ROZPOZNAWANIE GRANIC SŁOWA W SYSTEMIE AUTOMATYCZNEGO ROZPOZNAWANIA IZOLOWANYCH SŁÓW

Maciej Piasecki, Szymon Zyśko

Wydziałowy Zakład Informatyki
Politechnika Wrocławska
Wybrzeże Stanisława Wyspiańskiego 27
50-370 Wrocław

STRESZCZENIE

W artykule przedstawiono algorytm segmentacji sygnału mowy na odcinki odpowiadające poszczególnym słowom. Prezentowany algorytm został zastosowany w systemie rozpoznawania izolowanych słów. Działanie algorytmu opiera się na analizie dynamiki widma energetycznego sygnału mowy. W pracy przedstawiono ponadto krótką analizę algorytmu „z dołu do góry", który stał się podstawą wyjściową do konstrukcji prezentowanego rozwiązania. Artykuł zawiera również wyniki badań empirycznych nad przydatnością algorytmu testowanego w założonych typowych warunkach pracy.

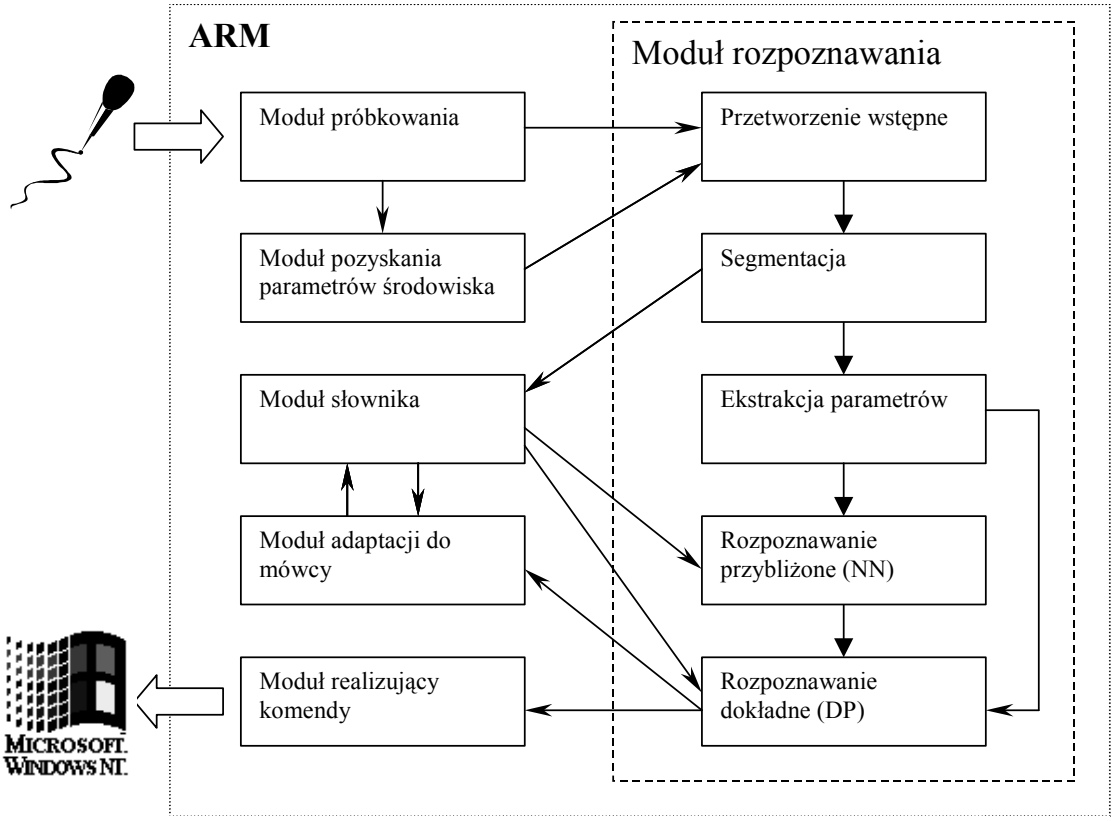
WPROWADZENIE

Segmentacja jest to procedura podziału sygnału na określone mniejsze odcinki znaczeniowe, lub jednostki o stałej długości. [1] W systemach automatycznego rozpoznawania mowy a w szczególności systemach rozpoznawania izolowanych słów po wstępnym przetworzeniu sygnału następuje zazwyczaj szczególny przypadek procesu segmentacji czyli określenie granic poszczególnych wyrazów. Precyzyjne ich określenie jest warunkiem koniecznym do przeprowadzenia prawidłowego procesu rozpoznawania. Dokładne określenie początku i końca słowa utrudnione jest jednak przez szumy otoczenia, *głoski niskoenergetyczne* na początku i końcu słowa, szum spowodowany otwarciem ust na początku wypowiedzi i wydechem na końcu, a także gwałtownymi przerwami (ang. *stop-gap*) spowodowane artykulacja głosek zwartych. [2] Specyfika języka polskiego z dużą ilością głosek niskoenergetycznych, niejednokrotnie występujących na początku i końcu słowa, powoduje, że metody oparte na progowej analizie energii sygnału nie zawsze dają wiarygodne rezultaty. Dotyczy to zwłaszcza sytuacji gdy algorytm nie działa w dobrych warunkach akustycznych.

Prezentowany algorytm został opracowany na potrzeby systemu rozpoznawania izolowanych słów umożliwiającego wydawanie prostych komend w systemie Windows NT 4.0 PL za pomocą mowy w języku polskim (np. zamknięcie okna, uruchomienie programu, kopiowanie, wstawianie, wycinanie). W implementowanym systemie z racji małego słownika rozpoznawanymi jednostkami leksykalnymi będą całe słowa dlatego też szczególnie istotne jest precyzyjne określenie ich granic. W rozpoznawaniu granic słowa biorą udział trzy moduły implementowanego systemu:

- moduł pozyskiwania parametrów środowiska,
- moduł przetwarzania wstępnego,
- moduł segmentacji (rys 1).

Szczególnie istotny jest tutaj moduł pozyskiwania parametrów środowiska, który oprócz procesu kalibracji mikrofonu w celu uzyskania maksymalnie silnego sygnału, nie przekraczającego jednak zakresu próbkowania, dokonuje analizy 5 sekundowego fragmentu ciszy, w celu obliczenia średniego widma ciszy, a następnie oblicza maksymalną energie widma ciszy wykorzystywaną do oszacowania parametrów biorących udział w algorytmie segmentacji. Istnienie tego modułu jest konieczne gdyż program z założenia ma działać u przeciętnego użytkownika komputerów PC, nie można więc założyć że występują dobre warunki akustyczne działania programu.



Rys 1. Schemat działania systemu rozpoznawania izolowanych słów

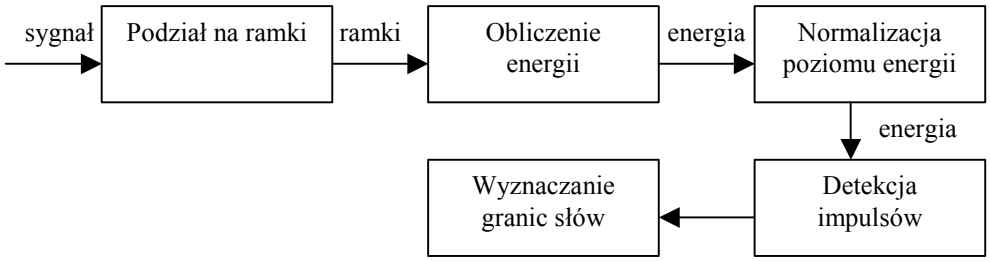
ALGORYTM „Z DOŁU DO GÓRY”

Schemat działania algorytmu „z dołu do góry” zaproponowanego przez Staroniewicza [2], a oparty na modelu Lamela i Rabinera [3] przedstawia rys 2. Sygnał wejściowy jest dzielony na okna a następnie każda ramka jest ważona oknem Hamminga. Po przetworzeniu liczona jest energia sygnału w ramce wg poniższego wzoru:

gdzie: l – jes numerem ramki

$$E(l) = 10\log_{10}\left(\sum_{n=0}^{N} x_l(n)^2\right)$$

Normalizacja poziomu energii polega na obliczeniu różnicy pomiędzy wartością energi w ramce a mnimalną wartością energi sygnału. Następnym etapem jest

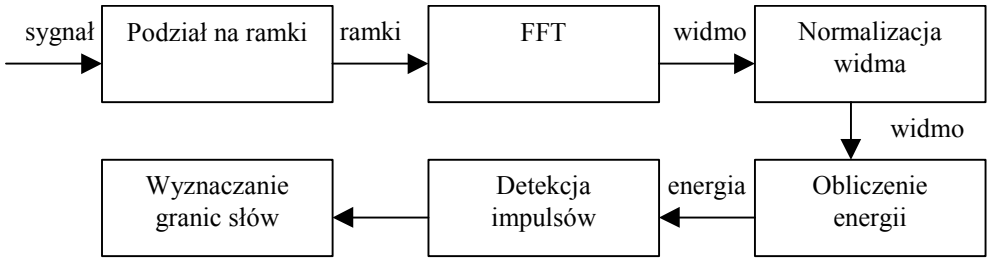


Rys 2. Schemat działania algorytmu „z dołu do góry"

wyznaczenie *impulsów energetycznych*, w tym celu zostały zdefiniowane *trzy progi energetyczne*: K_1, K_2, K_3. Jeśli energia wzrośnie powyżej progu K_1 a następnie nie spadając poniżej jego wartości wzrośnie powyżej progu K_2, oznaczany jest początek impulsu. Następnie gdy energia opadnie poniżej progu K_3 oznaczany jest koniec impulsu. Przyjmuje się, że jeden impuls odpowiada jednemu słowu.

ZMODYFIKOWANY ALGORYTM „Z DOŁU DO GÓRY"

Algorytm „z dołu do góry" nie sprawdził się jednak w praktycznym zastosowaniu w omawianym we wstępie systemie. Przyczyną błędnej segmentacji były głoski niskoenergetyczne występujące na początku słów a także gwałtowne przerwy energetyczne spowodowane wymową głosek zwartych. Usunięcie tych właśnie niedogodności przyświecało podczas dokonywania modyfikacji. Schemat działania zmodyfikowanego algorytmu „z dołu do góry" przedstawia rys 3.



Rys 3. Schemat działania zmodyfikowanego algorytmu „z dołu do góry"

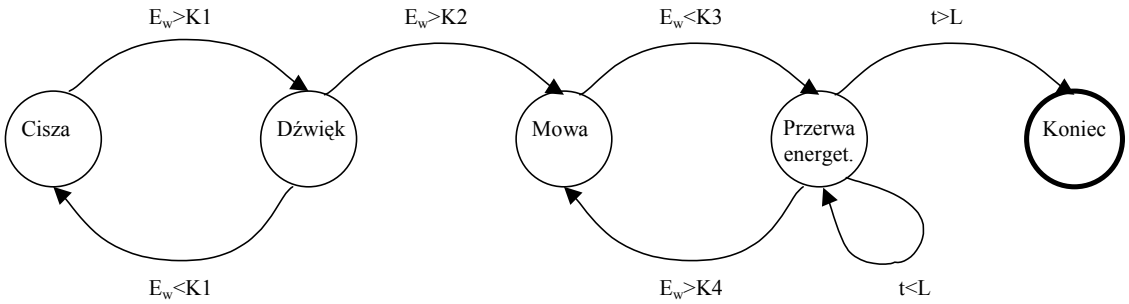
Podobnie jak w oryginalnym algorytmie sygnał wejściowy dzielony jest na ramki, następnie dla każdej z ramek obliczane jest widmo krótkoterminowe za pomocą szybkiej transformaty Fouriera [1], z wykorzystaniem funkcji Hamminga jako funkcji okna. Następnie dokonywana jest normalizacja poprzez odjęcie od obliczonego widma nagranego wcześniej i uśrednionego widma sygnału ciszy. Kolejnym krokiem jest obliczenie energii widma zgodnie z poniższym wzorem:

$$E_w(l) = 10\log_{10}\left(\sum_{f=\frac{F_h N}{2\omega}}^{\frac{F_l N}{2\omega}} |X_l(f)|\right)$$

gdzie:
F_h – wartość częstotliwości dla filtru górnoprzepustowego
F_l – wartość częstotliwości dla filtru dolnoprzepustowego
ω – częstotliwość próbkowania
N – rozmiar okna transformaty Fouriera
X(f) – FFT[x(n)] Szybka Transformata Fouriera
l – numer ramki

W celu wyznaczenia impulsów energetycznych używa się czterech parametrów K1, K2,

K3, K4 oraz parametru L. Początek impulsu oznaczany jest gdy energia widma wzrośnie powyżej progu K1, oraz nie spadając poniżej jego wartości wzrośnie powyżej parametru K2. Koniec impulsu oznaczany jest gdy energia spadnie poniżej progu K3 i



Rys 4. Graf przejść, metody detekcji impulsów energetycznych w zmodyfikowanym algorytmie „z góry do dołu"

w czasie L od momentu spadku nie wzrośnie powyżej progu K4 (rys 4).

BADANIA

Badania przeprowadzono na komputerze PC z kartą dźwiękową SB64 lub SB32. Materiał badawczy procesu segmentacji stanowił zestaw 40 izolowanych słów wypowiadanych z częstotliwością 0.5 słowa na sekundę przez 6 mówców. Stanowiło to łączny materiał 240 słów. Tylko jeden z mówców miał wcześniejsze doświadczenia z systemami automatycznego rozpoznawania mowy. Próbki nie były nagrywane w specjalnie stworzonych warunkach akustycznych, leczy w warunkach reprezentatywnych dla większości domowych użytkowników komputerów PC. Dane były próbkowane z częstotliwością 11025 Hz, ramka miała wielkość 1024 próbki (0.092 s.) a krok ustalono na 350 próbek (0.032 s.). Wartości parametrów K1, K2, K3, K4 ustalono na podstawie analizy sygnału ciszy w następujący sposób: K1 = 0.95*maxE_w, K2 = 1.10*maxE_w, K3 = 1.00*maxE_w, K4 = 1.05*maxE_w. Wartość współczynników dobrano w sposób eksperymentalny. Parametr L został ustalony na 0.15 s. a wartości filtrów F_h = 50 Hz, F_l = 5000 Hz.

Poniższa tabela zawiera dane na temat skuteczności prezentowanej metody dla poszczególnych mówców.

Tabela 1. Skuteczność segmentacji dla poszczególnych mówców

Mówca	1	2	3	4	5	6
Ilość poprawnie wyselekcjonowanych słów	40 (100%)	40 (100%)	40 (100%)	38 (95%)	40 (100%)	40 (100%)
Ilość dodatkowo wyselekcjonowanych słów	1 (2.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (2.5%)

W wyniku przeprowadzonych badań stwierdzono, iż przedstawiony algorytm segmentacji ma średnią skuteczność rozpoznawania wynosi 99.1%. Ilość dodatkowo wyselekcjonowanych impulsów nie będących słowami wypowiedzianymi przez mówce stanowiła 0.83%.

PODSUMOWANIE

Zmodyfikowany algorytm „z góry do dołu" wykazał się dużą skutecznością w przypadku rozpoznawania granic izolowanych słów. Zastosowanie widma sygnału zamiast amplitudy spowodowało lepsze rozpoznawanie głosek niskoenergetycznych znajdują-

cych się na początku i na końcu słowa. Zastosowanie parametru L odpowiadającego za długość sygnału ciszy oraz parametru określającego wartość ponownej aktywacji impulsu K4, pozwoliło na wyeliminowanie błędów procesu segmentacji związanymi z wewnątrzwyrazowymi przerwami energetycznymi. Wadą opisanego algorytmu jest większa złożoność obliczeniowa w stosunku do algorytmu wyjściowego związana z obliczeniem widma sygnału za pomocą szybkiej transformaty Fouriera. Rozwiązanie tego typu jest jednak do przyjęcia w kontekście realizowanego systemu, gdyż widmo energetyczne stanowi materiał bazowy dla dalszego procesu rozpoznawania opartego o analizę częstotliwościową. Problem może stanowić również fakt, iż prezentowany algorytm zakłada pewną wiedzę na temat środowiska akustycznego, w którym działa. Wiedza ta konieczna jest do normalizacji widma sygnały a w realizowanym systemie pobierana jest z modułu pozyskiwania parametrów środowiska, ponadto pozwala na oszacowanie wartości parametrów K1 do K4. Rozwiązanie tego typu dodatkowo pozwala na zmniejszenie wpływu szumów otoczenia przy założeniu, że nie podlegają one zmianom pomiędzy procesem pozyskiwania parametrów środowiska a procesem segmentacji.

W dalszych badaniach należało by sprawdzić przydatność algorytmu w systemach rozpoznawania mowy ciągłej. Można mieć nadzieje iż przyniesie on zadawalające rezultaty w przypadku zastosowania mniejszego rozmiaru okna oraz zmniejszeniu wartości parametru L. Dodatkowe wzbogacenie algorytmu o wykorzystywanie informacji semantycznej i syntaktycznej, tak jak w algorytmie „z góry do dołu" [2], dodatkowo może wzmóc jego skuteczność.

BIBLIOGRAFIA

[1] Basztura Cz., *Rozmawiać z komputerem*, Wydawnictwo Prac Naukowych „FORMAT", Wrocław 1992

[2] Staroniewicz Piotr, Majewski Wojciech, *Określanie granic wyrazów przy głosowym wybieraniu numeru telefonicznego*, XLI Otwarte Seminarium z Akustyki, Wrocław 1994

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