STUDIA INFORMATICA 2011
Volume 32 Number 2B (97)

Mariusz MĄSIOR, Bartosz ZIÓŁKO, Dawid SKURZOK, Tomasz JADCZYK
Akademia Górniczo-Hutnicza, Katedra Elektroniki

BAZA DANYCH SŁOWNIKA JĘZYKA POLSKIEGO ZE STATYSTYKAMI SŁÓW DLA SYSTEMU AUTOMATYCZNEGO ROZPOZNAWANIA MOWY

Streszczenie. Artykuł opisuje słownik języka polskiego zaimplementowany w postaci bazy danych na potrzeby systemu rozpoznawania mowy. Przedstawiono zastosowania słownika do poprawienia jakości rozpoznania przez modelowanie języka z wykorzystaniem danych przechowywanych w bazie. Zawarto także informacje na temat danych znajdujących się w bazie na koniec stycznia 2011 roku.

Słowa kluczowe: rozpoznawanie mowy, słownik języka polskiego, statystyki tekstów

A DATABASE OF POLISH DICTIONARY WITH WORD STATISTICS FOR AUTOMATIC SPEECH RECOGNITION

Summary. A dictionary of Polish implemented as a data base for automatic speech recognition is presented. The dictionary allows improvement of recognition by language modelling using statistics stored in the data base. The data currently kept in the database are presented as well.

Keywords: speech recognition, ASR, Polish dictionary, text statistics

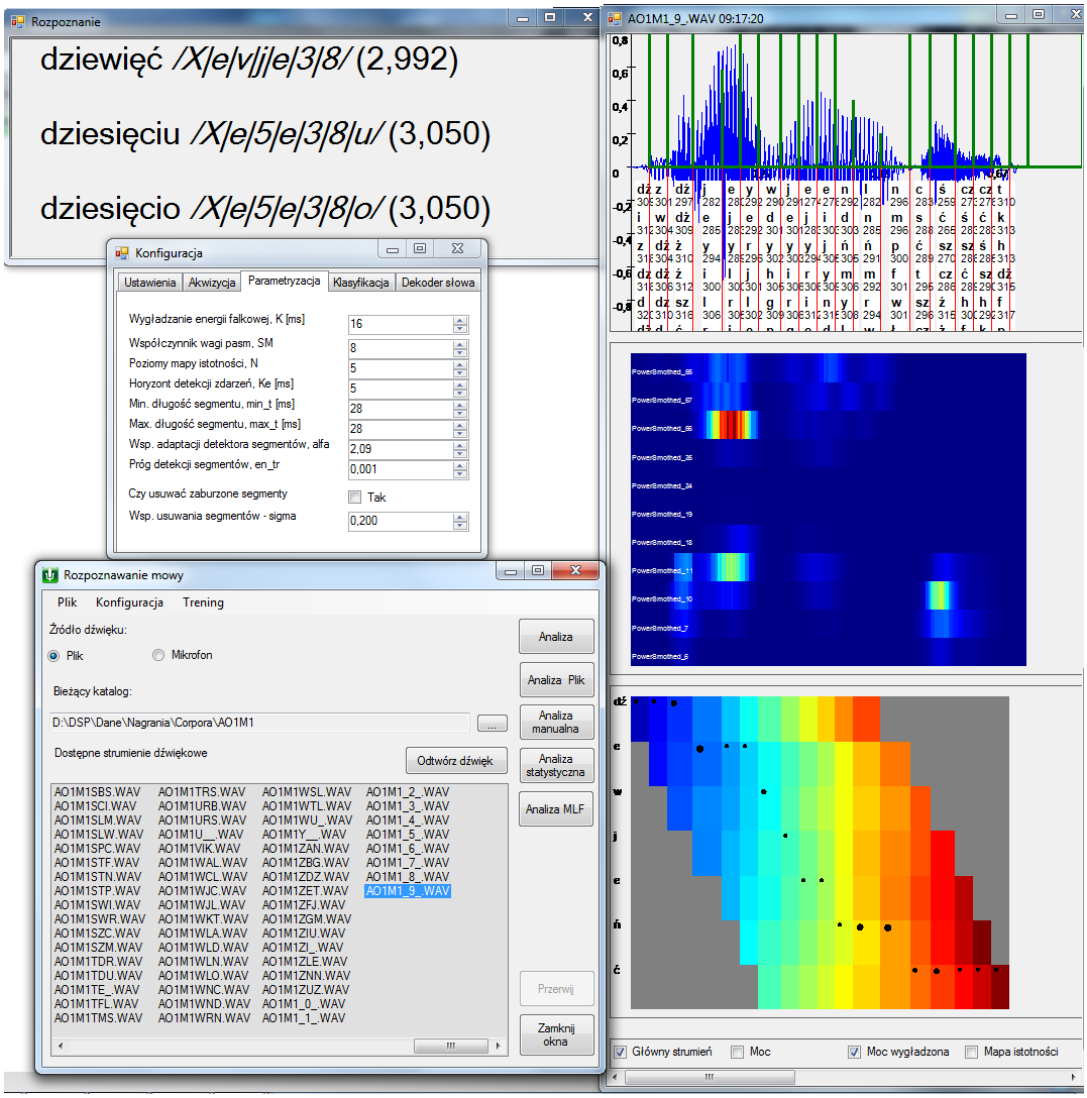
1. Wstęp

Zespół Przetwarzania Sygnałów Katedry Elektroniki Akademii Górniczo-Hutniczej w Krakowie zajmuje się zagadnieniami teoretycznymi i zastosowaniem teorii falek w rozpoznawaniu mowy, kompresji sygnałów i transmultipleksacji.

Prace z zakresu technologii polskiej mowy mają na celu opracowanie prototypowego systemu analizy mowy, połączonego z interfejsem głosowym [1]. Dzięki rozpoznawaniu mowy

będzie można wprowadzać teksty do komputera za pomocą mikrofonu zamiast klawiatury i za pomocą głosu sterować wybranymi funkcjami komputera. Analiza semantyczna tekstu umożliwi dobranie odpowiedniej odpowiedzi komputera i to zarówno w postaci pisanej, jak i głosowej.

Obecnie trwają prace nad system rozpoznawania mowy. Jego rozwój wchodzi w kluczowy etap optymalizacji i poprawy szybkości działania. Jednocześnie wciąż są prowadzone badania nad poprawą algorytmów rozpoznawania mowy ciągłej i całości fraz dyktowanych przez użytkownika. Rysunek 1 przedstawia aktualny wygląd systemu rozpoznawania mowy.



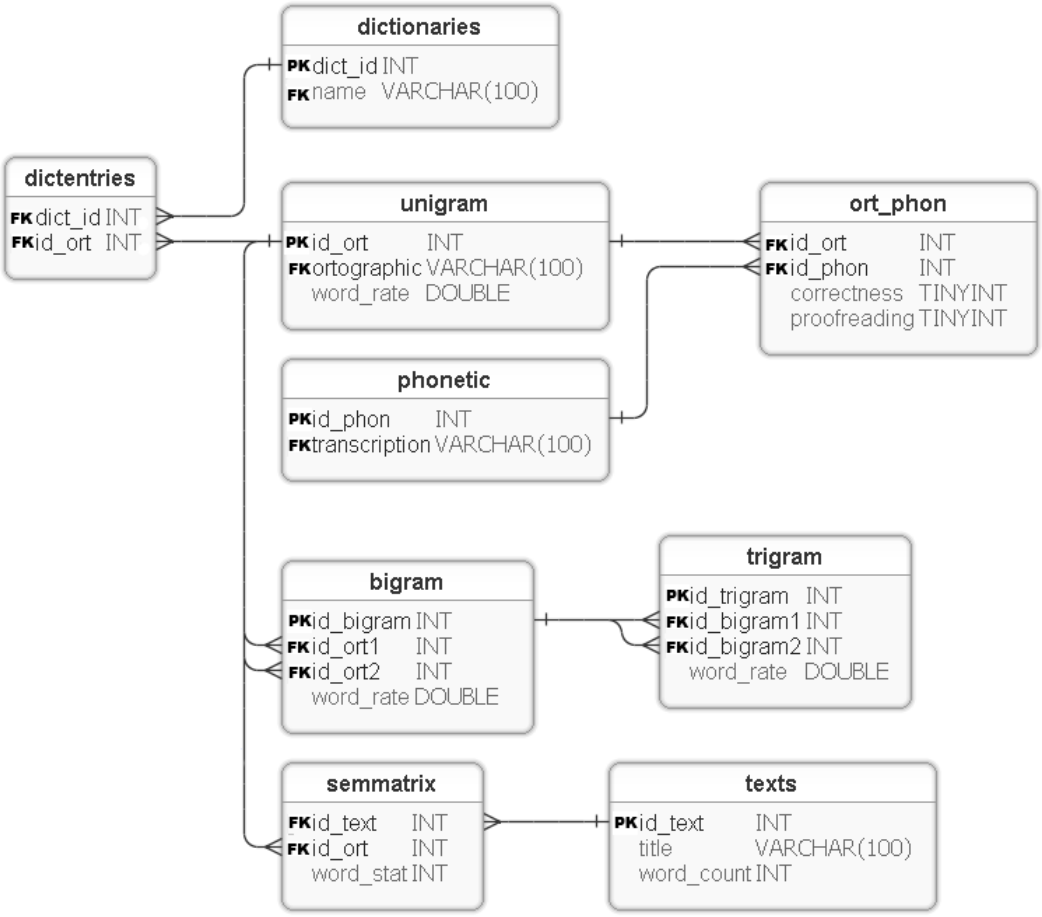
Rys. 1. Program do rozpoznawania mowy
Fig. 1. Program for speech recognition

Integralną częścią systemu jest baza danych zawierająca słowniki języka polskiego wraz z zapisami fonetycznymi słów. W bazie znajdują się również statystyki częstości występowania pojedynczych słów, ich par i trójek. Statystyki te w znacznym stopniu poprawiają możliwości rozpoznawania całych zdań wypowiadanych w mowie ciągłej.

Prędkość obsługi słowników jest kwestią kluczową w wydajności czasowej systemu rozpoznawania mowy, który w założeniu ma działać w czasie rzeczywistym. Z tego powodu dobór struktury bazy danych, jej rozmiar i zależności pomiędzy poszczególnymi tabelami są istotnymi decyzjami projektowymi.

2. System Bazodanowy

2.1. Struktura bazy danych



Rys. 2. Struktura bazy słownictwa w systemie rozpoznawania mowy polskiej
Fig. 2. Structure of the vocabulary database in speech recognition system for Polish

W obecnym stadium pracy nad systemem rozpoznawania mowy istotna jest możliwość badania działania systemu dla różnych podzbiorów słownictwa języka polskiego. Relacyjna baza danych o strukturze przedstawionej na rys. 2 umożliwia prowadzenie takich testów.

Baza danych jest grupą współpracujących ze sobą słowników, które są zbiorami nierozłącznymi (różne słowniki mogą zawierać te same słowa). Podział na słowniki wynika zarów-

no z tematyki (na przykład specjalizowane słowniki prawnicze i medyczne), jak i ilości słów (na przykład 1000 lub 10 000 najpopularniejszych słów w języku).

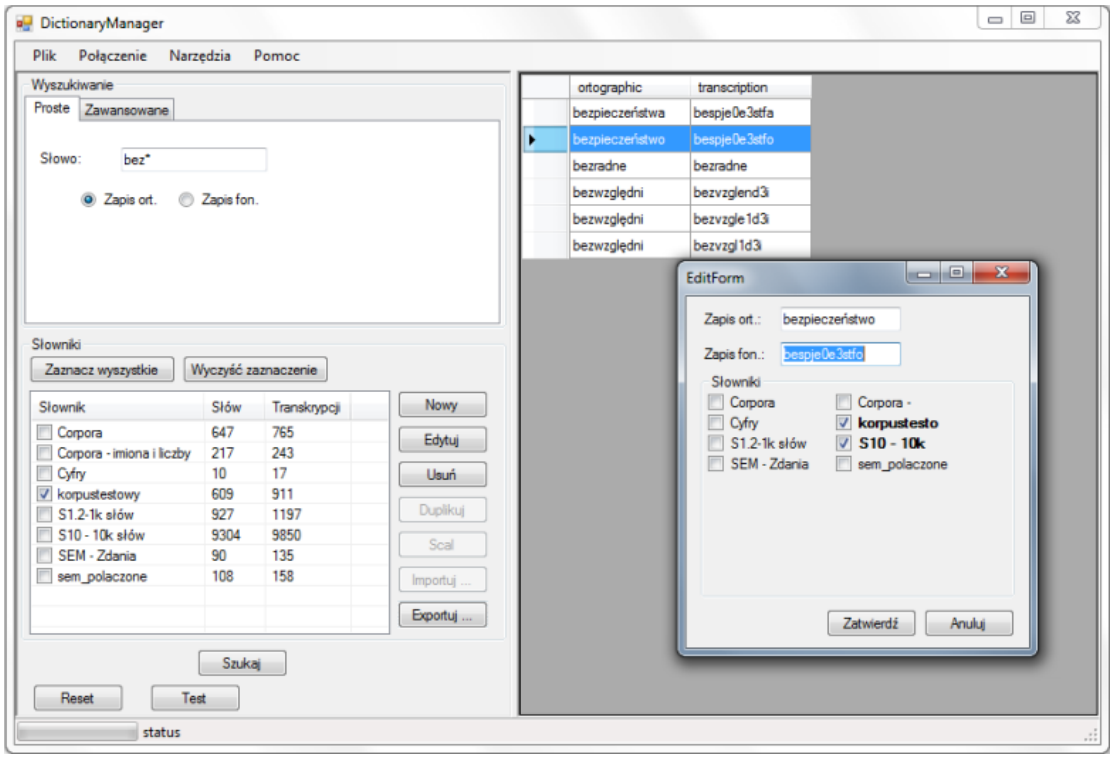
W tabeli 1 został przedstawiony aktualny podział używanego słownictwa.

Tabela 1

Spis słowników przechowywanych w bazie danych

Słownik	Ilość słów	Ilość transkrypcji
Corpora – zdania	656	767
Corpora – imiona i liczby	217	239
Cyfry	10	12
Referat na temat mowy – trening	1361	1741
Referat na temat mowy – test	689	864
S1.2	937	1300
S1 – prawnicze	988	1368
S10	9304	9850
SEM – zdania testowe	90	135
SEM	108	158

2.2. Transkrypcje słów w bazie



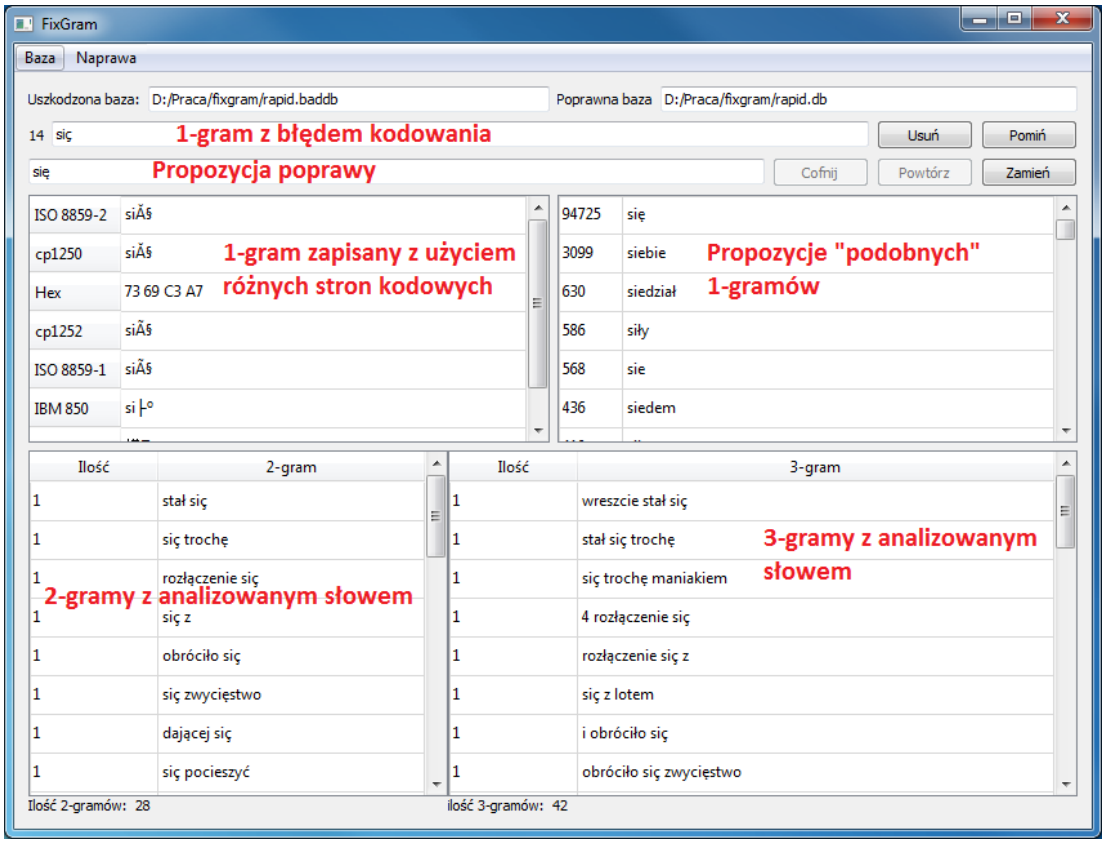
Rys. 3. DictionaryManager – program do edycji bazy danych słowników
Fig. 3. DictionaryManager – program for dictionary database edition

Zapis fonetyczny [2] każdego słowa może wymagać ręcznej korekty przez osobę posiadającą odpowiednią wiedzę z zakresu mowy polskiej. W celu uniknięcia wielokrotnego sprawdzania tych samych słów, w bazie danych zostały wprowadzone dodatkowe informacje o po-

prawności i korekcie danego słowa (kolumna „corectness” w tabeli „ort_phon”). Do takiej analizy wykorzystujemy zarówno autorskie oprogramowanie (rys. 3), jak i program standardowo dostępny z silnikiem bazy – MySQL Workbench.

Istotnym zagadnieniem jest również sposób kojarzenia zapisów ortograficznych i fonetycznych, gdyż w pewnych sytuacjach nie są to relacje jeden-do-jeden. Na przykład słowa „morze” i „może” dzielą ten sam zapis fonetyczny. Analogicznie, słowo posiadające jeden zapis ortograficzny może być wymawiane na kilka sposobów, w zależności od dialektu i staranności wymowy. Dlatego też wymagane było stworzenie tabeli „ort_phon”, realizującej wzorzec tabeli łączy dla relacji wiele-do-wielu (pomiędzy zapisami ortograficznymi i fonetycznymi).

2.3. Statystyki występowania słów



Rys. 4. FixGram – program do korekty słów przed umieszczeniem w bazie
Fig. 4. FixGram – program for words correction before inserting into database

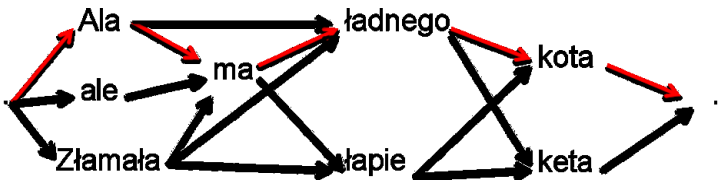
W bazie danych oprócz zapisów ortograficznych i fonetycznych przechowywane są także statystyki występowania słów (unigramy), ich par (bigramy) oraz trójek (trigramy). Statystyki te znacznie podnoszą skuteczność rozpoznawania mowy ciągłej oraz umożliwiają podjęcie decyzji, czy rozpoznane zdanie jest poprawne z punktu widzenia języka.

Struktura bazy danych (rys. 2) przewiduje także tabele („semmatrix” i „texts”) do przechowywania wiedzy semantycznej na temat poszczególnych słów. W obecnej wersji dane te nie są wykorzystywane. Planuje się ich użycie przez algorytm analizy hipotez, nieuwzględniający porządku słów w hipotezach, co umożliwi wnioskowania na podstawie dużo szerszego kontekstu [3, 4].

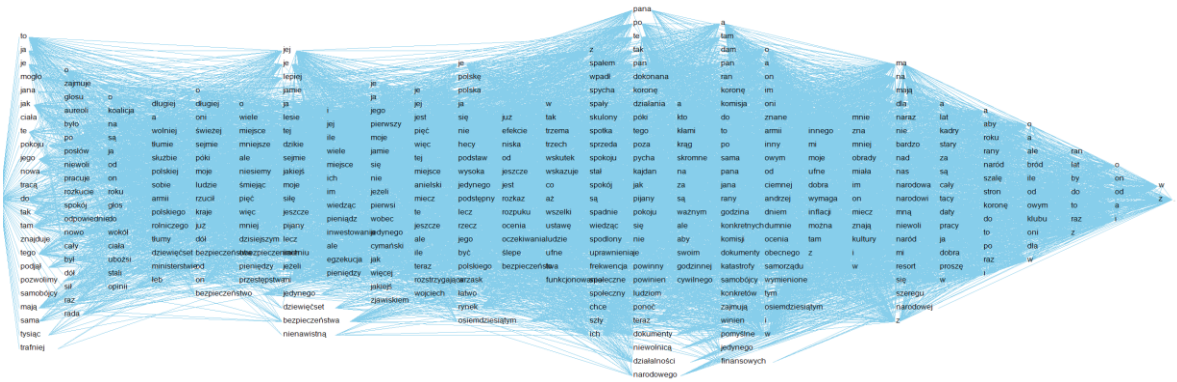
Przechowywanie statystyk w bazie wymusiło jej optymalizację pod tym kątem. Zapis ortograficzny słów „konsumuje” relatywnie dużo pamięci, co przy przechowywaniu kilku milionów kombinacji możliwych słów, stanowi istotny problem zużycia zasobów w systemie, a także ewentualnego udostępniania programu, na przykład na płycie CD lub DVD. Z tego powodu statystki zostały zbudowane w formie relacji pomiędzy tabelami „unigram”, „bigram” i „trigram” (rys. 2). Do zapisu ortograficznego (tabela „unigram”) przypisana jest statystyka pojedynczych słów (kolumna „word_rate”). Bigramy są generowane z podwójnej relacji do pojedynczych słów oraz analogicznie, aby oszczędzać pamięć, trigramy są generowane z pary rekordów tabeli z bigramami.

Statystyki przechowywane w opisanej bazie gromadzone są automatycznie z dużej liczby tekstów, jednakże w samej bazie umieszczane są jedynie słowa istniejące w języku polskim. Słowa budzące podejrzenia są wcześniej przekierowywane do korekty przez człowieka [5]. Służy temu program przedstawiony na rys. 4. Znajduje on słowa różniące się jedynie odmiennym zapisem ortograficznym (takim jak „rz” i „ż”), słowa zawierające w sobie niepolskie znaki oraz słowa nieodnalezione w słowniku MySpell. Ta ostatnia funkcjonalność nie została jeszcze wykorzystana ze względu na olbrzymią liczbę nowych słów w naszych statystykach w porównaniu do istniejących słowników. Wynika to głównie z nieuwzględniania nazw własnych w klasycznych słownikach. W przypadku naszego słownika, mającego wspomagać rozpoznawanie mowy, nazwy własne muszą być traktowane w sposób zbliżony do innych słów i zawierać się w słowniku.

Słownik zawiera dane statystyczne umożliwiające ocenę hipotez zdań. Rysunki 5 i 6 przedstawiają takie hipotezy. Są one grafami skierowanymi, w których węzłami są słowa, a krawędziami są możliwe połączenia słów. Dane 1-gramowe ze słownika przypisywane są węzłom, 2-gramowe krawędziom, natomiast 3-gramowe wartości wykorzystywane są do modyfikacji wag krawędzi przez algorytm zbliżony do algorytmu Dijkstry [6].



Rys. 5. Przykładowy graf hipotez rozpoznawanej frazy
Fig. 5. Example of hypotheses graph of recognized phrase



Rys. 6. Rzeczywisty przypadek grafu hipotez rozpoznawanej frazy
Fig. 6. A real case graph of recognized phrase's hypotheses

2.4. Zastosowana technologia

Do przechowywania słowników został użyty system MySQL. Jest on jednym z najpopularniejszych systemów zarządzania relacyjnymi bazami danych. Wybór tego systemu został podyktowany jego dostępnością i szybkością działania [7], co jest szczególnie istotne przy rozpoznawaniu mowy.

Przy projektowaniu bazy danych należało położyć szczególną uwagę na problemy i niestandardowe zachowania MySQL-a, dotyczące wartości domyślnych oraz kodowania bazy. Szczególnie ta ostatnia kwestia sprawiła trudność, gdyż zadaniem bazy danych jest przechowywanie zapisów słów w języku polskim, czyli zapisów z literami diakrytycznymi. W systemie rozpoznawania mowy wykorzystujemy kodowanie UTF-8, które nie jest jednak domyślne dla systemu MySQL. Także biblioteki łączności udostępniane z bazą danych mają domyślne kodowanie inne niż UTF-8. Dodatkowo MySQL domyślnie nie rozróżnia małych i wielkich liter, co jest istotne przy analizie zapisu fonetycznego, w którym każdy znak ma inne znaczenie.

Programy komputerowe, które powstały w trakcie pracy nad systemem, zostały stworzone w technologii .NET (C# i zarządzany C++). Łączność z bazą danych została zrealizowana za pomocą biblioteki sterownika łączności dostarczanego przez producenta MySQL. Sterownik ten działa na podstawie technologii .NET, dzięki czemu połączenie bazy danych z tworzonymi aplikacji nie było pracochłonne i zagwarantowało dobrą szybkość przesyłu danych.

3. Podsumowanie

W ramach pracy nad aplikacjami mowy, a w szczególności nad systemem automatycznego rozpoznawani mowy, powstała baza danych słownika języka polskiego, zawierająca staty-

styki występowania słów i ich zapis fonetyczny. Baza została zbudowana na podstawie systemu MySQL i aktualnie jest wypełniana danymi generowanymi z różnych tekstów literackich oraz stron internetowych.

Praca naukowa finansowana ze środków na naukę w latach 2008-2011 jako projekt rozwojowy MNISW OR00001905.

BIBLIOGRAFIA

1. Ziółko M., Gałka J., Ziółko B., Jadczyk T., Skurzok D., Wicijowski J.: Automatic Speech Recognition System Based on Wavelet Analysis – DEMO. Proceedings of Fourth IEEE International Conference on Semantic Computing (IEEE ICSC2010), Pitsburg, USA 2010.
2. Demenko G., Wypych M., Baranowska E.: Implementation of Grapheme-to-phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-speech Synthesis, Speech and Language Technology. PTFon, Poznań 2003.
3. Ziółko B., Manandhar S., Wilson R. C., Ziółko M.: Bag-of-words Modelling for Speech Recognition. Proceedings of International Conference on Future Computer and Communication (ICFCC 2009), Kuala Lumpur, Malezja 2009.
4. Wicijowski J., Ziółko B.: Analiza skupień i redukcja wymiarowości w hierarchicznym modelu korpusowym. Studia Informatica, Vol. 31, No. 2A (89), Wyd. Pol. Śl., Gliwice 2010, s. 133-145.
5. Ziółko B., Skurzok D., Michalska M.: Polish n-grams and their correction process. Proceedings of The 4th International Conference on Multimedia and Ubiquitous Engineering (MUE 2010), Cebu, Filipiny 2010.
6. Dijkstra E. W.: A Note on Two Problems in Connexion with Graphs, Numerische Mathematik, 1959.
7. Lerner R. M.: Open-Source Databases, Part III: Choosing a Database. Linux Jurnal, 2007.

Recenzenci: Dr inż. Małgorzata Bach
Prof. dr hab. inż. Tadeusz Wieczorek

Wpłynęło do Redakcji 31 stycznia 2011 r.

Abstract

A dictionary of Polish implemented as a data base for automatic speech recognition system (Fig. 1) was presented. The paper describes structure of the database (Fig. 2) and importance of each table. The dictionaries currently kept in the database are summarised in Table 1. Part of the challenge with making the dictionary is that it has to conatain phonetic transcriptions to be useful for speech recognition. Our speech recognition system connects to the database and a user can choose which of the exact dictionaries they want to use in particular tasks by operating a dictionary manager (Fig. 3). The dictionary stores statistics about words co-occurences which firstly can be corrected by a human operator using FixGram (Fig. 4). Then the statistics can be applied to n-gram model (Fig. 5 and 6) in the speech recognition system.

Adresy

Mariusz MĄSIOR: Akademia Górniczo-Hutnicza, Katedra Elektroniki, al. Mickiewicza 30, 30-059 Kraków, Polska, masior@agh.edu.pl.
Bartosz ZIÓŁKO: Akademia Górniczo-Hutnicza, Katedra Elektroniki, al. Mickiewicza 30, 30-059 Kraków, Polska, bziolko@agh.edu.pl.
Dawid SKURZOK: Akademia Górniczo-Hutnicza, Katedra Elektroniki, al. Mickiewicza 30, 30-059 Kraków, Polska.
Tomasz JADCZYK: Akademia Górniczo-Hutnicza, Katedra Elektroniki, al. Mickiewicza 30, 30-059 Kraków, Polska, jadczyk@agh.edu.pl.