AKADEMIA GÓRNICZO-HUTNICZA IM. S. STASZICA W KRAKOWIE

WYDZIAŁ INŻYNIERII MECHANICZNEJ I ROBOTYKI

INŻYNIERIA AKUSTYCZNA

SYLWIA BAŁAZY

TECHNOLOGIA MOWY

SPRAWOZDANIE Z LABORATORIÓW HTK

1. OPIS GRAMATYKI

Założeniem projektu była stworzenie systemu rozpoznawania mowy realizującego automatyczne kupowanie biletów Komunikacji Miejskiej w Krakowie. System taki można byłoby zastosować w automatach z biletami oraz w aplikacjach na telefony komórkowe, które aktualnie umożliwiają zakup biletu z poziomu aplikacji.

Gramatyka (plik 1) została przygotowana tak, by kupno biletów dokonywane było w formie grzecznościowej (na początku zdania znajdować się musi „poproszę") i aby zakupić bilet konieczne jest wypowiedzenie wszystkich cech biletu – najpierw użytkownik musi przekazać informacje na jaką strefę ma być bilet - strefowy (Miasto Kraków) lub aglomeracyjny (aglomeracja krakowska), następnie podać rodzaj ulgi, a dopiero potem typ biletu. Szyk musi być tak dobrany z uwagi na to, że dostępnych jest wiele rodzajów biletów. W związku z tym użytkownik musi podać wszystkie informacje, żeby możliwe było dokonanie zakupu.

W gramatyce zastosowano także kilka różnych możliwości dla danego typu biletu, np. aby zakupić bilet godzinny można również poprosić o bilet sześćdziesięciominutowy itp.

W gramatyce rozróżniono typy biletów możliwe do zakupu dla danej strefy – bilety w obrębie Miasta Krakowa mają dużo więcej typów niż te, które umożliwiają przejazd po całej aglomeracji.

```
$ulga = ulgowy | normalny;
$typS = jednoprzejazdowy | dwuprzejazdowy |
      grupowy | dwudziesto minutowy | czterdziesto minutowy |
      godzinny | szeSCdziesiEcio minutowy | dziewiECdziesiEcio minutowy |
      caLodniowy | dwudziestocztero godzinny |
      dwudniowy | czterdziestooSmio godzinny |
      trzydniowy | siedemdziesiEciodwu godzinny |
      tygodniowy | siedmiodniowy | weekendowy;
$typA = jednoprzejazdowy | dwuprzejazdowy |
      grupowy | godzinny | szeSCdziesiEcio minutowy |
      dziewiECdziesiEcio minutowy |
      caLodniowy | dwudziestocztero godzinny |
      tygodniowy | siedmiodniowy | weekendowy;

( SENT-START ( poproszE bilet strefowy $ulga $typS | poproszE bilet
aglomeracyjny $ulga $typA ) SEND-END)
```

PLIK 1 – GRAMATYKA

Do stworzenia pliku ze słownikiem (plik 2) użyto programu OrtFon. Każde słowo użyte w gramatyce zostało w tym programie przekonwertowane do notacji fonetycznej Corpora, a następnie w odpowiedni sposób zapisane w pliku ze słownikiem. Na końcu każdego słowa została zapisana cisza – 'sil', po to by zwiększyć rozpoznawalność słów w systemie.

```
poproszE                p o p r o sz e sil
bilet                   b i l e t sil
strefowy                s t r e f o w y sil
aglomeracyjny           a g l o m e r a c y j n y sil
ulgowy                  u l g o w y sil
normalny                n o r m a l n y sil
jednoprzejazdowy        j e d n o p sz e j a z d o w y sil
dwuprzejazdowy          d w u p sz e j a z d o w y sil
grupowy                 g r u p o w y sil
minutowy                m i n u t o w y sil
godzinny                g o dzi i n n y sil
dwudziesto              d w u dzi e s t o sil
czterdziesto            cz t e r dzi e s t o sil
szeSCdziesiEcio         sz e zi dzi e si e ni ci o sil
dziewiECdziesiEcio      dzi e w j e ni dzi e si e ni ci o sil
caLodniowy              c a l_ o d ni o w y sil
dwudziestocztero        d w u dzi e s t o cz t e r o sil
dwudniowy               d w u d ni o w y sil
czterdziestooSmio       cz t e r dzi e s t o o si m j o sil
trzydniowy              t sz y d ni o w y sil
siedemdziesiEciodwu     si e d e m dzi e si e ni ci o d w u sil
tygodniowy              t y g o d n i o w y sil
siedmiodniowy           si e d m j o d ni o w y sil
weekendowy              l_ i k e n d o w y sil
SEND-END [] sil
SENT-START [] sil
```

PLIK 2 – SŁOWNIK

2. OPIS NAGRAŃ

Zarówno nagrania treningowe, jak i testowe zwierają mowę ciągłą i wykonane zostały w warunkach domowych (niski poziom tła akustycznego oraz niewielki pogłos pomieszczenia) przy użyciu rejestratora Zoom H2. Oba nagrania treningowe trwają w sumie 3 minuty i 29 sekund, podczas których wypowiedziane zostały 53 zdania, składające się razem z 283 wyrazów[1]. W przypadku nagrań testowych zarejestrowanych zostało 12 zdań zawierających w sumie 65 wyrazów.

Nagrane przy pomocy rejestratora pliki miały format .wav, były nagrane z użyciem dwóch kanałów (stereo), a próbkowanie wynosiło 44,1 kHz. Do zmiany parametrów nagrania na tryb mono i próbkowanie 16 kHz użyto programu Audacity.

Po odpowiedniej anotacji nagrań na poziomie słów przy pomocy programu Anotator, anotacja na poziomie fonemów została przeprowadzona przez inż. Piotra Żelasko. Jednak z uwagi na wykorzystanie modelu z ciszą, należało ręcznie dodać do pliku „segmentacja.mlf" kilka przedziałów czasowych, w których występowała cisza.

[1] niektóre wyrazy zostały podzielone na dwie części, aby usprawnić rozpoznawanie – np. wyraz „dwudziestominutowy" został podzielony na dwa wyrazy: „dwudziesto" i „minutowy", ponieważ druga część wyrazu jest używana również z innymi przedrostkami „czterdziesto", „sześćdziesięcio" i „dziewięćdziesięcio".

3. Wyniki testów z HResults

Do tworzenia poszczególnych plików oraz do sprawdzenia w jaki sposób zaprojektowany system rozpoznaje mowę skorzystano z wiersza poleceń (rysunek 1).

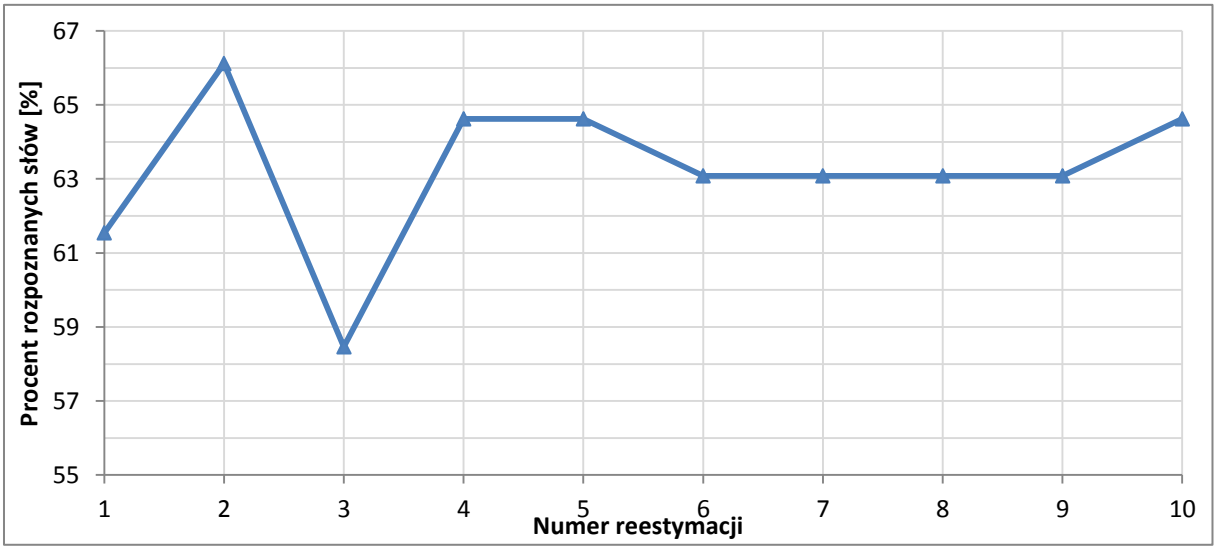
```
E:\TM\lab5>HResults -I testref.mlf monophones0 recout2.mlf
====================== HTK Results Analysis =======================
  Date: Thu Jan 02 15:12:02 2014
  Ref : testref.mlf
  Rec : recout2.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=8.33 [H=1, S=11, N=12]
WORD: %Corr=66.15, Acc=64.62 [H=43, D=3, S=19, I=1, N=65]
===================================================================
```

RYSUNEK 1 - FRAGMENT WIERSZA POLECEŃ Z UZYSKANYM WYNIKIEM DLA DRUGIEJ REESTYMACJI

Wyniki uzyskane tym sposobem zamieszczono w tabeli 1 oraz na wykresie (rysunek 2).

TABELA 1 - WYNIKI UZYSKANE Z HRESULTS

nr reestymacji	procent rozpoznanych zdań [%]	procent rozpoznanych słów [%]	D - usunięcia	S - zamiany	I - wstawienia
1	8,33	61,54	2	23	2
2	8,33	66,115	3	19	1
3	0	58,46	5	22	0
4	0	64,62	5	18	0
5	0	64,62	5	18	0
6	0	63,08	5	19	0
7	0	63,08	5	19	0
8	0	63,08	5	19	0
9	0	63,08	5	19	0
10	0	64,62	5	18	0



RYSUNEK 2 - WYNIKI UZYSKANE Z HRESULTS PRZEDSTAWIONE NA WYKRESIE

Najlepszą rozpoznawalność uzyskano dla drugiej reestymacji – ponad 66% rozpoznanych słów oraz jedno w całości poprawnie rozpoznane zdanie.

Na uwagę zasługuje też ogromny spadek rozpoznawalności dla trzeciej reestymacji – skuteczność systemu spada tu do najniższej wartości – 58%, podczas gdy kolejne reestymacje osiągają wyniki powyżej 63%.

W powyższej tabeli (tabela 1) zamieszczono także informacje o tym, co stało się z nierozpoznanymi słowami – większość z nich system zamienił na inne, nie więcej niż 5 słów zostało usuniętych i w przypadku dwóch pierwszych reestymacji niewielka ilość słów została wstawiona. Od 4 reestymacji ilość zamian utrzymuje się na podobnym poziomie (18 lub 19). Natomiast od 3 reestymacji jest stała zarówno ilość usunięć (5), jak i wstawień (0).

4. Analiza błędów rozpoznania

Na taki poziom rozpoznawalności duży wpływ miał stosunek słów stałych do zmiennych. Każde zdanie rozpoczyna się słowami: „Poproszę bilet”, po których następuje podanie informacji niezbędnych do zakupu odpowiedniego typu biletu. Na 65 wyrazów występujących w nagraniach testowych – 24 są stałe i za każdym razem zostały rozpoznane poprawnie.

Trudno jest na podstawie kilku nagrań testowych przeanalizować z rozpoznawaniem jakich słów system sobie nie radzi, dlatego przeanalizowano jedynie najczęściej występujące słowa przy reestymacji dającej najlepsze wyniki (reestymacja 2). Okazało się, że system dobrze rozpoznaje słowa „poproszę” i „bilet”, ponieważ jak zostało to wyżej wyjaśnione są one słowami stałymi. Poprawnie rozpoznawane są też słowa: „strefowy” oraz „normalny”. Natomiast systemowi prawie w ogóle nie udało się poprawnie odszyfrować słów: „aglomeracyjny” oraz „ulgowy”. Analizując zakończenie zdań testowych, występujące tam słowa system w większości przypadków rozpoznawał błędnie, a dodatkowo często wyświetlając słowo „trzydniowy”, jako ostatnie występujące w komendzie, mimo tego, że nie pojawiło się ono ani raz podczas nagrań testowych. W przypadku innych reestymacji system także często rozpoznawał, że na zakończenie zdań pojawiały słowa: „tygodniowy”, „weekendowy”, „godzinny”.

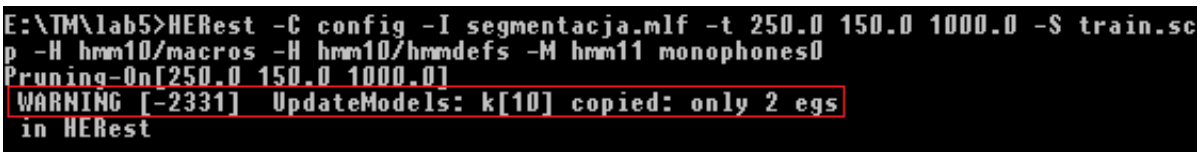
Na błędne rozpoznanie największy wpływ miała prawdopodobnie niepoprawna anotacja na poziomie fonemów, która została przeprowadzona automatycznie. Z uwagi na ilość fonemów występujących w nagraniu treningowym, trwającym ponad 3 minuty, nie możliwe było sprawdzenie każdego z fonemów po kolei. Dlatego do tego celu użyto programu (plik 3), który napisał Filip Hałon. Pozwalał on na odsłuchanie każdego fragmentu z nagrania treningowego, który został przypisany do danego fonemu. Przy użyciu tego programu możliwe było wygenerowanie nagrań zawierających określone fonemy, lub raczej to co zostało do nich zaanotowane.

```
[y, fs] = wavread('nagranie.wav');
fid = fopen('phonescheck.txt');
data = textscan(fid, '%f32 %f32 %s');
fclose(fid);

start = data{1, 1} .* 10 ^ -7 * fs;
meta = data{1, 2} .* 10 ^ -7 * fs;
phones = data{1, 3};
ph=0;

phone_search = find(ismember(phones, 'j')); % sprawdzany fonem 'j'

for i = 1 : 1 : length(phone_search)
    %wavplay(y(start(phone_search(i)) : meta(phone_search(i))), fs)
    ph=[ph;y(start(phone_search(i)) : meta(phone_search(i)))];
end
wavwrite (ph,fs,16, 'j'); % zapis nagrania z wszystkimi fonemami 'j'
```

PLIK 3 - PROGRAM NAPISANY W MATLABIE, UMOŻLIWIAJĄCY PRZETESTOWANIE ANOTACJI

Jak się okazało, niektóre fonemy trwają za długo i oprócz istotnej części zawierają też sąsiednie fonemy lub prawie całe słowa. Najgorzej wypadły fonemy „j", „dzi", „e" oraz „u". Niektóre z wygenerowanych nagrań zostały dołączone do folderu „błędy w anotacji".

Podczas pracy nad systemem w wierszu poleceń pojawiło się ostrzeżenie (rysunek 3) o tym, że fonem „k" występuje jedynie dwa razy w nagraniu treningowym, co jest zbyt małą liczbą, by mógł on z odpowiednim prawdopodobieństwem być rozpoznany prawidłowo. Spowodowane to było faktem, że fonem ten występuje tylko w jednym słowie znajdującym się w słowniku: „weekendowy". Co ciekawe słowo to, jak wspomniano wyżej, było często rozpoznawane przez system, mimo tego, że nie pojawiło się ani raz w nagraniach testowych.



RYSUNEK 3 - OSTRZEŻENIE O NIEWIELKIEJ ILOŚCI WYSTĄPIEŃ FONEMU "K"

5. ANALIZA RÓŻNYCH ROZWIĄZAŃ

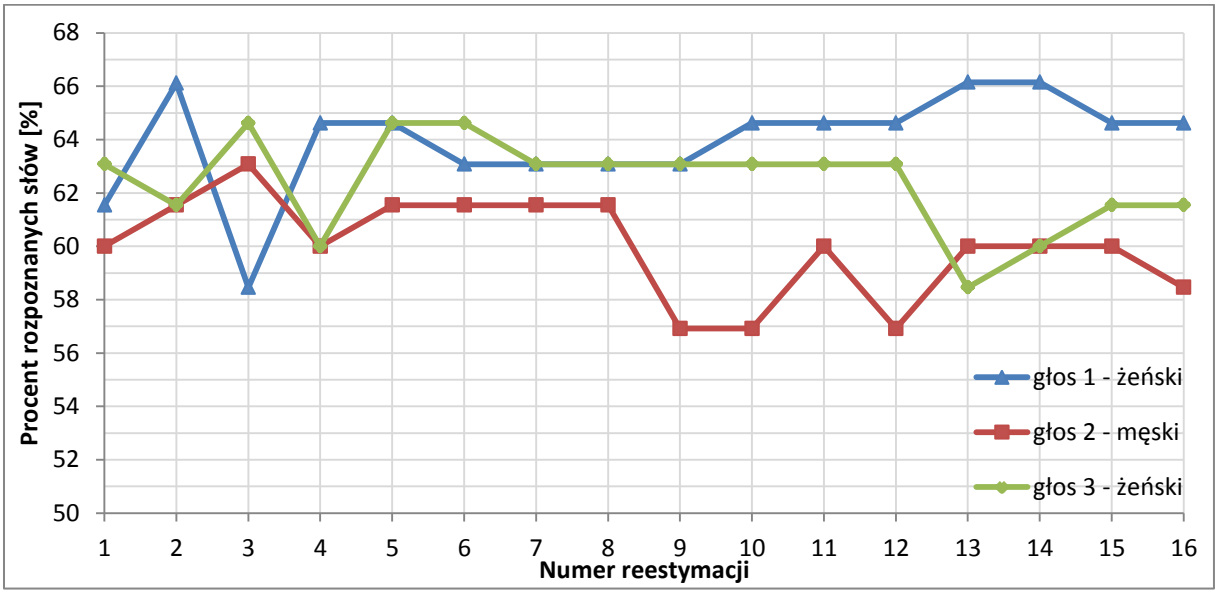
Z uwagi na to, że przypuszczalnie największy wpływ na niską rozpoznawalność ma zła anotacja na poziomie fonemów, sprawdzono co się stanie w przypadku usunięcia z pliku „segmentacja.mlf" tych, które trwają zbyt długo (zawierają w rzeczywistości więcej niż ten fonem, do którego zostały przypisane). Niestety, w przypadku tych fonemów, które sprawiają najwięcej problemów wszystkie są zaanotowane źle, a usunięcie ich sprawi, że system nie będzie działał w ogóle. Po usunięciu innych fonemów rozpoznawalność się nie poprawiła, więc powrócono do pierwotnej wersji pliku.

Następnie próbowano ręcznie poprawić ustawiony przedział czasowy dla problematycznych fonemów, ale było to czasochłonne i bardzo trudne w wykonaniu, więc po zmianie kilku linijek pliku „segmentacja.mlf" zaprzestano dalszego poprawiania.

Następnie postanowiono sprawdzić jak na rozpoznawalność wpłynie większa liczba reestymacji lub zmiana mówców w nagraniach testowych. Wyniki zamieszczono poniżej w tabeli 2 oraz na wykresach (rysunek 4 i 5).

TABELA 2 - WYNIKI UZYSKANE Z HRESULTS

nr reestymacji	głos 1 - żeński[2]		głos 2 – męski[3]		głos 3 - żeński[4]	
	procent rozpoznanych zdań [%]	procent rozpoznanych słów [%]	procent rozpoznanych zdań [%]	procent rozpoznanych słów [%]	procent rozpoznanych zdań [%]	procent rozpoznanych słów [%]
1	8,33	61,54	0	60	0	63,08
2	8,33	66,115	0	61,54	0	61,54
3	0	58,46	0	63,08	0	64,62
4	0	64,62	0	60	0	60
5	0	64,62	0	61,54	0	64,62
6	0	63,08	0	61,54	0	64,62
7	0	63,08	0	61,54	0	63,08
8	0	63,08	0	61,54	0	63,08
9	0	63,08	0	56,92	0	63,08
10	0	64,62	0	56,92	0	63,08
11	0	64,62	0	60	0	63,08
12	0	64,62	0	56,92	0	63,08
13	0	66,15	0	60	0	58,46
14	0	66,15	0	60	0	60
15	0	64,62	0	60	0	61,54
16	0	64,62	0	58,46	0	61,54



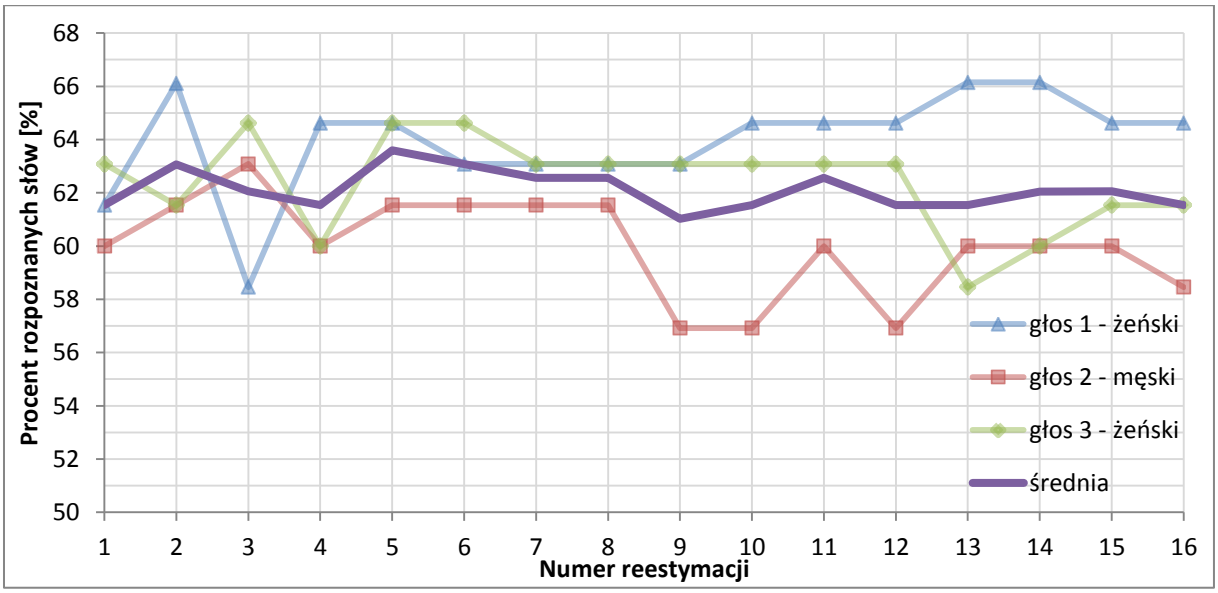
RYSUNEK 4 - WYNIKI UZYSKANE Z HRESULTS PRZEDSTAWIONE NA WYKRESIE

[2] głos 1 – głos podstawowy, wykorzystywany do nagrania treningowego
[3] głos 2 – głos używany jedynie w nagraniach testowych
[4] głos 3 – głos używany jedynie w nagraniach testowych

Po wzięciu pod uwagę większej liczby reestymacji, okazuje się, że dla 13 i 14 reestymacji wynik rozpoznania słów jest taki sam, jak dla drugiej, które do tej pory była najlepszą reestymacją. Jednak do wniosków takich można dojść jedynie biorąc pod uwagę głos podstawowy, wykorzystany do nagrania treningowego. W przypadku głosu 2 powyżej 8 reestymacji pojawia się duży spadek rozpoznawalności i spore wahania wyników. Dla głosu 3 spadek pojawia się dopiero przy 12 reestymacji.

Ponadto mimo tego, że głosy 2 i 3 są „nowe" dla systemu ich rozpoznawalność jest porównywalna z głosem podstawowym i do 8 reestymacji wyniki są dość dobre (powyżej 60%). Dla 3 reestymacji wyniki innych głosów są nawet lepsze niż te uzyskane dla podstawowego.

Z powyższych danych można także zauważyć, że głosy żeńskie są rozpoznawane zdecydowanie lepiej. Wynikać to może z podobnej częstotliwości tonu podstawowego w przypadku kobiet.



RYSUNEK 5 - WYNIKI UZYSKANE Z HRESULTS, Z UWZGLĘDNIENIEM WARTOŚCI ŚREDNIEJ DLA WSZYSTKICH GŁOSÓW

Na powyższym wykresie (rysunek 5) zamieszczona została średnia z wyników uzyskanych dla trzech mówców. Dzięki temu można zauważyć, że optymalna reestymacja dla wszystkich głosów to reestymacja 5 – średnia dla niej wynosi powyżej 63%.

Ciekawe jest również to, że głosy 2 oraz 3 uzyskały najwyższy wynik przy 3 reestymacji, podczas gdy głos podstawowy uzyskał tam wynik najniższy. Co oznacza, że gdyby rozważać wyniki uzyskane jedynie przez głosy inne niż podstawowy, to 3 reestymacja byłaby najlepszą.

W przypadku głosów 2 i 3 sytuacja z nierozpoznanymi słowami jest podobna, jak dla głosu podstawowego – większość z nich system zamienił na inne, ok. 5 słów w każdej komendzie zostało usuniętych i ani jedno nie zostało wstawione.

6. WNIOSKI

Rozpoznawalność na poziomie ok.60% jest niewystarczająca dla tego typu systemu, szczególnie, że nie rozpoznaje on końcówek komend, w związku z czym nie jest w stanie odczytać o jaki dokładnie typ biletu chodzi mówcy.

Do usprawnienia systemu konieczna jest poprawa anotacji na poziomie fonemów. Dopiero potem można byłoby przeprowadzać analizy rozpoznawalności i szukać innych, ewentualnych błędów. Na problemy z anotacją mogło mieć wpływ zastosowanie systemu z anotowaniem ciszy na końcu zdań w nagraniu treningowym, podczas gdy w anotacji na poziomie fonemów nie została ona uwzględniona.

Można również zauważyć, że nie jest potrzebna większa liczba reestymacji niż 10, ponieważ od pewnego momentu dla wyższej wartości reestymacji spadała efektywność systemu. W przypadku tego systemu wystarczyło już 5 reestymacji, aby uzyskać najwyższy wynik.

Co jest zadziwiające rozpoznawalność innych mówców, niż ten którego głos wykorzystano do stworzenia systemu, jest równie wysoka. Jednak na to może mieć wpływ fakt, że mówcy byli w podobnym wieku, więc nie wiadomo jak rozpoznawane byłyby osoby młodsze lub starsze.

Część każdej komendy stanowiły te same słowa „poproszę bilet”, co zwiększyło poziom rozpoznawalności słów. Gdyby gramatyka w tym systemie była bardziej skomplikowana i ilość stałych słów mała lub wręcz zerowa, efektywność byłaby zdecydowanie niższa.

Aby usprawnić system należałoby też przygotować większą liczbę nagrań treningowych, aby w szczególności słowa kończące zdania były powiedziane więcej niż dwukrotnie.

Gdyby przygotowana została większa liczba nagrań testowych, można byłoby uzyskać więcej wniosków odnośnie rozpoznawalności. Być może okazałoby się też, że słowa ze słownika, które nie były używane w nagraniach testowych są rozpoznawalne dobrze, przez co skuteczność systemu byłaby większa.

Co do samej gramatyki, to można byłoby wprowadzić większą dynamikę, aby przystosować system do tych mówców, którzy nie są zaznajomieni z gramatyką. Zamiast słowa „poproszę” można byłoby użyć słowa „proszę” lub w ogóle je pominąć. Aby ułatwić komunikację systemu z osobami korzystającymi z niego można byłoby także skrócić najczęściej występujące komendy, np. zamiast „bilet strefowy normalny jednoprzejazdowy” wprowadzić komendę: „bilet normalny” lub nawet „normalny”, ponieważ bez uściślania z jakiej strefy i jakiego typu ma to być bilet, kupujący ma zazwyczaj na myśli bilet strefowy jednoprzejazdowy – jest to najczęściej kupowany rodzaj biletu.