Inżynieria Akustyczna – Technologia mowy , rok 2013

Krzysztof Drewniany

Sprawozdanie z laboratoriów HTK

1. W ramach laboratoriów wykonano system rozpoznawania mowy realizujący automatyczne zamawianie dań w budce z kebabem. Gramatyka systemu została dobrana aby zamówienie było dokonywane w grzecznościowej formie i nie pozostawiało żadnych wątpliwości co do oczekiwanego produktu. Użytkownik musi znać odpowiedni szyk zdania aby zamówienie zostało poprawnie rozpoznane, losowe wypowiedzi będą rozpoznawane błędnie co zostało wykazane w testach.

2. Do nagrań korpusu użyto dynamicznego mikrofonu Sony F-V120 o paśmie przenoszenia 60 Hz – 12 kHz. Nagrań dokonano w sześciennym pokoju mieszkania w bloku przy braku obecności innych osób. Wypowiedź wykorzystana w korpusie była mową ciągłą. Format pliku to mono wave PCM, 16 bit o częstotliwości próbkowania 16 kHz.

3. Nagrania testowe wykonano w tych samych warunkach co nagrania korpusu. Zarejestrowane zostało 8 zdań wykorzystujących wszystkie słowa ze słownika:

a) Poproszę mega kebab drobiowy w tortilli z sosem czosnkowym oraz dużą fantę.
b) Poproszę duży kebab drobiowy w bułce z sosem łagodnym oraz małą kolę.
c) Poproszę dużego kebaba drobiowego w tortilli z sosem średnim oraz dużą kolę.
d) Poproszę średni kebab wołowy w bułce z sosem ostrym oraz małą fantę.
e) Poproszę średniego kebaba wołowego w tortilli z sosem czosnkowym oraz dużą fantę.
f) Poproszę mały kebab wegetariański w bułce z sosem łagodnym oraz małą kolę.
g) Poproszę małego kebaba wegetariańskiego w tortilli z sosem średnim oraz dużą kolę.
h) Poproszę mega kebab wegetariański w bułce z sosem ostrym oraz małą fantę.

Zrezygnowano z użycia programu HResults na rzecz dokładniejszej analizy ręcznej.

4. Wyniki testów zestawiono w poniższej tabeli gdzie:

- **lw** to liczba wystąpień danego słowa,
- **lprddie** to liczba poprawnych rozpoznań dla danej ilości reestymacji,
- **pprds** to procentowa poprawność rozpoznania danego słowa,
- **SUMA** to suma poprawnie rozpoznanych słów dla danej reestymacji,
- **ppr** to procentowa poprawność rozpoznania dla danej reestymacji.

	lw	lprddie: 4	lprddie: 5	lprddie: 6	lprddie: 7	lprddie: 8	lprddie: 9	lprddie: 10	lprddie: 11	lprddie: 12	lprddie: 13	lprddie: 14	lprddie: 15	pprds [%]
bułce	4	4	4	4	4	4	4	4	4	4	4	4	4	100
czosnkowym	2	2	0	1	1	2	2	2	2	2	2	2	2	83,3
drobiowego	1	1	1	1	1	1	1	1	1	1	1	1	1	100
drobiowy	2	2	2	2	2	2	2	2	2	2	2	2	2	100
dużą	4	4	4	4	4	4	4	4	4	4	4	4	4	100
dużego	1	1	1	1	1	1	1	1	1	1	1	1	1	100
duży	1	1	1	1	1	1	1	1	1	1	1	1	1	100
fantę	4	3	3	2	2	2	2	2	2	2	2	2	2	54,2
kebab	5	5	5	5	5	5	5	5	5	5	5	5	5	100
kebaba	3	1	1	1	1	1	1	1	1	1	1	1	1	33,3
kolę	4	4	4	4	4	4	4	4	4	4	4	4	4	100
łagodnym	2	2	2	2	2	2	2	2	2	2	2	2	2	100
małą	4	4	4	4	4	4	4	4	4	4	4	4	4	100
małego	1	0	0	0	0	0	0	0	0	0	0	0	0	0
mały	1	1	1	1	1	1	1	1	1	1	1	1	1	100
mega	2	1	0	0	0	0	2	2	2	2	2	2	2	62,5
oraz	8	8	8	8	8	8	8	8	8	8	8	8	8	100
ostrym	2	2	0	1	1	1	2	2	2	2	2	2	2	79,2
poproszę	8	8	8	8	8	8	8	8	8	8	8	8	8	100
sosem	8	8	8	8	8	8	8	8	8	8	8	8	8	100
średni	1	1	1	1	1	1	1	1	1	1	1	1	1	100
średniego	1	0	0	0	0	0	0	0	0	0	0	0	0	0
średnim	2	2	2	2	2	2	2	2	2	2	2	2	2	100
tortilli	4	4	4	4	4	4	4	4	4	4	4	4	4	100
w	8	8	8	8	8	8	8	8	8	8	8	8	8	100
wegetariańskiego	1	0	0	0	0	0	0	0	0	0	0	0	0	0
wegetariański	2	2	1	0	0	0	0	0	0	0	1	2	1	29,2
wołowy	1	1	1	1	1	1	1	1	1	1	1	1	1	100
wołowego	1	0	0	0	0	0	0	0	0	0	0	0	0	0
z	8	8	8	8	8	8	8	8	8	8	8	8	8	100
SUMA	96	88	82	82	82	83	86	86	86	86	87	88	87	
ppr [%]		91,7	85,4	85,4	85,4	86,5	89,6	89,6	89,6	89,6	90,6	91,7	90,6	

Wykonano 15 reestymacji. Wbrew oczekiwaniom najwyższa poprawność została osiągnięta dla 14 a nie dla około 9 reestymacji. Taką samą poprawność osiągnięto dla 4 lecz wydaję się to być przypadkiem. Dla 15 poprawność rozpoznania zaczyna spadać co można tłumaczyć przetrenowaniem modelu.

Słowa, które nie zostały ani razu poprawnie rozpoznane to: „małego”, „średniego”, „wegetariańskiego” i „wołowego”. Były mylone odpowiednio z „mały”, „średni”, „wegetariański” i „wołowy”. Przyczyną takiego stanu jest fonetyczne podobieństwo, słowa różnią się tylko końcówką. Ponadto liczba występowań tych słów w korpusie jest stosunkowo niewielka, „wołowy” występuję 10 razy, bardzo podobny „wołowego” już tylko 8 co wydaję się i tak niewielką liczbą przy 50 występowaniach słowa „poproszę”. Pozostałe wyrazy, które nie zostały rozpoznane w 100% poprawnie również wykazywały podobieństwo do innych lub posiadały niedostateczną ilość występowań w korpusie.

5. W ramach dalszych testów sprawdzono odporność systemu na:

a) jakość nagrań.

Do 3 zawsze poprawnie rozpoznawanych nagrań testowych dodano szum biały, testu dokonano dla optymalnych 14 reestymacji. Słowa koloru czerwonego zostały rozpoznane niepoprawnie:

Poproszę **duży** kebab **drobiowy** w **bułce** z sosem **łagodnym** oraz małą kolę.
Poproszę **dużego kebaba drobiowego** w tortilli z sosem **średnim** oraz **dużą** kolę.
Poproszę **średni** kebab wołowy w bułce z sosem ostrym oraz małą **fantę.**

Jak widać obniżenie jakości nagrania przez jego zaszumienie istotnie pogarsza poprawność rozpoznania.

b) losowość wypowiedzi.

Z najlepiej rozpoznawanych słów zbudowano 3 zdania, których szyk nie było zgodny z gramatyką systemu. Nagrań dokonano w takich samych warunkach jak korpusu. Zdania koloru czerwonego zostały rozpoznane przez system przy 14 reestymacjach natomiast zielone słowa nie pokrywały się ze zdaniem testowym:

Duży kebab drobiowy z sosem łagodnym w tortilli poproszę oraz małą kolę.
Poproszę duży kebab drobiowy w tortilli z sosem **ostrym** oraz małą kolę.
Mały wołowy kebab poproszę w bułce z sosem ostrym oraz dużą fantę.
Poproszę mały kebab wołowy w bułce z sosem ostrym oraz dużą fantę.
Małą kolę oraz kebab wegetariański w tortilli z sosem czosnkowym poproszę.
Poproszę **mega kebaba drobiowego** w tortilli z sosem **ostrym** oraz małą kolę.

Wykazano, że błędna gramatyka obniża poprawność rozpoznania lecz nie w stopniu tak znaczącym jak niewystarczająca ilość reestymacji czy jakość nagrania.

6. Wnioski.

Praktycznie każdy element automatycznego systemu rozpoznawania mowy ma wpływ na poprawność jego rozpoznania a tym samym jakość. Dla określonego celu należy odpowiednio dobrać obszerność korpusu, jakość jego nagrania oraz liczbę reestymacji. Gramatyka systemu powinna być jak najbardziej elastyczna, umożliwiając poprawne rozpoznanie jak największej ilości wypowiedzi. Należy również zadbać aby rozpoznawana mowa była rejestrowana w możliwie zbliżonych warunkach do tych, w jakich nagrywano korpus. Dla niskiej jakości sprzętu użytego przy realizacji zadania oraz wybranych narzędziach (HTK) 91,7% poprawności rozpoznania wydaję się być zadowalającym wynikiem.

Wyrażam zgodę na dołączenie moich nagrań do korpusu mowy AGH.