Inżynieria Akustyczna – Technologia Mowy, rok 2013

**Marcin Pasternak**

# Sprawozdanie z laboratoriów HTK

## 1. Opis gramatyki – Symulator kierowcy rajdowego

Postanowiłem podjąć się zadania napisania systemu, który być może byłby przydatny do gier wyścigowych w przypadku prawdziwych fanatyków tego gatunku. Miał on na celu rozpoznawać najbardziej charakterystyczne i najczęściej spotykane komendy wydawane przez pilota rajdowego, w którego miałby się wcielać gracz, i za ich pośrednictwem pokonywać poszczególne trasy z komputerem jako kierowcą. Niestety, ciężko tutaj mówić o jakiejkolwiek naturalności – rola pilota rajdowego polega na wydawaniu szybkich komend, które zostaną szybko i poprawnie rozpoznane przez kierowcę rajdowego i przełożą się na jego odpowiednią reakcję. Wymagana jest przynajmniej średniozaawansowana wiedza z tego zakresu. Poniżej plik gram.txt, ukazujący gramatykę:

*$ost = ostroZnie | uwaga ;*

*$odl = sto | siedemdziesiAt | piECdziesiAt | trzydzieSci ;*

*$trud = szeSC | piEC | cztery | trzy | dwa ;*

*$jaki = otwarty | dLugi ;*

*$rodzaj = lewy | prawy ;*

*$dodatek = tnij | wAsko | zaC ;*

*$info = dziury | wybija | syf | szczyt | hopa ;*

*$stastop = start | stop ;*

*( SENT-START ( ($stastop) | ($ost ($odl | $trud) $rodzaj zakrEt [bardzo] $jaki [nie] [$dodatek] [$info]) | ($ost $odl prosta) | ($ost ($odl | $trud) $rodzaj nawrOt [nie] [$dodatek] [$info]) SENT-END ))*

Komendy start/stop służyłyby do odpowiednio rozpoczęcia i zakończenia jazdy.

Istnieje naprawdę ogromna ilość kombinacji, które można utworzyć (przykładowe prawdopodobieństwo wypadnięcia określonej sekwencji wynosi 1:4320). Spowodowane jest to tym, że program zamiast analizować 2-3 wyrazowe sekwencje musi dokonywać analizy nieraz bardzo rozbudowanego zdania. Byłem zmuszony utworzyć nieco zubożoną wersję, nastawioną na konkretne sytuacje w celu podniesienia progu rozpoznawalności. O wynikach będzie jeszcze mowa w dalszej części sprawozdania.

2**. Opis nagrań**

Nagrania dokonano , w przypadku nagrywania treningowego oraz testowego autorskiego, przy użyciu mikrofonu pochodzącego z zestawu słuchawkowego firmy Creative z serii Fatal1ty. Parametry mikrofonu:

-Redukcja szumów oraz piankowa osłona mikrofonu pozwalająca na przyzwoitą filtrację niepożądanych dźwięków;
- Pasmo przenoszenia – 100Hz – 18kHz
- Impedancja - <2.2 kOhm
- Czułość (1kHz) - -39dBV/Pa
- Pojedynczy kabel z miedzi OFC, pokryte złotem wtyki

Na uwagę zasługuje fakt, że w przypadku moich nagrań testowych używałem innego komputera, na którym musiałem użyć sprzętowego wzmocnienia mikrofonu +20dB.

W przypadku obcych nagrań testowych autorstwa Adriana Sekuły, użyto mikrofonu firmy Tracer o bliżej nieokreślonej specyfikacji.

Nagrania odbywały się w pokoju w kamienicy, przy bezwzględnej ciszy (w przypadku nagrań treningowych i obcych nagrań testowych) oraz przy szumie i lekkim przeziębieniu (w przypadku moich nagrań testowych). W obu przypadkach słowa były izolowane, łączny czas nagrań treningowych wynosił 3:12.

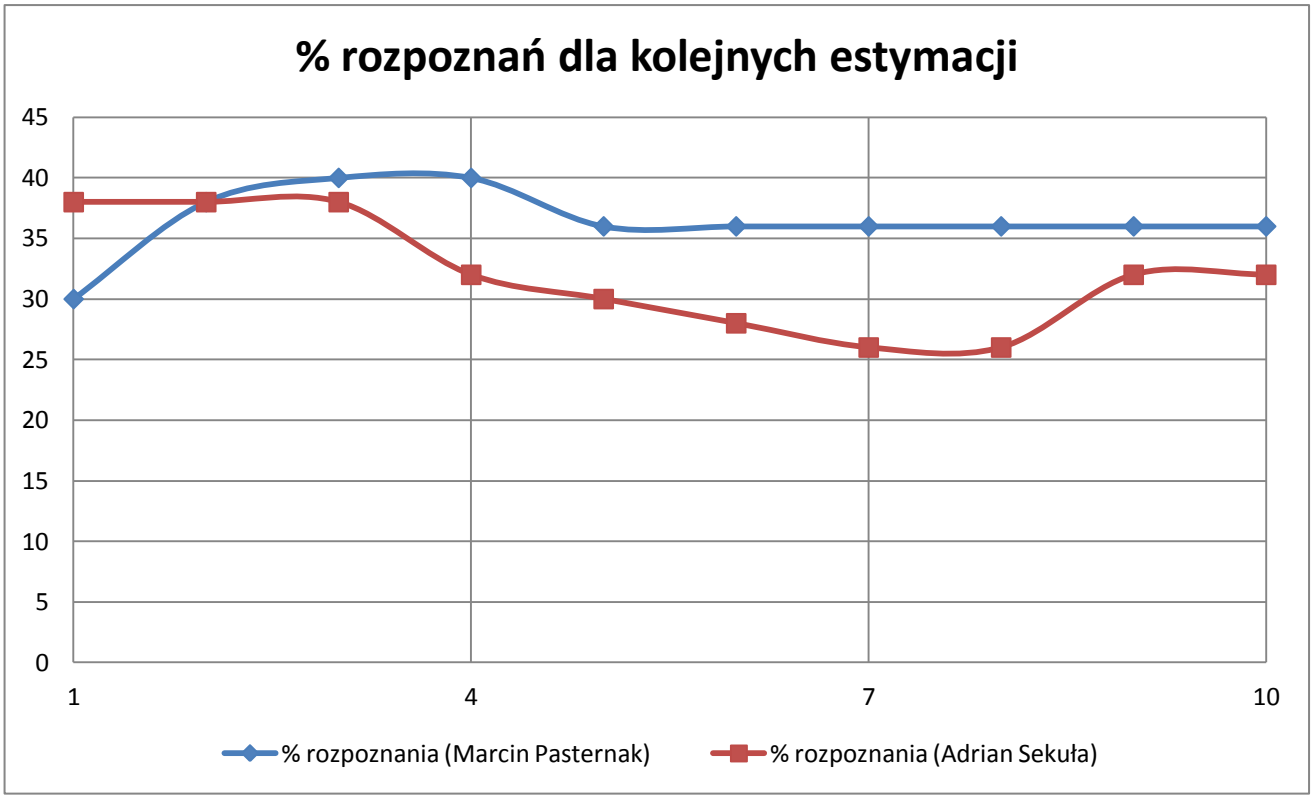
W przypadku nagrań testowych sporządzono 11 zdań – komend.

3. W**yniki testów z HResults**

Poniżej w tabeli zestawiono wyniki rozpoznania dla dwóch mówców na podstawie 11 zdań testowych dla kolejnych estymacji:

| nr estymacji | % rozpoznania (Marcin Pasternak) | % rozpoznania (Adrian Sekuła) |
|---|---|---|
| **1** | 30 | **38** |
| **2** | **38** | **38** |
| **3** | **40** | 38 |
| **4** | **40** | 32 |
| **5** | **36** | 30 |
| **6** | **36** | 28 |
| **7** | **36** | 26 |
| **8** | **36** | 26 |
| **9** | **36** | 32 |
| **10** | **36** | 32 |

Oraz wykres:




4. **Analiza błędów rozpoznania oraz wnioski**

Najczęściej rozpoznawano słowa ostrzeżenia (ostrożnie, uwaga). Najbardziej wrażliwym punktem programu był zawsze trzeci bądź czwarty wyraz (określenie, czy jedziemy prosto, mamy przed sobą nawrót albo zakręt). Jeśli program je błędnie rozpoznał, wtedy przeważnie cała komenda ulega przekłamaniu, co jest znaczącą wadą tego systemu. Dlatego też ostatnie słowa, które mają szanse paść (z grupy $info oraz $dodatek), z uwagi na ich miejsce występowania i ilości kombinacji, są rozpoznawane najgorzej.

Z pewnością zabrakło tutaj dodatkowych kilku godzin nagrań treningowych, dzięki którym system miałby zdecydowanie większe szanse na poprawne działanie. Poziom skomplikowania systemu (komendy nawet do sekwencji 8 wyrazów!) jest niewspółmierny do ilości nagrań testowych. Jednak według mnie, 40% rozpoznawania nie jest najgorszym wynikiem, ale na pewno nie jest to wynik zadowalający. W przyszłości system mógłby wymagać m.in. zwiększenia ilości nagrań treningowych oraz ulepszonej gramatyki i rozszerzenia słownika.

Jak widać na załączonych wynikach,  osiągnąłem lepsze rezultaty niż mój kolega Adrian. Może być to spowodowane używanym przeze mnie sprzętem (zgodnym z tym, na którym nagrywane były nagrania treningowe). Wydaje mi się, że gdybym nagrywał dane testowe w dobrym stanie zdrowotnym, wynik mógłby być lepszy. Z nagrań Adriana można też zauważyć akcentowanie świadczące o końcu komendy (obniżanie głosu), podczas, gdy moje nagrania testowe, jak i treningowe, są tego pozbawione. Głos Adriana na nagraniach jest zdecydowanie bardziej naturalny, co było zamierzonym zabiegiem.

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