Inżynieria Akustyczna – Technologia Mowy, rok 2013

Sebastian Dziadzio

Sprawozdanie z laboratoriów HTK

1. Opis projektu i zastosowanej gramatyki

Projekt polegał na stworzeniu systemu rozpoznawania mowy, który w uproszczony sposób symulował sterowanie samochodem. Zaimplementowane funkcje to: zwiększanie i redukcja biegu, zmiana prędkości o zadaną wartość, skręt o zadany kąt, włączanie lub wyłączanie świateł, silnika i tempomatu. Podstawowym kryterium podczas projektowania gramatyki była jej naturalność pod względem językowym. W tym celu konieczne było na przykład ograniczenie dokładności przy zmianie prędkości czy skręcie do pięciu jednostek. Udało się jednak otrzymać gramatykę nie wymagającą od użytkownika specjalnych przygotowań do pracy z systemem. Komendy są intuicyjne i poprawne językowo, przykładowo: „przyspiesz (trzydzieści kilometrów/mil)", „zwiększ/zredukuj bieg", „(wyłącz/włącz) światła drogowe", „skręć (dwadzieścia stopni) w lewo".

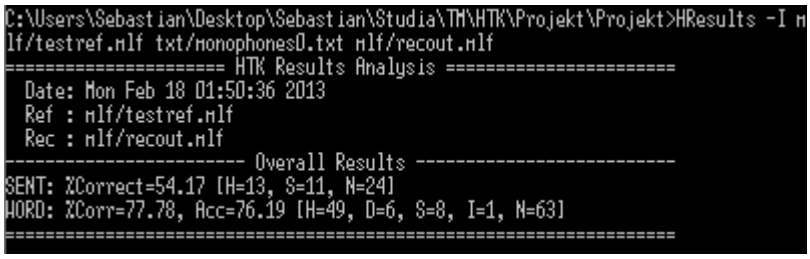
2. Opis nagrań

Nagrania sporządzono w pokoju o powierzchni 15m^2 za pomocą mikrofonu wbudowanego w kamerkę internetową Logitech C310, wyposażonego w funkcję redukcji szumów. Hałas tła był znikomy, najgłośniejszym słyszalnym zakłóceniem był pracujący komputer stacjonarny. Rejestrowano mowę ciągłą z przerwami pomiędzy kolejnymi wypowiedziami. Następnie przekonwertowano pliki na format mono, podpróbkowano do 16 000 Hz i zapisano jako waveform audio file. Operacje te wykonano z użyciem programu Audacity. Początkowy łączny czas nagrań treningowych to 3 minuty i 26 sekund. Ze względu na stosunkowo małą skuteczność działania programu, inkorporowano nagrania pochodzące z innego systemu. Wówczas czas nagrań treningowych wzrósł do 6 minut i 41 sekund. Nagrania testowe sporządzono w tych samych warunkach. Nagrania w folderach *test, test_snr20, test_snr30 i test_snr40* pochodzą od tego samego mówcy co nagrania treningowe, przy czym w trzech ostatnich folderach zostały one sztucznie zniekształcone za pomocą pakietu Matlab poprzez dodanie szumu białego o stosunku sygnału do szumu wartości odpowiednio 20, 30 i 40 dB. Folder *test_inny_mowca* zawiera nagrania sporządzone przez innego mowcę.

3. Wyniki testów

Początkowo system przetestowano za pomocą dwudziestu czterech nagrań testowych pochodzących od tego samego mówcy co nagrania treningowe. Wykonano 20 reestymacji i nie uwzględniono modelu ciszy. System nie działał poprawnie, rozpoznanie słów było znikome (poniżej 10%). Po zamodelowaniu ciszy (jako osobnego fonemu sp), RR% dla słów wzrósł do około 20%. Ze względu na tak niski wynik, zdecydowano się na wcielenie nagrań treningowych pochodzących z innego projektu (dzięki uprzejmości Krzysztofa Drewnianego), czego śladem jest plik *phones0_dr.mlf* w folderze *mlf*. Wbrew oczekiwaniom, nie wpłynęło to znacznie na skuteczność systemu. Podjęto próbę zmiany parametrów *word insertion penalty* oraz *grammar scale factor*, jednak nawet znaczne zmiany ich wartości pozostawały bez wpływu na działanie programu, więc zaprzestano na kilkunastu próbach. System nie działał nawet gdy testowano go bezpośrednio z wykorzystaniem nagrań treningowych,

więc skonstruowano go od nowa. Zrezygnowano z wykorzystania dodatkowych nagrań treningowych. Wykryto przy tym nieznaczne błędy w gramatyce oraz nieprawidłowe parametry w nagraniach testowych (ich częstotliwość próbkowania wynosiła 44100 Hz zamiast 16000 Hz). Po podpróbkowaniu oraz zmianie rozdzielczości bitowej, RR% dla słów przekroczył 30%. Kluczowa dla poprawy skuteczności działania programu była zmiana liczby reestymacji. Najprawdopodobniej doszło do przetrenowania modeli, gdyż okazało się, że przyjęcie danych otrzymanych po dwóch reestymacjach skutkuje znacznie lepszym działaniem programu niż wykorzystanie modeli otrzymanych po piętnastu czy dwudziestu reestymacjach. W ten sposób udało się otrzymać wynik 77%:



Współczynnik rozpoznania dla zdań jest stosunkowo wysoki (54%), co wynika z faktu, że zdania w systemie są bardzo krótkie, składają się maksymalnie z pięciu-sześciu słów. Łącznie wykonano testy z użyciem 120 nagrań testowych, o różnych parametrach akustycznych. Zostało to bardziej szczegółowo omówione w podpunkcie piątym.

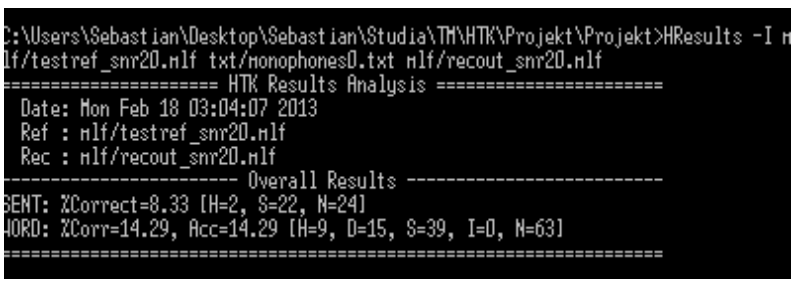
4. Analiza błędów rozpoznania

Najczęstsze zaobserwowane błędy w działaniu systemu to zamiana wyrazów o podobnej strukturze fonetycznej lub identycznej funkcji gramatycznej, np. włącz-wyłącz, trzydzieści-czterdzieści, pięćdziesiąt-sześćdziesiąt, zwolnij-przyspiesz, lewy-prawy. Można poradzić sobie z tym problemem na etapie projektowania gramatyki, poprzez unikanie wyrazów o podobnym brzmieniu, jeśli mają pełnić w zdaniu tę samą funkcję (na przykład *włącz* można zastąpić słowem *uruchom*). Niestety zmniejsza to naturalność gramatyki oraz intuicyjność interfejsu głosowego, która jest przecież jedną z jego podstawowych zalet.

5. Analiza różnych rozwiązań

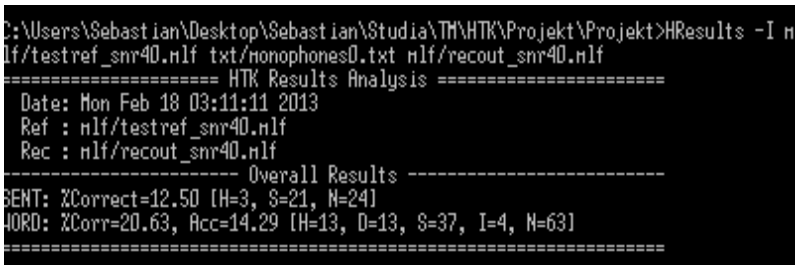
Jak napisano w punkcie trzecim, łączna liczba nagrań testowych to 120: 24 nagrania mówcy A, 24 nagrania mówcy A z szumem o SNR 20 dB, 24 nagrania mówcy A z szumem o SNR 30 dB, 24 nagrania mówcy A z szumem o SNR 40 dB oraz 24 nagrania mówcy B. Poniżej zaprezentowano wyniki oraz ich zgrubną analizę w zależności od liczby reestymacji.

a) SNR 20dB

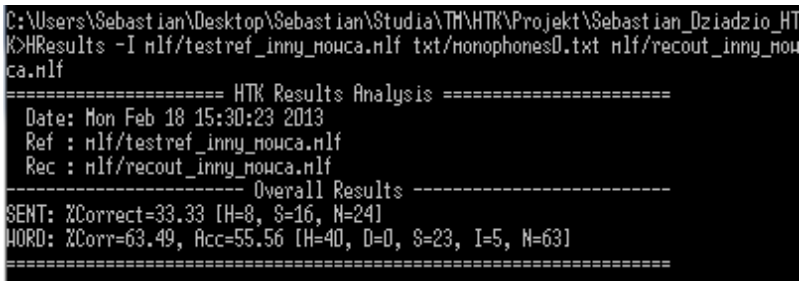


b) SNR 30 dB

c) SNR 40 dB



d) Mówca B



Dla zaszumionych nagrań najlepszy współczynnik rozpoznania osiągnięto dla trzech reestymacji. Przy dziesięciu, piętnastu i dwudziestu skuteczność systemu gwałtownie spadała i wynosiła poniżej dziesięciu procent. Dla obcego mówcy optymalną liczbą reestymacji było sześć. Dla dziesięciu, piętnastu i dwudziestu skuteczność nieznacznie spadała (do około 50%), natomiast dla dwóch, trzech i czterech wynosiła poniżej 20%.

6. Wnioski

Działanie systemu w bardzo dużym stopniu zależy od warunków akustycznych. Ze względu na ewentualne zastosowanie tego typu oprogramowania, skuteczność rozpoznania zdań powinna wynosić co najmniej 95% dla różnych mówców, nawet w złych warunkach akustycznych (duża ilość zakłóceń). Odstęp sygnału od szumu rzędu 40 dB jest zapewne trudny do uzyskania we wnętrzu samochodu (pracujący silnik, hałas aerodynamiczny, muzyka, inni mówcy) a nawet dla tak "komfortowych" warunków akustycznych skuteczność systemu wynosi zaledwie 20%. W idealnych warunkach udało się uzyskać 77 % rozpoznawalności słów. Przy tak mało zróżnicowanym zbiorze treningowym (zaledwie jeden mówca) dość zaskakujący był stosunkowo niewielki spadek efektywności przy zmianie użytkownika. Należy jednak zaznaczyć, że mówca B to osoba o podobnej skali głosu, akcencie i intonacji oraz tej samej płci (brat mówcy A). Do stworzenia systemu niezależnego od mówcy potrzebny byłby dużo bardziej reprezentatywny korpus audio. Główną

przyczyną niewielkiej rozpoznawalności (zaledwie 77%) jest zapewne bardzo zła jakość nagrań treningowych. Jak okazało się podczas drugiego podejścia do stworzenia systemu, prawie wszystkie nagrania treningowe są zniekształcone (doszło do przesterowania mikrofonu). Również sprzęt rejestrujący nie był zbyt dobrej jakości.

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