przykład 3 poziomowej dekompozycji sygnału (wg dokumentacji Matlab)

Zastosowanie dyskretnej transformacji falkowej prowadzi do zmniejszenia ilości współczynników o połowę, wraz ze wzrostem poziomu dekompozycji sygnału (Rys. 2). W trakcie dekompozycji sygnału jest on rozkładany na części dolnoprzepustową (oznaczana jako A) i górnoprzepustową (oznaczana jako D). Operacje dalszej dekompozycji dokonywane są zawsze z częścią dolnoprzepustową otrzymaną w poprzednim etapie. Proces syntezy sygnału dokonywany jest według tego samego algorytmu w odwrotnym kierunku.

Gdzie podstawowy sygnał można przedstawić zgodnie z formułą (2). Zapis ten przedstawia ideę przeprowadzania dekompozycji i syntezy oryginalnego sygnału. Znak plus należy tu bowiem traktować jako syntezę, a nie matematyczną operację dodawania.

$$S = A_1 + D_1 = A_2 + D_2 + D_1 = A_3 + D_3 + D_2 + D_1 \qquad (2)$$

4. SEGMENTACJA SYGNAŁU MOWY Z UŻYCIEM TRANSFORMACJI FALKOWEJ

Alani i Deriche (1999) do wyznaczenia segmentów sygnału mowy wykorzystali dekompozycję falkową łącznie z segmentacją równomierną w oparciu o okno Hamminga i 25% overlapping. Badali euklidesową odległość pomiędzy ramkami sygnału. Zastosowanie ramkowania sygnału prowadziło jednak do spadku rozdzielczości czasowej otrzymanej sekwencji.

Tan i inni (1994) zastosowali dekompozycję falkową do wydzielenia pięciu poziomów dekompozycji, niosących informację z różnych przedziałów częstotliwości, które odpowiadają rożnym rodzajom fonemów(nosowe, szczelinowe itd.). Następnie po użyciu 0,8ms okna Hamminga, wyznaczono standardową dewiację sąsiednich ramek, przyjmując że średnia wartość energii każdej ramki dąży do zera. Pozwoliło to na wydzielenie granic poszczególnych dźwięków na poziomach im odpowiadających.

Gałka (2008) zaproponował wykorzystanie widma mocy o jednorodnej długości na każdym z poziomów dekompozycji falkowej do wykonania segmentacji sygnału mowy. Wykazując jednocześnie, że energia widma falkowego jest równa energii sygnału, czyli transformacja falkowa zachowuje energię podstawowego sygnału.

Ponieważ liczba współczynników dekompozycji zależy od poziomu na którym jest liczona, w celu wyrównania długości wierszy, na każdym z poziomów dekompozycji energię wyznaczano biorąc 2^k kolejnych współczynników (k=1, 2, 3, …). Tym sposobem można było zbudować macierz energii liczonej z dekompozycji sygnału, gdzie liczba wierszy jest równa *M+1*. Następnie wykorzystywany był złożony, filtr Tukey'a do określenia granic segmentu sygnału akustycznego. Widmo wyznacza się wykorzystując formułę (3). Na najwyższym poziomie dekompozycji wyznaczano energię tej samej liczby współczynników dla obu składowych (A i D analogicznie).

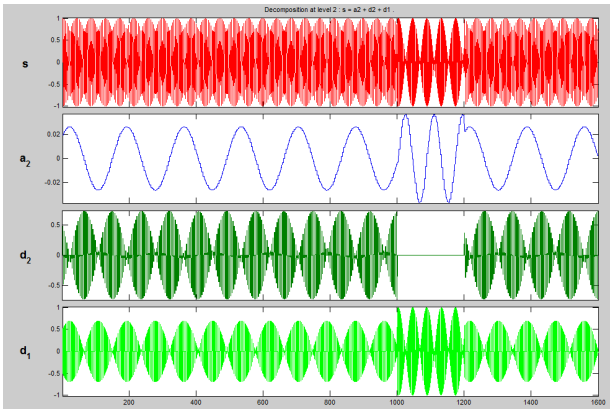
$$d_m^p[k] = \sum_{n=1+(k-1)\cdot 2^{M-m+1}}^{k\cdot 2^{M-m+1}} d_m^2[n], \quad \text{dla } m = 1, 2, ...M\ , \qquad (3)$$

gdzie: *m=1,..*, *M* oznacza użyty poziom dekompozycji, d_{mn} – *n*-ty współczynnik dekompozycji otrzymany w *m*-tym wektorze.

Ważnym etapem jest zebranie doświadczeń z użycia odpowiedniej rodziny falek, jak i doboru poziomu dekompozycji. Na wstępie stworzono w Matlabie testowy sygnał akustyczny o długości 1600 próbek, z częstotliwością próbkowania 4096Hz tonu 1000Hz, zawierający wstawkę długości 50ms tonu 2000Hz poczynając od 1001 próbki. Pozwolił on też na ocenę opracowanej metody segmentacji z uwagi na pełną wiedzę o badanym sygnale *a priori*. Przy takim sygnale po zastosowaniu falki db1 poziomu 2 otrzymujemy bardzo wyraźną formę sygnału wejściowego, przedstawioną na rysunku (3). Dokonano prób z wieloma falkami i poziomami, jednak najlepszą formę sygnału wejściowego podczas dekompozycji, uzyskano właśnie przy użyciu w/w falki.

Z sygnału długości 1600 próbek otrzymano współczynniki falkowe, odpowiednio na każdym poziomie dekompozycji: D1 – 800 współczynników, D2 i A2 – 400.

Uwzględniając częstotliwość próbkowania każdy poziom dekompozycji odpowiada jednemu pasmu częstotliwości: **D1** – 1024÷2048Hz, **D2** – 512÷1024Hz i **A2** – 0÷512Hz.

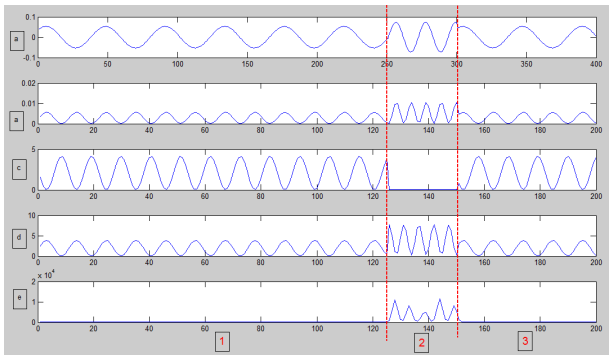


Rys. 3. Dekompozycja sygnału testowego falką db1 poziom 2

Na podstawie współczynników dekompozycji według wzoru (3) można wyznaczyć macierz o rozmiarze 3x200 ze współczynnikami energii widma. Uwzględniając opisany wyżej sposób dekompozycji wzór (1) winien ewoluować do postaci (4).

$$F[k] = \frac{d_1^p[k]}{a_M^p[k] + \sum_{m=2}^{M} d_m^p[k]}, \tag{4}$$

Na podstawie macierzy ze współczynnikami energii widma dokonano porównania energetycznego pasma 1024÷2048Hz do 0÷1024Hz. Ponieważ oceniamy proces zmiany stosunku energii obu pasm w sygnale, można zastosować wydzielenie obwiedni tej zmiany – Rys. (4). Do wydzielenia obwiedni sygnału użyto filtru dolnoprzepustowego ze skończoną odpowiedzią impulsową (ang. Finite Impulse Response filter – FIR).

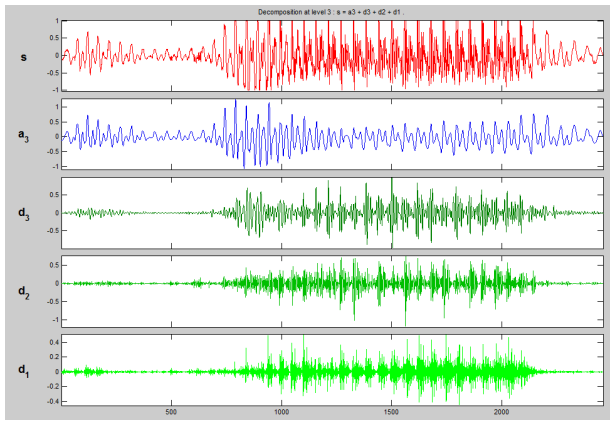


Rys. 4. Zmiany energii widma sygnału; a)współczynniki A2 dekompozycji; b)energia widma z poziomu A2; c) energia widma z poziomu D2; d)energia widma z poziomu D1; e)obwiednia F[k]

Zastosowana metodyka pozwoliła wydzielić z sygnału akustycznego trzy fragmenty oddzielone liniami przerywanymi, które odpowiadają trzem fragmentom o różnej częstotliwości próbnego sygnału.

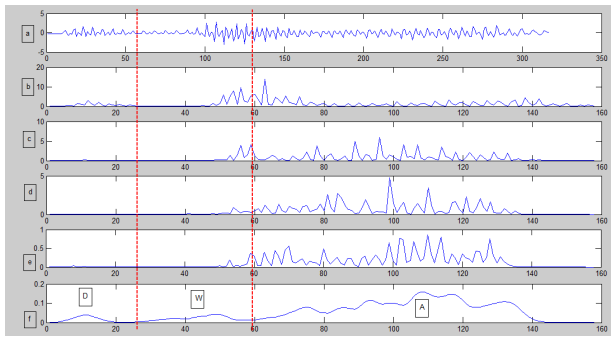
5. PRZYKŁAD

Następnie badaniom poddano słowo „dwa” wypowiedziane przez mężczyznę, które zostało zarejestrowane dyktafonem cyfrowym firmy Sony model ICD-P28 z częstotliwością próbkowania 8[kHz]. Przy wyborze rodziny falek wykorzystano doświadczenie zdobyte w trakcie badań nad wydzieleniem słów z otaczającej ciszy (Gosiewski i Tarasiuk, 2009), jednocześnie sprawdzając możliwość wykorzystania różnych falki i ich poziomów. Czytelna dekompozycja została wykonana falką db6 poziom 3 (Rys. 5). Jest to najniższy poziom oddający prawidłowo kształt badanego sygnału, bez zakłóceń.



Rys. 5. Dekompozycja słowa „dwa” falką db6 poziom 3

Powtórzono wszystkie operacje opisane wcześniej (przy obróbce testowego sygnału akustycznego), które pozwoliły na dokonanie graficznego podziału sygnału na segmenty (Rys. 6). Linie pionowe oddzielają poszczególne, opisane fragmenty słowa. W tym przypadku porównywano energię z pasma 2÷4kHz do energii z pasma 0÷2kHz. Na podstawie wyznaczonych granic dokonano podziału sygnału akustycznego, po odsłuchaniu ich stwierdzono, że słowo „dwa” zostało podzielone poprawnie.



Rys. 6. Zmiany energii widma słowa „dwa”; a)współczynniki A3 dekompozycji; b)energia widma z poziomu A3; c)energia widma z poziomu D3; d)energia widma z poziomu D2; e)energia widma z poziomu D1; f)obwiednia *F[k]*

W każdym wydzielonym segmencie następuje wzrost wartości stosunku energii pasma górnego do dolnego od wartości zerowej do maksimum lokalnego. Dalej następuje jego spadek do wartości zerowej przy końcu segmentu.

Powtarza się to w każdym segmencie. Występują chwilowe spadki jego wartości(fragment z samogłoską A), jednak nie świadczy to o końcu segmentu, a tylko o chwilowej zmienności sygnału.

Ponieważ dokonywano porównania energii górnej części pasma do dolnej, dokonano sprawdzenia czy podobny rezultat zostanie osiągnięty przy wykorzystaniu współczynników dekompozycji sygnału falką db6 poziom 1. W tym przypadku porównywana będzie też energia z pasma 2÷4kHz do energii z pasma 0÷2kHz. Wykonanie dekompozycji na tym poziomie skutkowało jednak otrzymaniem większej liczby współczynników falkowych w każdym z dwóch pasm. Wyznaczanie energii według wzoru (3) prowadziło w istocie do wyznaczenia energii okna obserwacji, o różnej długości na każdym z poziomów dekompozycji. Dokonano więc wyznaczenia energii okna zgodnie ze wzorem (5), co skutkowało porównywaniem takiej samej liczby próbek wartości energii, jak na Rys. 6.

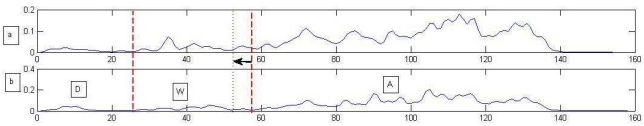
$$d_m^p[k] = \sum_{n=1}^{2^M} d_m^2[n], \quad \text{dla } m = 1, \tag{5}$$

gdzie: *M*-czytelny poziom dekompozycji badanego sygnału. Uwzględniając opisany wyżej sposób wyznaczenia energii wzór (4) uprości się do postaci (6).

$$F[k] = \frac{d_1^p[k]}{a_1^p[k]}, \tag{6}$$

Na Rys. 7 znajduje się porównanie obwiedni *F[k]* otrzymanych z różnych dekompozycji falkowych tego samego sygnału.

Pierwsza granica pomiędzy 1 i 2 segmentami znajduje się w tym samym miejscu w obu przypadkach. Następuje natomiast przesunięcie (oznaczone strzałką) granicy podziału pomiędzy 2 i 3 segmentami. Należy podkreślić, że odsłuchowe sprawdzenie obu sposobów segmentacji, nie pozwala stwierdzić, która z nich jest właściwsza.



Rys. 7. Obwiednia F[k] słowa „dwa"; a)otrzymana na podstawie dekompozycji db6 poziom 1; b)otrzymana na podstawie dekompozycji db6 poziom 3

W następnym etapie dokonano sprawdzenia porównawczego 20 nagrań słowa „dwa" wypowiedzianych przez tego samego mówcę, zarejestrowanych tym samym urządzeniem. Okazało się że 17 nagranych słów, czyli 85% zostało podzielonych na segmenty właściwie.

6. PODSUMOWANIE

Przeprowadzone badania wykazały , że opracowany algorytm segmentacji wykorzystujący transformację falkową, jest ciekawą metodą, która po dalszych badaniach, może stanowić cenne narzędzie analizy mowy.

W dalszych pracach badawczych z większą liczbą słów różnej długości i wypowiadanych przez osoby o zróżnicowanej płci i wieku, należy zweryfikować prawidłowość działania opracowanej metody, jak i ocenić z jakiego poziomu dekompozycji winna być wyznaczana. Ponieważ jej obecne stadium pozwala jedynie na ręczne dokonywanie podziału sygnałów mowy, należy w ich trakcie opracować sposób umożliwiający podział sygnału bez udziału człowieka, w celu późniejszego wykorzystania w automatycznym systemie dokonującym podziału słów na segmenty. Po tym etapie winno nastąpić opracowanie całościowego algorytmu programu, automatycznie wydzielającego słowa z otaczającej ciszy, jak i dzielącego zarejestrowane sygnały mowy na segmenty oraz wybór platformy do jego implementacji.

LITERATURA

1. **Alani A., Deriche M**. (1999), A Novel Approach to Speech Segmentation Using The Wavelet Transform, *Proceedings of The 5th International Symposium on Signal Processing and Applications*, Brisbane, Australia.
2. **Demuynck K., Laureys T.** (2002), A Comparison of Different Approaches to Automatic Speech Segmentation, *Proceedings of the 5th International Conference on Text, Speech and Dialogue*, Lecture Notes in Computer Science, Vol. 2448.
3. **Driaunys K., Rudzionis V., Zvinys P.** (2005), Analysis of vocal phonemes and fricative consonant discrimination based on phonetic acoustic features, *Information Technology and Control*, Vol. 34, No. 3
4. **Gałka J.** (2008), *Optymalizacja parametryzacji sygnału w aspekcie rozpoznawania mowy polskiej*, rozprawa doktorska, Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki, Akademii Górniczo Hutniczej w Krakowie.
5. **Gosiewski Z., Tarasiuk M.** (2009), Preliminary study of the automatic speech recognition for devices supporting the ill and disabled, *Journal of Vibroengineering*, Vol. 11, No 3, 497-503.
6. **Fang X.** (1994), *Automatic Phoneme Segmentation of Continuous Speech Signal*, IEEE Transactions.
7. **Mallat S.** (1999), *A Wavelet Tour of Signal Processing*, Academic Press.
8. **Siafarikas M.**, **Mporas I., Ganchev T., Fakotakis N.** (2008), Speech Recognition using Wavelet Packet Features, *Journal of Wavelet Theory and Applications*, Number 1, 41–59
9. **Tan B. T., Lang R., Schroder H., Spray A., Dermody P.** (1994), Applying Wavelet Analysis to Speech Segmentation and Classification, Wavelet Applications – *Proceedings of SPIE*.
10. **Zieliński T. P.** (2002), *Od teorii do cyfrowego przetwarzania sygnałów*, Zakład Poligraficzny Uniwersytetu Jagielońskiego.

SPEECH SEGMENTATION IN POLISH LANGUAGE BY WAVELET TRANSFORMATION

Abstract: This article introduces an conception on polish spoken words segmentation using wavelet transformation. There was suggested an algorithm and presented achievements made during researches. Spoken words were then divided and their segmentation correctness was verified with use of mentioned above method. This study provides a base platform for further development of the automatic speech recognition system. Research and calculations were executed in MATLAB.