Inżynieria Akustyczna – Technologia Mowy 2013
Wojciech Binek

Sprawozdanie z laboratoriów HTK

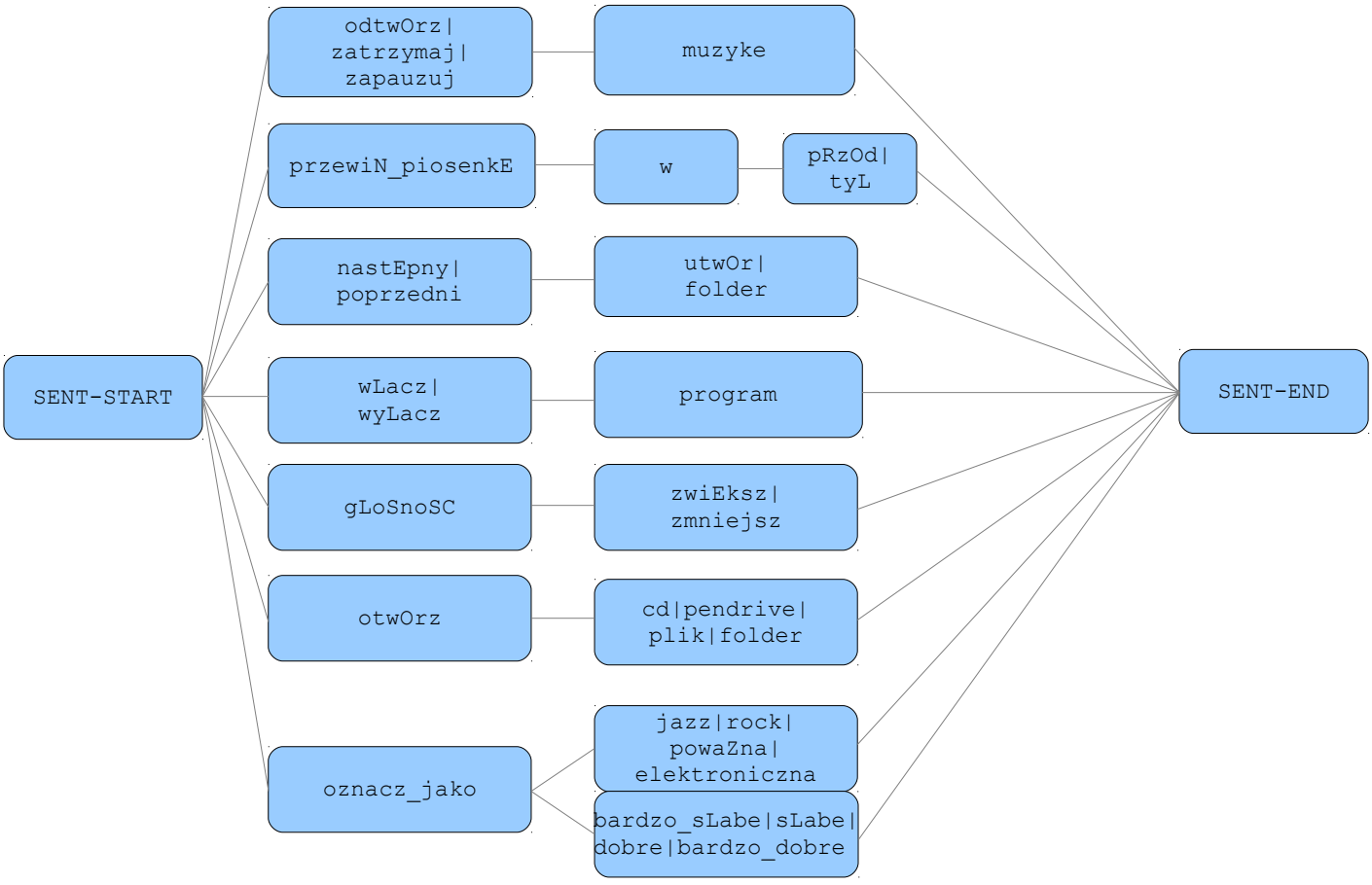
1. Opis gramatyki

System rozpoznawanie mowy był projektowany do obsługi odtwarzacza muzyki. Zawiera on podstawowe komendy umożliwiające odtwarzanie i zatrzymywanie piosenek, przełączanie ich, wybór źródła muzyki oraz dodawanie prostych tagów. W gramatyce nie występują słowa, które pojawiają się w każdym zdaniu. Większość komend zawiera jedno stałe słowo i zestaw czynności które można z nim wykonać np. muzykę odtwórz|zatrzymaj|zapauzuj. Słownik składa się łącznie z 33 słów, niektóre z nich posiadają kilka zapisów fonetycznych, w celu uwzględnienie różnej wymowy danego słowa.

Plik opisujący gramatykę:

```
$startstop = odtwOrz | zatrzymaj | zapauzuj;
$kierunekpios = przOd | tyL;
$kierunekutw = nastEpny | poprzedni;
$onoff =wLacz | wyLacz;
$coprzewin = folder | przewiN_piosenkE;
$zwzm = zwiEksz | zmniejsz;
$zrodlo = cd | pendrive | plik | folder;
$podobalnosc = bardzo_sLabe | sLabe | dobre | bardzo_dobre;
$gatunek = jazz | rock | powaZna | elektroniczna;


(SENT-START ($startstop muzyke | przewiN_piosenkE w $kierunekpios
| $kierunekutw (utwOr |  folder)| $onoff program | gLoSnoSC $zwzm
| otwOrz $zrodlo | oznacz_jako ($gatunek|$podobalnosc)) SENT-END)
```



2. Opis nagrań

Nagrania testowe i treningowe wykonano w zamkniętym i umeblowanym pomieszczeniu przy użyciu:

Mikrofon – Behringer XM8500 (dynamiczny)
Rejestrator – Line6 POD X3 Live

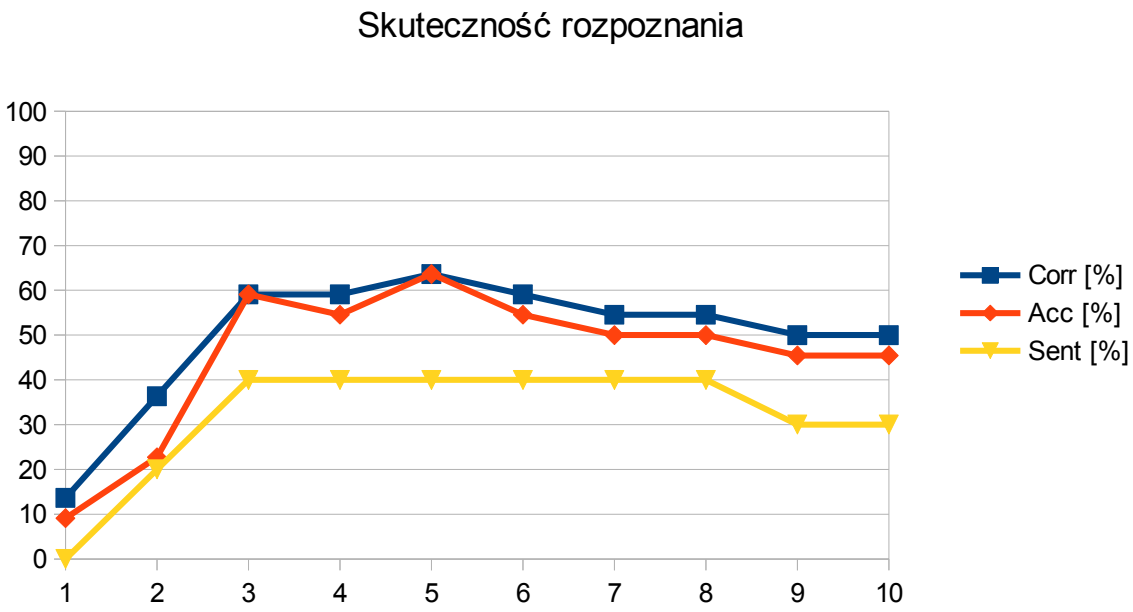
Nagranie treningowe zawierało każde możliwe zdanie wypowiedziane trzykrotnie, całość trwała niewiele ponad 3 minuty. Nagrania testowe stanowiło 10 nagrań zarejestrowanych przez tego samego mówcę, w tych samych warunkach.

3. Wyniki

Początkowo najlepsze wyniki oscylowały około 45% poprawnie rozpoznanych słów i 20% poprawnie rozpoznanych zdań. Po modyfikacjach słownika (dodaniu lub zmianie zapisu fonetycznego niektórych słów) udało się uzyskać następujące rezultaty.

Reestymacja	Corr [%]	Acc [%]	Sent [%]
1	13,64	9,09	0
2	36,36	22,73	20
3	59,09	59,09	40
4	59,09	54,55	40
5	63,64	63,64	40
6	59,09	54,55	40
7	54,55	50	40
8	54,55	50	40
9	50	45,45	30
10	50	45,45	30

Tab. 1 Wyniki rozpoznania nagrań testowych



rys. 1 Wyniki rozpoznania nagrań testowych

4. Wnioski

Zarówno rozpoznanie poszczególnych słów jak i całych zdań jest na akceptowalnym poziomie. Stosunkowo wysoki procent poprawnie rozpoznanych zdań wynika z mało skomplikowanej gramatyki – większość komend składa się z tylko dwóch słów.

Analizując hipotezy rozpoznania zwracane przez program, można zauważyć, że w większości przypadków słowa, które są źle rozpoznane nie są rozpoznawane wcale (nie mieszczą się w pierwszych 5 hipotezach). Świadczy to o błędnym zapisie fonetycznym. Niestety, jak się później okazało, stworzony słownik nie zawierał wszystkich potrzebnych fonemów (brakowało m.in. ą,ę,c) i niemożliwe okazało się poprawienie niektórych błędnych zapisów fonetycznych. Należy więc szczególnie uważnie i dokładnie wykonywać transkrypcje fonetyczne, bo etap ten jest najbardziej kluczowy dla jakości rozpoznania.

Wyrażam zgodę na dołączenie naszych nagrań do korpusu mowy AGH. Nagrania mogą być odtwarzane ale wyłącznie bez podawania tożsamości mówców (np. w celu prezentacji jakości, rodzaju nagrań itd.).