Technologia Mowy	**Rozpoznawanie mowy przy wykorzystaniu HTK**	*Dominika Behounek*
05.01.2014r		WIMiR Inżynieria Akustyczna III rok

I. Gramatyka

Celem projektu było opracowanie systemu rozpoznawania mowy korzystając z HTK. Stworzona gramatyka symulowała wizytę w kinie, tym samym więc zawierała informacje na temat rodzaju biletu wraz z ich ilością, tytuł filmu, dzień tygodnia, w którym odbywa się seans oraz przybliżoną porę dnia.

System tego typu mógłby okazać się przydatny we wszelkich obiektach, gdzie występuje konieczność zakupu biletu wstępu: teatry, sale koncertowe, kino czy nawet muzea.

Poniżej prezentowana jest realizowana gramatyka:

```
$ilosc = dwa | trzy | cztery;
$rodzajjeden = ulgowy | normalny;
$rodzajwiele = ulgowe | normalne;
$film = miS | rejs | pianista | kiler | amator | rewers | pasja | komornik | rOZa |
przesLuchanie | przypadek | katedra | dzieN_Swira | sami_swoi;
$dzien = poniedziaLek | wtorek | SrodE | czwartek | piAtek | sobotE | niedzielE;
$pora = rano | wieczOr;

( SENT-START ( poproszE ((jeden bilet $rodzajjeden)  | ($ilosc bilety $rodzajwiele))
na film $film na $dzien $pora ) SEND-END )
```

Zasadniczą wadą systemu jest fakt, że wymusza on na użytkowniku konkretną składnię wypowiadanego zdania. Każde ze zdań składa się z dziesięciu słów, z czego słów „dowolnych" jest pięć, co stanowi dokładnie 50% całości zdania.

Aby rozpocząć pracę z systemem należało przygotować plik zawierający transkrypcje fonetyczne wszystkich wyrazów, które wykorzystuje podana gramatyka. Zawartość tego pliku prezentowana jest poniżej.

poproszE	p o p r o sz e sil
jeden	j e d e n sil
dwa	d w a sil
trzy	t sz y sil
cztery	cz t e r y sil
bilet	b i l e t sil
bilety	b i l e t y sil
ulgowy	u l g o w y sil
ulgowe	u l g o w e sil
normalny	n o r m a l n y sil
normalne	n o r m a l n e sil
na	n a sil
film	f i l m sil
miS	m i si sil
rejs	r e j s sil
pianista	p j a ni i s t a sil
kiler	k i l e r sil
amator	a m a t o r sil
rewers	r e w e r s sil
pasja	p a s j a sil
komornik	k o m o r ni i k sil
rOZa	r u rz a sil
przesLuchanie	p sz e s l_ u h a ni e sil
przypadek	p sz y p a d e k sil
katedra	k a t e d r a sil
dzieN_Swira	dzi e ni si f i r a sil
sami_swoi	s a m i s f o j i sil
poniedziaLek	p o ni e dzi a l_ e k sil
wtorek	f t o r e k sil
SrodE	si r o d e sil
czwartek	cz f a r t e k sil
piAtek	p j o n t e k sil
sobotE	s o b o t e sil
niedzielE	n i e dzi e l e sil
rano	r a n o sil
wieczOr	w j e cz u r sil
SENT-START	[] sil
SEND-END	[] sil

II. Nagrania: sprzęt oraz warunki

Podczas nagrywania plików treningowych użyto następujących urządzeń i oprogramowania:

- mikrofon: Shure Beta 58a
- procesor wokalny: TC-Helicon VoiceLive 2
- oprogramowanie: Audacity

Nagranie wykonane zostało dla częstotliwości próbkowania wynoszącej 16kHz przy rozdzielczości równej 16 bitów.

Plik dźwiękowy zawierał nagrania 42 zdań ułożonych w taki sposób, aby każde ze słów gramatyki wypowiedziane zostało podczas treningu przynajmniej trzy razy. Tym sposobem uzyskano plik trwający 03:07 minuty, z czego wynika, że na każde zdanie przypadało około 4,5 sekundy nagrania. Widać więc, że tempo wypowiadanych zdań było dosyć szybkie, bardzo zbliżone do prędkości wypowiedzi w warunkach „naturalnych".

Kolejnym krokiem było anotowanie poszczególnych wypowiedzianych słów w nagraniu treningowym. Do tego posłużono się programem AudioDescription. Następnie każde ze słów podzielone zostało na poszczególne fonemy.

Nagrania dziesięciu zdań mających testować skuteczność stworzonego systemu rozpoznawania mowy nagrano dokładnie tym samym sprzętem, w tych samych warunkach zachowując bardzo podobne tempo wypowiedzi. Jedyną drobną różnicą między tymi nagraniami jest mniejsza głośność nagrań zdań testowych.

Pomieszczenie, w którym nagrywano pliki dźwiękowe było bardzo ciche, tak więc nagranie pozbawione jest większych zakłóceń pochodzących z zewnątrz. Szumy nie zostały całkowicie wyeliminowane, jednak w przypadku trzyminutowego nagrania treningowego odległość rzeczywistego sygnału od szumu jest bardzo duża. Trochę gorsza „czystość" nagrania została uzyskana dla nagrań testowych, gdyż mowa jest znacznie cichsza, jednak zrozumiałość wypowiedzi i tak jest zadowalająca.

III. Otrzymane wyniki

Poniżej prezentowane są wyniki zestawione w tabeli zawierającej procentowe rozpoznanie dla wszystkich słów, zmiennych słów oraz ilość zdań rozpoznanych w całości.

Numer reestymacji	Ilość rozpoznanych w całości zdań (na 10)	Rozpoznanie wszystkich słów [%]	Rozpoznanie zmiennych słów [%]
1	1	67.33	50
2	5	91.09	84
3	8	97.03	96
4	8	97.03	96
5	9	98.02	98
6	9	98.02	98
7	8	97.03	96
8	8	97.03	96
9	8	97.03	96
10	8	97.03	96

Na podstawie prezentowanych powyżej wyników można stwierdzić, że skuteczność zaprojektowanego systemu, choć nie jest idealna, jest bardzo zadowalająca. Już w przypadku piątej i szóstej reestymacji uzyskujemy rozpoznanie na poziomie 98%, co oznacza, że na 10 wypowiedzianych zdań (a więc 100 wypowiedzianych wyrazów) zaledwie jedno słowo zostało błędnie zidentyfikowane przez system.

Wśród używanych wyrazów można zauważyć takie, które najrzadziej (lub w ogóle) są rozpoznawane przez program. Wśród nich znajdują się:

- „Przesłuchanie" – 0% rozpoznania
  Jedyne słowo, które nigdy nie zostało rozpoznane. Nagranie zawierające ten tytuł filmu zdaje się nie odbiegać jakością ani sposobem wymowy słowa od innych. Co dziwniejsze system identyfikował to słowo zawsze jako „Komornik". Zastanawiający jest fakt, iż słowa nie zdają się być do siebie zbliżone. Błędem mogącym wpływać na taki stan rzeczy może być błędna transkrypcja fonetyczna, którą prawdopodobnie (po znastanowieniu, znając skutki obecnego stanu rzeczy) powinno się zmienić z:
  p sz e s l_ u h a ni e na *p sz e s u h a ni i e.*

- „Dzień Świra" - 60% rozpoznania
  Wyraz (właściwie dwa wyrazy, jednak w systemie zaimplementowane były one jako zawsze występujące obok siebie) mylony był najczęściej ze słowem „Kiler". Powód takiego błędu jest nieznany. Transkrypcja fonetyczna zdaje się być poprawna, a oba wyrazy nie mają ze sobą wiele wspólnego (przede wszystkim znacznie różnią się długością), ciężko więc stwierdzić, w jaki sposób można by próbować eliminować tę pomyłkę.

- Dni tygodnia – rozpoznawalność przedstawiona w tabeli poniżej

Dzień tygodnia	Rozpoznawalność [%]
poniedziałek	100
wtorek	80
środa	80
czwartek	100
piątek	95
sobota	80
niedziela	85

IV. Wnioski

Skuteczność stworzonego systemu rozpoznawania mowy można określić jako bardzo dobrą. Oprogramowanie tego typu byłoby znakomitym dodatkiem do samoobsługowych automatów do kupowania biletów, które już teraz znajdują się w większości nowych kin. Tym sposobem można by ułatwić komunikację maszyn np. z osobami niewidomymi lub niedowidzącymi, mającymi problemy z korzystaniem z graficznych interfejsów, w które zaopatrzone są maszyny tego typu.

Oczywiście takie rozwiązanie wymagałoby rozbudowania bazy plików treningowych, gdyż w tym przypadku całość testu wykonywana była dla jednej osoby, natomiast w warunkach naturalnych system musiałby wykazywać się bardzo wysoką skutecznością niezależnie od mówcy.

Aby zwiększyć skuteczność tego konkretnego systemu, można by rozważyć zrealizowanie nagrań w lepszych warunkach, jednak biorąc pod uwagę względy praktyczne i faktyczne zastosowanie – nie miało by to większego sensu, gdyż w warunkach rzeczywistych nie jest możliwe uzyskanie „sterylnego" nagrania audio, ze względu na poziom szumów tła.

Największą napotkaną trudnością był najprawdopodobniej fakt, że pliki treningowe zawierały nagrania mowy ciągłej, w związku z czym już na etapie anotacji ciężko było momentami wyizolować koniec jednego słowa i początek następnego. To, wraz z możliwym niedokładnym oznaczeniem czasów trwania konkretnych fonemów (w pliku *transkrypcja_fonetyczna.mlf*) mogło wpłynąć na spadek skuteczności przy identyfikacji słów.

Jednym z rozwiązań wprowadzonych aby zwiększyć skuteczność rozpoznań była próba zwiększenia liczby reestymacji, jednak rozwiązanie to nie okazało się skuteczne, gdyż dla większej ich ilości procentowe rozpoznanie nadal utrzymywało się na poziomie około 97-98%.