Raport z systemu rozpoznawania mowy HTK

Katarzyna Baruch, Karolina Kolber Inżynieria Akustyczna – Technologia Mowy

1. Wstęp:

Stworzono dwa systemy rozpoznawania mowy. Omówiono tutaj głównie wyniki z drugiej wersji systemu, która dała większy procent rozpoznania.

2. Opis nagrań:

Nagrania testowe wykonano za pomocą mikrofonu z zestawu A4 TECH model HS-7P na komputerze Lenovo ThinkPad Edge 11 (NVY3LPB) ze zintegrowaną kartą muzyczną, wykorzystując program Samplitude natomiast treningowe na komputerze TOSHIBA Satellite L500 (mikrofon, program j.w.). Wykonano 20 minut 10 sekund nagrań treningowych mowy ciągłej, w której słowa wypowiadane były z krótką przerwą pomiędzy (łącznie ok. 250 zdań). Warunki wykonania nagrań treningowych i testowych– cichy pokój mieszkalny.

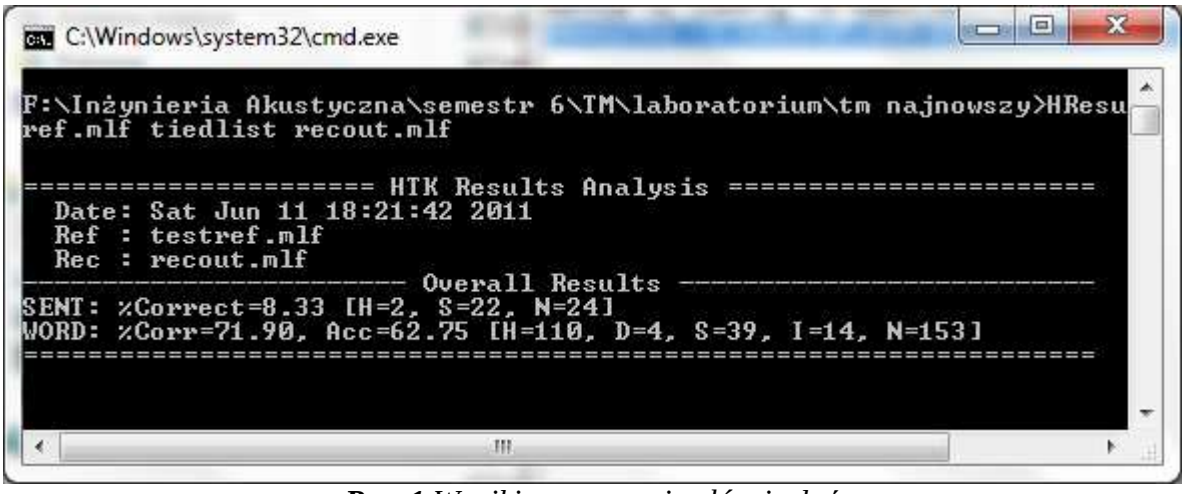
Nagrania znajdują się w katalogu kobiety/KB1. *(nagrania dosyłane były ponownie w późniejszym terminie, więc możliwe jest, że znajdują się w innym katalogu)*

3. Opis gramatyki:

```
$powitanie = dzieN_dobry | witam;
$czasownik = zamawiam | proszE | chciaLabym;
$co = [ mroZonA | zimnA | rozpuszczalnA ] kawE | latte;
$dodatek = [lodami | czekoladA | amaretto ];
(sent-start $powitanie ($czasownik ($co [z ($dodatek) [i ($dodatek)]) ])) sent-end)
```

System mógłby służyć do automatycznego telefonicznego przyjmowania zamówień na kawę np. w biurze dużej firmy. Osoba zamawiająca miałaby kartę z wypisanymi możliwymi kawami do wyboru. U nas byłyby to kawa, kawa rozpuszczalna, mrożona, zimna, latte i możliwe dodatki tj. lody, czekolada i amaretto. Osoba zamawiająca nie musiałaby poświęcać wiele czasu na naukę składni. System rozpoznaje zdania rozpoczynające się od powitania: dzień dobry lub witam, kolejnym wymaganym elementem jest orzeczenie: zamawiam, proszę, chciałabym, następnie zamówienie: kawę, kawę mrożoną, rozpuszczalną, zimną (dowolna kolejność przymiotnik – rzeczownik lub rzeczownik – przymiotnik), latte i opcjonalnie żaden, jeden lub dwa dodatki spośród trzech. W gramatyce celowo nie zapisano opcji, która wyklucza powtórzenie się dwóch dodatków, ponieważ uznano za korzystniejsze, gdy ktoś otrzyma kawę bez jednego z zamówionych dodatków niż otrzyma ten dodatek, którego nie zamawiał.

4. Wyniki z HResults:

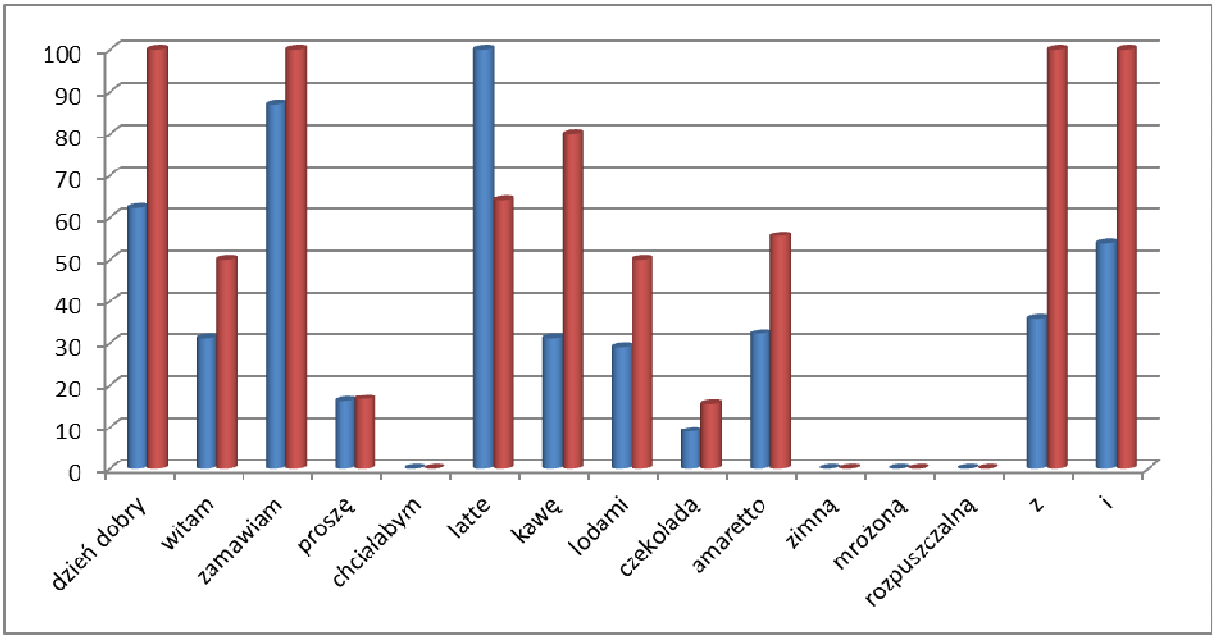


Rys. 1 *Wyniki rozpoznania słów i zdań dla II systemu rozpoznawania mowy – 3 reestymacja*

Najwyższy procent rozpoznawalności słów dla tego systemu uzyskano dla 3 reestymacji i osiągnął on 71,9%.

5. *Analiza błędów rozpoznania:*

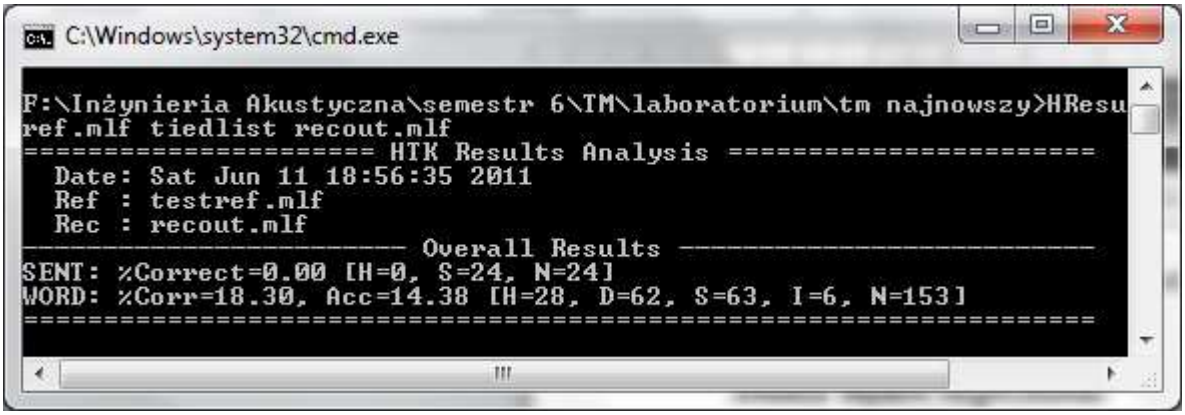
Najwyższą rozpoznawalność uzyskały słowa, które w z zdaniach treningowych pojawiały się najczęściej np. „dzień dobry” (100%), „zamawiam” (100%), „i” (100%), „z” (100%) oraz „kawę” (80%). Natomiast najniższą rozpoznawalność mają słowa „mrożoną” (0%), „zimną” (0%), „rozpuszczalną” (0%) a także „chciałabym”. Główną przyczyną braku rozpoznania tych słów jest najprawdopodobniej zbyt mała częstotliwość występowania danego słowa w bazie treningowej systemu oraz wyższy stopień rozbudowania zdania, w którym takie słowo występuje.



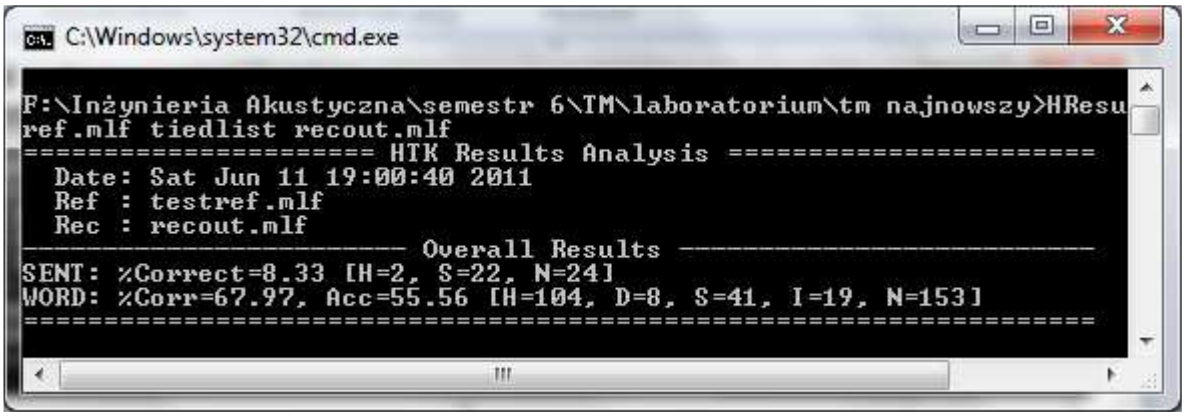
Rys. 2 *Zestawienie rozpoznawalności poszczególnych wyrazów w II systemie rozpoznawania mowy*

Wykres (Rys. 2) przedstawia wyrażoną w procentach wielkość rozpoznania poszczególnych słów (kolor czerwony), oraz stopień rozpoznania słów w zależności od częstotliwości ich występowania w bazie treningowej (kolor niebieski) – wartości te zostały odniesione do słowa „latte”, rozpoznanego w 64% przy 56 wystąpieniach w nagraniach treningowych.

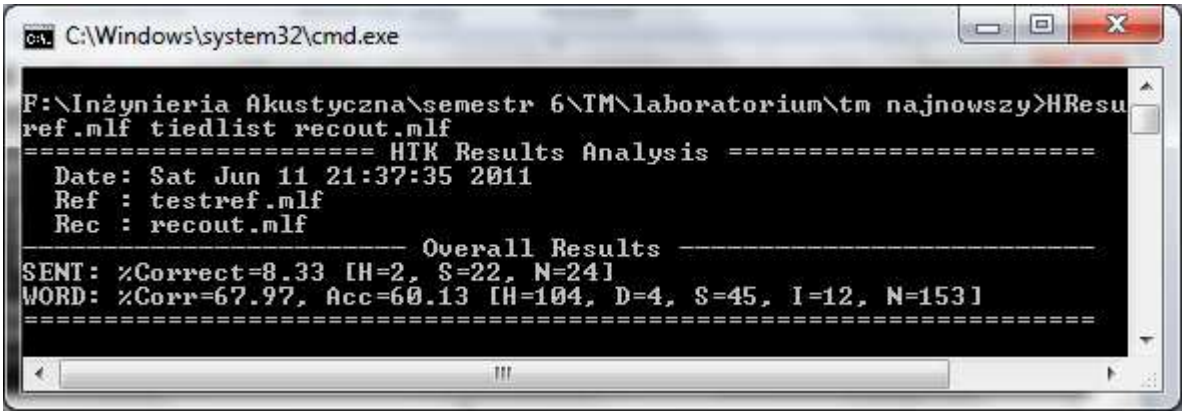
6. *Analiza różnych rozwiązań (np. różna liczba reestymacji):*



Rys. 3 *Wyniki rozpoznania słów i zdań dla II systemu rozpoznawania mowy – 1 reestymacja*



Rys. 4 *Wyniki rozpoznania słów i zdań*
dla II systemu rozpoznawania mowy – 2 reestymacja

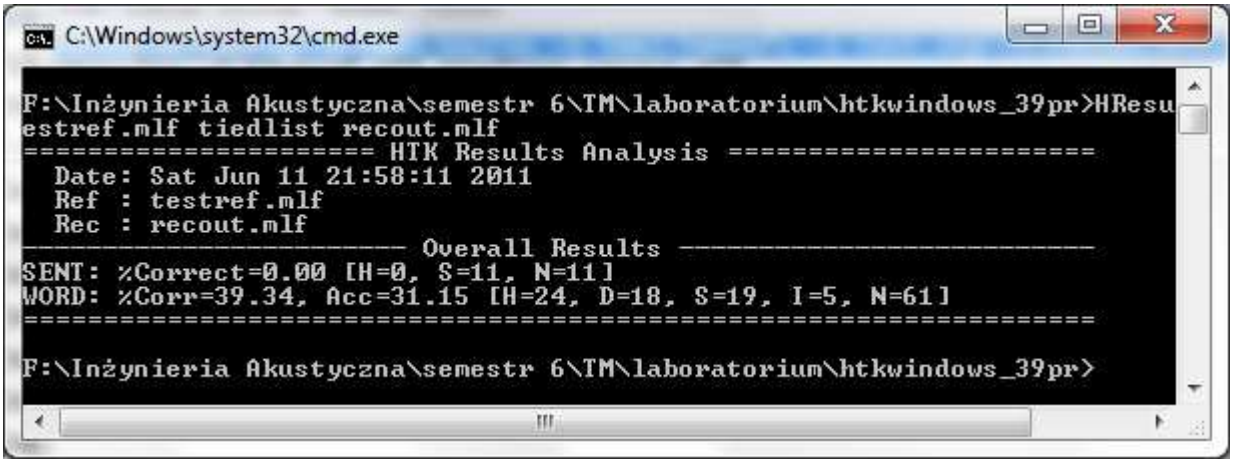


Rys. 5 *Wyniki rozpoznania słów i zdań*
dla II systemu rozpoznawania mowy – 4 reestymacja

Najlepsza rozpoznawalność słów w systemie została zaobserwowana dla 3 reestymacji. Każda kolejna dała niższą rozpoznawalność – przyczyną jest tzw. przetrenowanie systemu.

Za każdy m razem najwięcej błędów rozpoznania wynikało z podstawienia (S) - program w miejsce prawidłowego słowa wstawiał inne. Natomiast najmniej błędów było rezultatem całkowitego usunięcia wyrazu (D) z wypowiedzianego zdania.

7. Pierwszy system rozpoznawania mowy:



Rys. 6 *Wyniki rozpoznania słów i zdań*
dla I systemu rozpoznawania mowy – 8 reestymacja.

Pierwszy, stworzony przez autorów system do rozpoznawania mowy opierał się na 4 minutach 18 sekundach nagrań treningowych (67 zdań), na 1 minucie 1 sekundzie nagrań testowych (11 zdań) - mowa ciągła. Słownik zawierał 34 słowa. Zarówno tematyka, struktura gramatyczna jak i przeznaczenie tego systemu było takie samo jak dla systemu II.

Rozpoznawalność słów, w porównaniu z II systemem rozpoznawania mowy, była niemalże dwukrotnie mniejsza.

8. Wnioski:

Decydującym czynnikiem wpływającym na jakość rozpoznawania mowy jest stosunek długości nagrań treningowych do liczby słów, które mają być rozpoznawane. Inną bardzo ważną kwestią jest jakość nagrań treningowych (np. stosunek sygnału do szumu), a także sposób wymawiania poszczególnych zdań (mowa ciągła, słowa izolowane).

Wyrażamy zgodę na dołączenie nagrań do korpusu mowy AGH.

Załączniki:
Pliki testowe wav wraz z testref.mlf i lab.
Katalogi hmm i inne elementy systemu użyte do testów.