Inżynieria Akustyczna – Technologia Mowy, rok 2012

Paulina Bugiel, Sławomir Raab

Sprawozdanie z laboratoriów HTK

1. Opis gramatyki i jej uzasadnienie jako realizacja jakiejś wirtualnej potrzeby wraz z ocenę naturalności i dostosowania się do użytkownika:

Nasz projekt ma za zadanie realizować automatyczne zamawianie pizzy. Przyjęliśmy sytuację, w której dostępne jest 12 rodzajów pizzy oraz opcjonalnie 3 sosy. Każda pizza ma swoją nazwę oraz przypisany numer w menu, można ją zamówić w jednym z trzech różnych rozmiarów.

Lista słów użytych w naszym systemie wraz z transkrypcją fonetyczną według słownika Grocholewskiego znajduje się w pliku slownik1.txt

Poniżej przedstawiamy konstrukcję zdania użytą w projekcie:

(SENT-START ($rozmiar $nazwa [sos $sos]) | ($nazwa $rozmiar [sos $sos]) | ($rozmiar [pizza] numer $numer [sos $sos]) | ([pizza] numer $numer $rozmiar [sos $sos]) SENT-END)

Powyższa deklaracja zdania pokrywa się z najczęściej stosowanymi zwrotami przy składaniu zamówienia. Przykładowe zdania zgodne z powyższą gramatyką:

- *Duża Margherita sos pomidorowy*
- *Napolitana mała sos pomidorowy*
- *Duża pizza numer dwa sos tabasco*
- *Pizza numer siedem mała*

Dzięki powyższej konstrukcji, klient nie musi uczyć się konkretnej składni, a swobodnie dobierać wypowiadane zdania.

Jedynym zastosowanym uproszczeniem, jest uniknięcie odmiany słów przez przypadki, co mogłoby zmniejszyć rozpoznawalność oraz zbytnio skomplikować konstrukcję systemu.

2. Opis nagrań:

Nagrania zostały wykonane przy użyciu mikrofonu ShureSM58 z zastosowaniem pop filtru oraz karty dźwiękowej Focusrite Scarlett 2i2. Nagrania zostały przeprowadzone w umeblowanym pokoju. Rejestrowana była mowa ciągła. Łączny czas nagrań treningowych wynosi około 4,5 minuty. Nagrywany był głos kobiecy.

3. Warunki testowania i wyniki testów z HResults:

Nagrania testowe nie są częścią bazy użytej do trenowania modeli. Zostały one nagrane przez tą samą osobę, w tych samych warunkach i za pomocą tego samego sprzętu co nagrania treningowe. Baza testowa składa się z jedenastu zdań o różnej długości i składni. Łączna liczba słów wynosi 41. Najlepsze wyniki rozpoznawalności mowy uzyskaliśmy dla ósmej reestymacji i są one następujące:

- Rozpoznawalność zdań: 27,27%
- Rozpoznawalność słów: 56,1%

4. Analiza błędów rozpoznania, (które słowa lub frazy gorzej się rozpoznają, z hipotezą dlaczego akurat te):

Z 41 słów 23 zostały poprawnie rozpoznane, wystąpiło 5 błędów usunięcia oraz 13 podstawienia słowa. Największy problem stanowiło rozpoznanie pierwszego słowa w zdaniu testowym. Możliwe, że jest to spowodowane szumami na początku nagrań. Niezbyt dobrze były rozpoznawane niektóre nazwy pizzy np. Capriciosa lub Califfo. System nie poradził sobie również z rozpoznaniem słowa ‘czosnkowy’ oraz frazy ‘pizza numer’. Słowo ‘czosnkowy’ mogło nie być dostatecznie dobrze wytrenowane, jako że pojawiło się tylko w kilku zdaniach treningowych. Również słowo ‘pomidorowy’ jest do niego bardzo podobne, co mogło spowodować błędy. C o do reszty błędów, nie jesteśmy pewnie dlaczego wystąpiły.

5. Analiza różnych rozwiązań - różna liczba re estymacji oraz obcy mówca:

Dokonaliśmy porównania rozpoznawalności dla kilku reestymacji (od 3 do 15). Rozpoznawalność była optymalna dla ośmiu reestymacji, a kolejne powodowały przetrenowanie modelu i spadek rozpoznawalności. Poniżej znajdują się wyniki uzyskane dla kilku różnych reestymacji (dla tej samej bazy nagrań testowych co w pkt 3):

- hmm3:
  - Rozpoznawalność zdań - 27,27%
  - rozpoznawalność słów - 48,78%
- hmm5
  - Rozpoznawalność zdań - 27,27%
  - rozpoznawalność słów - 48,78%
- hmm7
  - Rozpoznawalność zdań – 18,18%
  - rozpoznawalność słów – 43,9%
- hmm8
  - Rozpoznawalność zdań – 27,27%
  - rozpoznawalność słów – 56,1%
- hmm9
  - Rozpoznawalność zdań – 27,27%
  - rozpoznawalność słów – 53,66%

- hmm10
    - Rozpoznawalność zdań – 27,27%
    - rozpoznawalność słów – 53,66%
- hmm11
    - Rozpoznawalność zdań – 27,27%
    - rozpoznawalność słów – 53,66%
- hmm15
    - Rozpoznawalność zdań - 27,27%
    - rozpoznawalność słów - 48,78%

Sprawdzenie rozpoznawalności dla obcego mówcy:

Nagrania zostały wykonane w tych samych warunkach i za pomocą tego samego sprzętu co nagrania treningowe, lecz przez mężczyznę. Treść nagrań testowych w tym przypadku była taka sama jak poprzednio. Najlepszą rozpoznawalność osiągnęliśmy również dla hmm8 i wynosiła ona:

- Rozpoznawalność zdań – 27,27%
- rozpoznawalność słów – 56,1%

Sprawdzaliśmy również rozpoznawalność dla kilku wytrenowań modelu. Uzyskane wyniki testów są prawie takie same jak przy testowaniu głosem kobiecym, z jednym wyjątkiem. Dla hmm5 uzyskaliśmy następujące wyniki:

- Rozpoznawalność zdań - 27,27%
- rozpoznawalność słów – 51,22%

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