Inżynieria Akustyczna – Technologia Mowy 2013

Jakub Antoniuk

Sprawozdanie z laboratoriów HTK

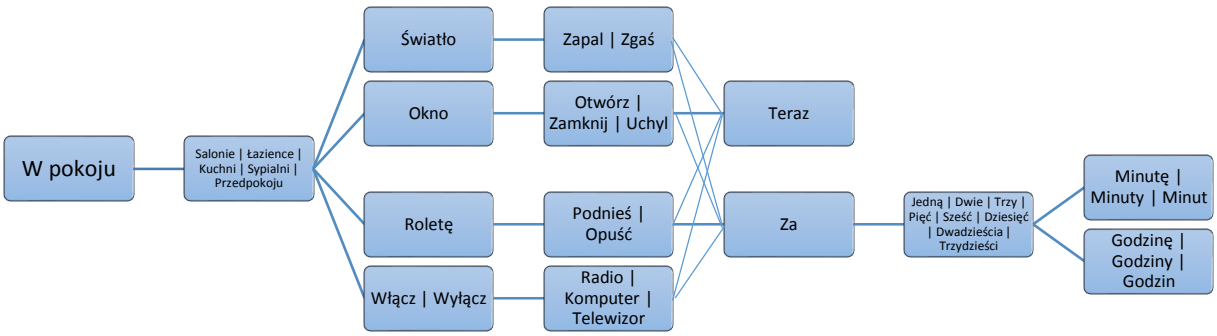
1.Opis gramatyki

System był projektowany w celu obsługi inteligentnych instalacji w domach. Istnieją systemy pozwalające kontrolować w prosty sposób światło, rolety, zestawy kina domowego itp. za pomocą jednego pilota. Gramatyka została zaprojektowana tak, aby użytkownik najpierw mówił, w którym pomieszczeniu chce wykonać daną akcję, a następnie, z czym związana jest dana akcja i samą akcje. Na końcu dodano możliwość odłożenia akcji w czasie, czyli można wykona akcję teraz lub za dany odstęp czasu.

Plik opisujący gramatykę:

$pokOj = salonie | Lazience | kuchni | sypialni | przedpokoju;
$okno = otwOrz | zamknij | uchyl;
$roleta = podnieS | opuSC;
$SwiatLo = zapal | zgaS;
$media = radio | telewizor | komputer;
$ile = dziesiEC | dwadzieScia | trzydzieSci | jednA | dwie | trzy | piEC | szeSC | dwanaScie;
$czego = minut | godzin;

(SENT-START (w_pokoju <$pokOj> (okno $okno | roletE $roleta | SwiatLo $SwiatLo | wLAcz $media | wyLAcz $media)) (teraz | za $ile $czego) SENT-END)



Dla ułatwienia odmiany wartości czasu (tzn. godzin, godzinę, godziny oraz minut, minutę, minuty) zostały oznaczone, jako jedno słowo (minut lub godzin) z trzema możliwymi transkrypcjami fonetycznymi, ponieważ dla systemu ważna jest ilość, czyli poprzedzające słowo. Ta operacja ma eliminować błędy w rozpoznaniu (np. program rozpoznał godziny, podczas gdy powiedziano godzinę), które nie wpłyną negatywnie na ogólny wynik treści zdania.

2.Opis nagrań

Zarówno nagrania treningowe i testowe wykonano w zamkniętym, umeblowanym pomieszczeniu mieszkalnym przy użyciu następujących urządzeń:

Mikrofon: Behringer XM8500
Rejestrator: Line6 POD X3 Live

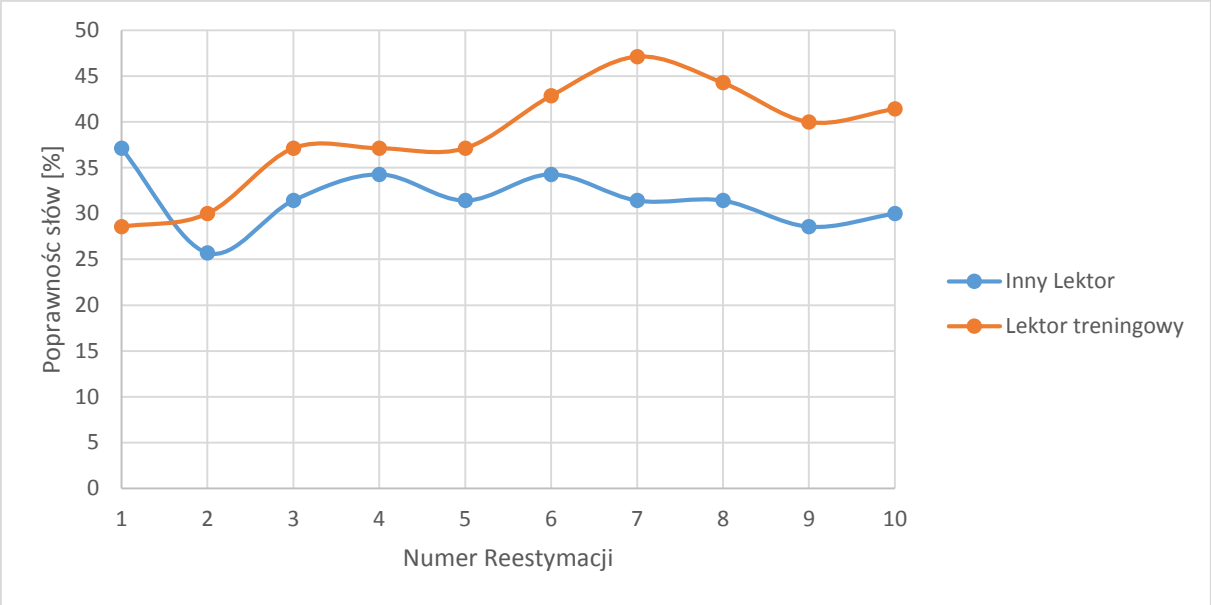
Nagranie treningowe zawierało ponad 3 minuty całych zdań w ten sposób, aby każde słowo było wypowiedziane co najmniej trzykrotnie.
Zbiór testowy składał się z 10 nagrań pojedynczych zdań wypowiedzianych przez tego samego lektora, w tych samych warunkach, co nagrania treningowe.

3.Wyniki

Reestymacja	Lektor treningowy		Inny Lektor	
	Corr	Acc	Corr	Acc
1	28,57%	25,71%	37,14%	32,86%
2	30,00%	30,00%	25,71%	25,71%
3	37,14%	34,29%	31,43%	31,43%
4	37,14%	34,29%	34,29%	34,29%
5	37,14%	34,29%	31,43%	31,43%
6	42,86%	40,00%	34,29%	34,29%
7	47,14%	47,14%	31,43%	31,43%
8	44,29%	44,29%	31,43%	31,43%
9	40,00%	40,00%	28,57%	28,57%
10	41,43%	41,43%	30,00%	30,00%

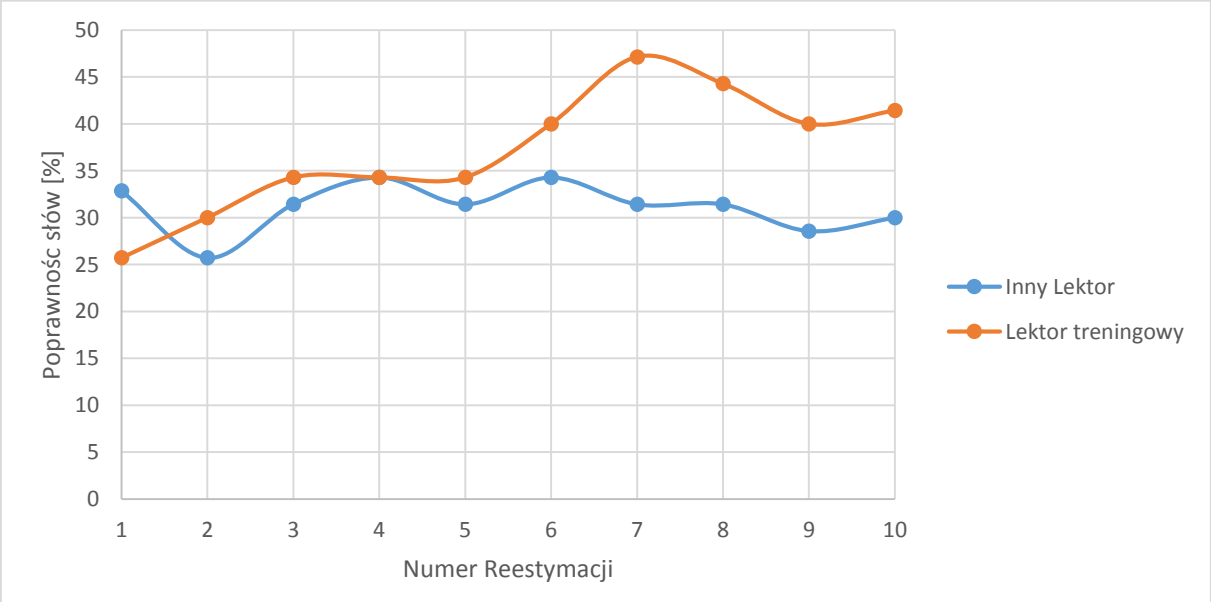
Tab. 1 Wyniki rozpoznania nagrań testowych.

W tabeli zamieszczono wyniki rozpoznania nagrań testowych lektora, który wykonywał nagranie treningowe oraz innego lektora, również płci męskiej. W każdym przypadku rozpoznawalność zdań wyniosła 0%. Najwyższa rozpoznawalność wyniosła 47.14% i wystąpiła u lektora treningowego przy siódmej reestymacji. Najwyższa rozpoznawalność drugiego lektora wyniosła 37.14% i wystąpiła przy pierwszej reestymacji.



Wykres 1. Poprawność rozpoznania (Corr) w zależności od krotności reestymacji.

Na wykresie 1. można zaobserwować, że poprawność słów w nagraniu lektora treningowego, poza pierwszą reestymacją jest, o co najmniej 5% wyższa niż u drugiego lektora.



Wykres 2. Poprawność rozpoznania (Acc) w zależności od krotności reestymacji.

4.Analiza wyników i wnioski

Zerowa rozpoznawalność zdań może wynikać z dość skomplikowanej gramatyki. W zaproponowanym systemie występuje 35 różnych słów, a poszczególne zdania w nagraniu testowym są stosunkowo długie (5 do 9 słów).

Analizując rozpoznane zdania można zauważyć, że program bardzo często błędnie rozpoznaje żądany czas wykonania akcji. Powodem tego zapewne jest niefortunnie zbudowana gramatyka, w której ważne jest rozpoznanie słowa „za”, ponieważ po nim wypowiadany jest żądany odstęp czasu. Zamiast tego program w większości przypadków rozpoznaje słowo „teraz”, które jest zakończeniem zdania.

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.).