Jubileusz 90 – lecia Akademii Górniczo-Hutniczej
im. Stanisława Staszica w Krakowie
28.05 – 30.05. 2009

Rozpoznawanie Mowy Polskiej

Polish Speech Recognition

Jakub Gałka, Bartosz Ziółko, Mariusz Ziółko
Katedra Elektroniki, Akademia Górniczo-Hutnicza, Al. Mickiewicza 30, 30-059 Kraków
{jgalka,bziolko,ziolko}agh.edu.pl

Streszczenie. *Artykuł podsumowuje badania nad rozpoznawaniem mowy polskiej prowadzone w AGH w zakresie segmentacji, parametryzacji z wykorzystaniem transformacji falkowych oraz modelowania akustycznego, gramatycznego i semantycznego.*

Abstract. *The paper presents research on Polish speech recognition conducted at AGH on segmentation, parameterisation applying discrete wavelet transform and acoustic, grammar and semantic modelling.*

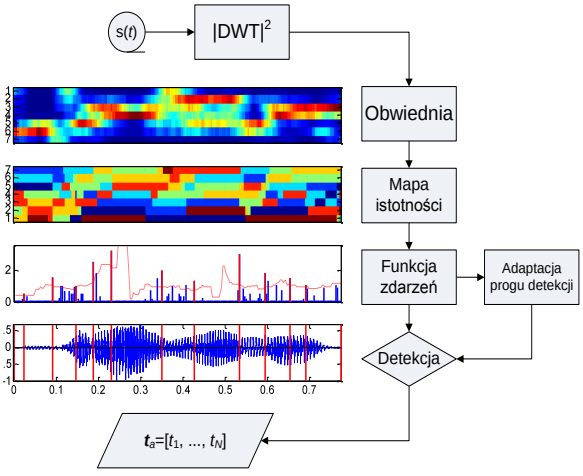
1. Rozpoznawanie mowy. Problematyka automatycznego rozpoznawania mowy (ASR, ang. *Automatic Speech Recognition*) od wielu lat jest przedmiotem zainteresowania ośrodków badawczych na całym świecie. Od kiedy rozwój techniki obliczeniowej pozwolił na implementację skutecznych systemów rozpoznawania, powstało kilka znaczących projektów, których wynikiem są zarówno komercyjne (Dragon, ViaVoice), jak i badawcze (HTK, Sphinx) systemy rozpoznawania mowy (Ziółko, 2008). Większość rozwiązań stworzona została dla języków dominujących, takich jak angielski, chiński, hiszpański, niemiecki czy francuski. Wciąż brak jest niestety efektywnych, ogólnie dostępnych rozwiązań dla języka polskiego. Postęp cywilizacyjny, rozwijająca się gospodarka i nieuchronna globalizacja powodują, że konieczne staje się szybkie rozwiązanie tego zapóźnienia.

Prace Zespołu Przetwarzania Sygnałów w Katedrze Elektroniki, Wydziału Elektrotechniki, Automatyki, Informatyki i Elektroniki w AGH, pod kierownictwem Prof. dr hab. inż. Mariusza Ziółki ukierunkowane są na opracowanie pierwszego polskiego systemu rozpoznawania mowy ciągłej z dużym słownikiem (LVCSR, ang. *Large Vocabulary Continuous Speech Recognition*).

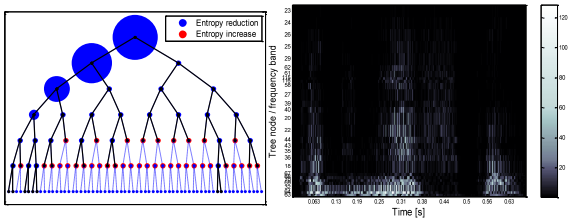
2. Przetwarzanie sygnału. Proces automatycznego rozpoznawania mowy można podzielić na dwa główne etapy. Pierwszym z nich jest rejestracja i przetwarzanie akustycznego sygnału mowy do postaci umożliwiającej jego rozpoznanie. Prowadzone prace doprowadziły do powstania nowoczesnych narzędzi segmentacji i parametryzacji sygnału mowy w oparciu o złożone transformacje falkowe (Ziółko, 2003).

Opracowany algorytm segmentacji nierównomiernej, którego poglądowy schemat przedstawiono na Rys. 1, pozwala na skuteczne izolowanie jednorodnych akustycznie fragmentów sygnału celem ich parametryzacji (Gałka, 2008).

Parametryzacja jest procesem zmniejszania redundancji sygnału i w kontynuowanym projekcie realizowana jest za pomocą złożonych transformacji falkowych, m. in. Paczkowej Transformacji Falkowo-Kosinusowej WPCT



Rysunek 1. Schemat algorytmu segmentacji nierównomiernej sygnału mowy.



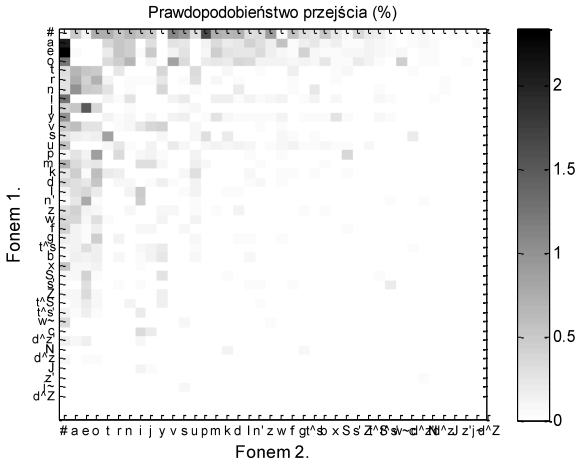
Rysunek 2. Drzewo dekompozycji falkowej uzyskane metodą MBB oraz adekwatne widmo wypowiedzi „*Agnieszka*"z korpusu mowy polskiej.

(ang. *Wavelet Packet Cosine Transform*) oraz nowych opracowanych algorytmów selekcji drzewa dekompozycji - MBB (ang. *Mean Best Basis*). Efektem tych operacji są wektory cech, które w efektywny sposób charakteryzują istotne z psychoakustycznego punktu widzenia właściwości sygnału mowy (Gałka, 2006). Na Rys. 2 przedstawiono przykładowe drzewo dekompozycji oraz adekwatne widmo falkowe sygnału.

3. Modelowanie języka. W nowoczesnych systemach rozpoznawania mowy proces klasyfikacji jest nierozerwalnie złączony z modelowaniem akustycznym elementów języka (np. fonemów) oraz w dalszej kolejności z modelowaniem struktur obszerniejszych (słów, wypowiedzi). Najczęściej wykorzystywanym klasyfikatorem są Niejawne Modele Markowa – HMM (ang. *Hidden Markov Models*), a modelowanie wysokopoziomowe realizowane jest w oparciu o statystyki fonemów oraz *N*-gramy.

W ramach realizowanego projektu stworzone zostały obszerne statystyki zarówno wystąpień fonemów (uni-, di-, tri-fonów) jak i słów (uni-, bi-, tri-gramy), które wykorzystywane są do klasyfikacji oraz korekcji rozpoznawanych sekwencji słów (Ziółko, 2009). Zgromadzony materiał jest prawdopodobnie największym statystycznym obrazem języka polskiego w kontekście wspomnianych wielkości (Rys. 3). Jego przygotowanie wymagało zgromadzenia ponad 2 GB tekstu (m. in. Literatury i Rzeczpospolitej) oraz jego systematycznego przetworzenia przez klaster obliczeniowy ACK Cyfronetu. Statystyki fonetyczne uzyskane zostały z wykorzystaniem oprogramowania PolPhone oraz specjalnie w tym celu przygotowanych programów.

Modelowanie wysokopoziomowe wiąże się z wykorzystaniem metod analizy semantycznej za pomocą macierzy temat/słowo opracowanej w zespole, która w istotny sposób poprawia skuteczność rozpoznawania.



Rysunek 3. Statystyki difonów języka polskiego.

4. Podsumowanie. Zaprojektowane metody są obecnie implementowane jako aplikacja środowiska Win32. Najważniejsze wyniki badań oraz opracowane algorytmy referowane były na wielu konferencjach międzynarodowych. Efektem pracy badawczej są również trzy ukończone doktoraty oraz duża ilość publikacji naukowych w czasopismach oraz materiałach konferencyjnych. Podjęta została także współpraca z różnymi ośrodkami naukowymi. Projekt realizowany jest przy wsparciu Ministerstwa Nauki i Szkolnictwa Wyższego, a także we współpracy z Polską Platformą Bezpieczeństwa Wewnętrznego (http://www.ppbw.pl/).

Literatura

Ziółko Mariusz, Kępiński Michał, Gałka Jakub, (2003), Wavelet-Fourier Analysis of Speech Signal, Proc. of the Workshop on Multimedia Communications and Services, Kielce

Gałka Jakub, (2006), Distance Measures for Wavelet representation of Speech, Proc. of the 12th National Conference on Application of Mathematics in Biology and Medicine - KKZMBM XII, Koninki

Gałka Jakub, Ziółko Mariusz, (2008), Wavelets in Speech Segmentation, Proc. of The 14th IEEE Mediterranean Electrotechnical Conference MELECON'08, Ajaccio

Ziółko Bartosz, Manandhar Suresh, Wilson Richard C., Ziółko Mariusz, Gałka Jakub, (2008), Application of HTK to the Polish Language, Proc. of IEEE International Conference on Audio, Language and Image Processing , Shanghai

Ziółko Bartosz, Gałka Jakub ,Ziółko Mariusz, (2009), Phone, Diphone and Triphone Statistics for Polish Language, Proc. of SPECOM'09, Sankt Petersburg