EAIiE, AGH	**Technologia Mowy – Laboratorium**	Bartosz Nowacki Paweł Synowiec
Raport końcowy	Budowa i testy systemu rozpoznawania mowy opartego o oprogramowanie HTK	Ocena:

System miał za zadanie rozpoznawać treści zamówień telefonicznych w lokalach podających pizzę i napoje. Gramatyka zakłada możliwość wybrania numeru pizzy z oferty, rozmiaru, opcjonalnie sosu do pizzy, oraz ilość napojów i ich rodzaje. System zakłada podawanie np. numeru pizzy w specyficzny sposób (cyframi, tj. 'jeden dwa' zamiast 'dwanaście'), co z jednej strony jest mało intuicyjne dla użytkownika, natomiast znacznie upraszcza budowę systemu i mocno redukuję ilość spodziewanych słów.

Wyniki z HResults

Do testów wykorzystano 2 zestawy po 7 plików z zarejestrowanymi zdaniami spełniającymi gramatykę. W obu zestawach te same zdania testowe zarejestrowali różni mówcy.

Najlepszy wynik dla mówcy pierwszego:

```
------------------------ Sentence Scores ------------------------
====================== HTK Results Analysis =====================
                  Date: Mon May 09 14:35:35 2011
                       Ref : testref.mlf
                      Rec : out1/recout1.mlf
-------------------------- File Results --------------------------
 test2.rec:   27.27( 27.27)  [H=   3, D=  0, S=  8, I=  0, N= 11]
 test3.rec:   60.00( 20.00)  [H=   6, D=  0, S=  4, I=  4, N= 10]
 test4.rec:   50.00( 41.67)  [H=   6, D=  0, S=  6, I=  1, N= 12]
 test5.rec:   36.36( 27.27)  [H=   4, D=  0, S=  7, I=  1, N= 11]
 test6.rec:   33.33( 33.33)  [H=   4, D=  3, S=  5, I=  0, N= 12]
 test7.rec:   45.45(  9.09)  [H=   5, D=  2, S=  4, I=  4, N= 11]
 test8.rec:   46.15( 46.15)  [H=   6, D=  1, S=  6, I=  0, N= 13]
------------------------ Overall Results ------------------------
               SENT: %Correct=0.00 [H=0, S=7, N=7]
    WORD: %Corr=42.50, Acc=30.00 [H=34, D=6, S=40, I=10, N=80]
```

Najlepszy wynik dla mówcy drugiego:

```
------------------------ Sentence Scores ------------------------
====================== HTK Results Analysis =====================
                  Date: Mon May 09 14:35:37 2011
                       Ref : testref2.mlf
                      Rec : out2/recout11.mlf
------------------------- File Results --------------------------
 test22.rec:   27.27( 27.27)  [H=   3, D=  2, S=  6, I=  0, N= 11]
 test33.rec:   50.00( 50.00)  [H=   5, D=  1, S=  4, I=  0, N= 10]
 test44.rec:   41.67( 41.67)  [H=   5, D=  2, S=  5, I=  0, N= 12]
 test55.rec:   27.27( 27.27)  [H=   3, D=  2, S=  6, I=  0, N= 11]
 test66.rec:   16.67( 16.67)  [H=   2, D=  3, S=  7, I=  0, N= 12]
 test77.rec:   45.45( 45.45)  [H=   5, D=  2, S=  4, I=  0, N= 11]
 test88.rec:   15.38( 15.38)  [H=   2, D=  4, S=  7, I=  0, N= 13]
------------------------ Overall Results ------------------------
               SENT: %Correct=0.00 [H=0, S=7, N=7]
    WORD: %Corr=31.25, Acc=31.25 [H=25, D=16, S=39, I=0, N=80]
================================================================
```

Najlepszy wynik dla pierwszego mówcy osiągnięto dla 1tej reestymacji modelu Markowa, dla drugiego mówcy także dla 1szej reestymacji.

Opis użytych nagrań

Nagranie testowe pierwszego mówcy zarejestrowano podczas laboratorium mikrofonem komputerowym firmy Logitech, przy dość dużym poziomie hałasu tła. Nagrania drugiego mówcy zarejestrowano w mieszkaniu, we względnej ciszy, również mikrofonem komputerowym do Skypa. W obu wypadkach rejestrowano polecenia wydawane mową ciągłą, z dłuższymi odstępami pomiędzy frazami. Model został wytrenowany dla obu mówców jednocześnie, jako jeden zestaw danych testowych.

Nagrania treningowe rejestrowano mikrofonem elektretowym podłączonym do komputera, w ciszy.
Czas nagrania treningowego: **00:03:33 – połączone nagrania obu mówców**
Czas nagrań testowych: **00:01:24 – 1szy mówca**
00:00:39 – 2gi mówca

Nazwy katalogów z nagraniami

Nagranie treningowe: /train/testowy2.wav

Nagrania pierwszego mówcy: /test1/test*.wav

Nagrania drugiego mówcy: /test2/test**.wav

Opis gramatyki

Gramatyka została zaprojektowana jako ciąg dwóch prostych zdań występujących w ściśle określonej kolejności: zamówienie pizzy poprzez podanie jej numeru w menu, rozmiaru, zamawianego sosu, następnie podanie ilości kubków, rodzaju napoju i rozmiaru kubka. Gramatyka została maksymalnie uproszczona, aby ułatwić analizę i rozpoznanie zamówienia. Jednocześnie skutkuje to koniecznością wypowiadania ściśle określonych zdań gdzie jedyną decyzją użytkownika systemu jest zmiana parametrów takich jak numer, rozmiar czy rodzaj napoju. Konstrukcja i kolejność wyrazów w zdaniu musi być zachowana, ale jego gramatyka jest zbliżona do naturalnej wypowiedzi kierowanej np. w restauracji do kelnera.

Analiza błędów rozpoznania

Dla pierwszego mówcy źle rozpoznają się słowa „Proszę, chciałbym, zamawiam", gdzie w jednym rozpoznaniu wszystkie są rozpoznane jako proszę, a w innym jako zamawiam. Dla najlepszego wyniku nie rozpoznano słowa „zamawiam". Mylone są także cyfry, błędnie rozpoznane jako jeden, siedem lub osiem. Często zamieniane są wyrazy mała z mały, duża z duży, co jest spowodowane zbyt małą liczbą danych treningowych. Mimo tego, ten błąd nie wpływa na poprawność rozpoznania dla tego zestawu słów. W niewielkiej liczbie przypadków nieprawidłowo rozpoznano rozmiar, w szczególności słowo „mega" nie zostało rozpoznane ani razu. W połowie nagrań testowych została użyta alternatywna wersja gramatyki poprzez zamianę kolejności fraz z podaniem rozmiaru i zamówieniem sosu. Spowodowało to większość błędów rozpoznania. Wraz ze wzrostem liczby reestymacji modelu ilość prawidłowo rozpoznanych wyrazów spadała w przypadku obu mówców, co świadczy o różnicach w brzmieniu próbek treningowych i testowych. Pomocne było by także zwiększenie długości nagrań treningowych.

Już więcej niż jedna reestymacja powodowała spadek skuteczności systemu, przetrenowanie modelu następowało bardzo szybko, zapewne ze względu na słabej jakości nagrania treningowe. Dla drugiej reestymacji dla mówcy pierwszego i drugiego rozpoznawalność spadła z odpowiednio 42.5% i 31.25% na:

- 37.5% oraz 28.75% dla dwóch reestymacji
- 27.5% oraz 25% dla trzeciej reestymacji
- 35% oraz 26.25% dla czterech reestymacji
- 31.25% oraz 25% dla pięciu reestymacji
- 33.75% 22.5% dla sześciu reestymacji
- 28.75% oraz 22.5% dla siedmiu reestymacji

Jak widać, dla trzech reestymacji poprawność rozpoznania spadła, dla pierwszego mówcy nawet o 10%. Po tym spadku, widać poprawę, przetrenowanie modelu następuje po szóstej reestymacji, jednak nawet w najbardziej optymalnym przypadku, tj. dla piątej – szóstej reestymacji, skuteczność systemu jest niższa, stąd też zdecydowaliśmy na pozostanie przy pierwszej reestymacji.

Zaburzona zależność skuteczności od ilości reestymacji i fakt, że system posiada maksymalną skuteczność dla pierwszej reestymacji, wynika m.in. z wyżej wymienionych przyczyn, tj. głównie słabej jakości nagrań, małej ich ilości i być może złego sposobu wymawiania zdań treningowych (np. niewyraźnie, nienaturalnie, bez wyraźnych odstępów pomiędzy wyrazami)

Wyrażamy zgodę na włączenie nagrań treningowych i testowych do Korpusu Mowy AGH

Załączone pliki: gram.txt, dict.txt, config1 pliki wav i lab testowe, pliki lab, wav i mlf treningowe