Kinga Frydrych

Wydział Inżynierii Mechanicznej i Robotyki

Inżynieria Akustyczna, rok III, 2013/2014

Sprawozdanie z ćwiczeń laboratoryjnych z Technologii mowy

Rozpoznawanie mowy za pomocą HTK

1. Opis gramatyki

Prezentowany system rozpoznawania mowy powstał przy użyciu HTK Speech Recognition Toolkit. Jego zadanie polega na rozpoznawaniu treści zamawiania odzieży. Użytkownik ma możliwość wyboru rodzaju ubrania, jego koloru oraz rozmiaru.

Gramatyka wygląda następująco:

```
$prosba = poproszE | chciaLbym_kupiC | chciaLabym_kupiC ;

$ubranie = spodnie | sukienkE | spOdnicE | bluzkE | koszulE | krawat | garnitur ;

$kolor = biaLym | ZOLtym | czerwonym | niebieskim | zielonym | czarnym ;

$rozmiar = iks_es | es | em | el | iks_el ;

(SENT-START ( $prosba $ubranie w_kolorze $kolor rozmiar $rozmiar ) SENT-END)
```

Na sześć słów, które system ma rozpoznać, dwa z nich („w_kolorze”, „rozmiar”) występują w każdej wypowiedzi (oprócz wypowiedzi z krawatem, gdzie nie podaje się rozmiaru), zatem już na wstępie rozpoznanie wynosi ok. 30%.

Gramatykę skompilowano oraz utworzono słownik z zapisem fonetycznym użytych słów (w załączniku).

2. Nagrania treningowe

Kolejnym krokiem było nagranie wypowiedzi treningowych w oparciu o plik *gram*. Rejestracji dokonano przy użyciu mikrofonu wbudowanego w laptop Packard Bell Easy Note TK PEW92 (Realtek High Definiton Audio), w programie Audacity. Zdania znalazły się w monofonicznym pliku *TMprobka.wav* (w załączniku), trwającym około trzech minut, częstotliwość próbkowania: 16 kHz, 16 bit. Nagranie powstało w warunkach domowych, w umeblowanym pokoju. Rejestrowany był głos kobiecy.

3. Anotacja

Korzystając z powstałego na AGH programu Anotator (Audio Descriptor) dokonano ręcznie anotacji na poziomie słów. Następnie automatycznie podzielono nagranie na fonemy.

4. Modele HMM

Utworzono prototypowy model HMM, po czym dokonano dziesięciu jego reestymacji.

5. Testowanie systemu

Nagrano dziesięć wypowiedzi testowych, przy użyciu tego samego sprzętu, co nagrania treningowe. Przepuszczono je przez utworzony system rozpoznawania mowy. Uzyskano pliki *recoutn.mlf* (gdzie „n" oznacza numer reestymacji), które zawierają słowa, jakie HTK udało się rozpoznać dla każdego pliku testowego.

Następnie sprawdzono (dla wszystkich reestymacji), jakie jest procentowe rozpoznanie zdań oraz słów w nagraniach:

```
C:\Users\Domowy\Desktop\rozpmow>HResults -I testref.mlf monophones0 recout1.mlf
====================== HTK Results Analysis =======================
  Date: Wed Dec 25 21:17:17 2013
  Ref : testref.mlf
  Rec : recout1.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=48.28, Acc=44.83 [H=28, D=0, S=30, I=2, N=58]
===================================================================

C:\Users\Domowy\Desktop\rozpmow>HResults -I testref.mlf monophones0 recout2.mlf
====================== HTK Results Analysis =======================
  Date: Wed Dec 25 21:17:23 2013
  Ref : testref.mlf
  Rec : recout2.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=44.83, Acc=41.38 [H=26, D=0, S=32, I=2, N=58]
===================================================================

C:\Users\Domowy\Desktop\rozpmow>HResults -I testref.mlf monophones0 recout3.mlf
====================== HTK Results Analysis =======================
  Date: Wed Dec 25 21:17:35 2013
  Ref : testref.mlf
  Rec : recout3.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=48.28, Acc=44.83 [H=28, D=0, S=30, I=2, N=58]
===================================================================

C:\Users\Domowy\Desktop\rozpmow>HResults -I testref.mlf monophones0 recout4.mlf
====================== HTK Results Analysis =======================
  Date: Wed Dec 25 21:17:41 2013
  Ref : testref.mlf
  Rec : recout4.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=48.28, Acc=44.83 [H=28, D=0, S=30, I=2, N=58]
===================================================================

C:\Users\Domowy\Desktop\rozpmow>HResults -I testref.mlf monophones0 recout5.mlf
====================== HTK Results Analysis =======================
  Date: Wed Dec 25 21:17:45 2013
  Ref : testref.mlf
  Rec : recout5.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=48.28, Acc=44.83 [H=28, D=0, S=30, I=2, N=58]
```

```
C:\Users\Domowy\Desktop\rozpmow>HResults -I testref.mlf monophones0 recout6.mlf
====================== HTK Results Analysis =======================
  Date: Wed Dec 25 21:17:52 2013
  Ref : testref.mlf
  Rec : recout6.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=50.00, Acc=46.55 [H=29, D=0, S=29, I=2, N=58]
===================================================================

C:\Users\Domowy\Desktop\rozpmow>HResults -I testref.mlf monophones0 recout7.mlf
====================== HTK Results Analysis =======================
  Date: Wed Dec 25 21:17:55 2013
  Ref : testref.mlf
  Rec : recout7.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=50.00, Acc=46.55 [H=29, D=0, S=29, I=2, N=58]
===================================================================

C:\Users\Domowy\Desktop\rozpmow>HResults -I testref.mlf monophones0 recout8.mlf
====================== HTK Results Analysis =======================
  Date: Wed Dec 25 21:17:59 2013
  Ref : testref.mlf
  Rec : recout8.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=50.00, Acc=46.55 [H=29, D=0, S=29, I=2, N=58]
===================================================================

C:\Users\Domowy\Desktop\rozpmow>HResults -I testref.mlf monophones0 recout9.mlf
====================== HTK Results Analysis =======================
  Date: Wed Dec 25 21:18:02 2013
  Ref : testref.mlf
  Rec : recout9.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=50.00, Acc=46.55 [H=29, D=0, S=29, I=2, N=58]
===================================================================

C:\Users\Domowy\Desktop\rozpmow>HResults -I testref.mlf monophones0 recout10.mlf

====================== HTK Results Analysis =======================
  Date: Wed Dec 25 21:18:05 2013
  Ref : testref.mlf
  Rec : recout10.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=48.28, Acc=44.83 [H=28, D=0, S=30, I=2, N=58]
```

Rys. 1. Wyniki rozpoznania zdań i słów

Najlepsze rozpoznanie to 50% (zawiera wcześniej wymienione 30%).

6. Analiza błędów rozpoznania i wnioski

W każdym nagraniu, dla każdej reestymacji prawidłowo rozpoznane są stałe elementy („w_kolorze”, „rozmiar”). Na podstawie uzyskanych wcześniej plików *recout.mlf* (w załączniku) zaobserwowano następujące rozpoznania dla poszczególnych słów:

Tabela 1. Słowa oraz ich rozpoznanie [%]

Słowo	spodnie	sukienkE	spodnicE	bluzkE	garnitur	krawat	koszulE	poproszE	chciaLbym_kupiC	chciałabym_kupiC
Rozp.	0	0	0	0	10	60	40	97,5	0	10
iks_es	es	em	el	iks_el	ZOLtym	biaLym	czarnym	zielonym	czerwonym	niebieskim
10	45	0	50	0	50	0	0	10	70	0

Uwagi:

- System dla każdej reestymacji zamiast słowa „spodnie” wypisuje w rozpoznaniu słowo „koszulE”
- Zamiast słowa „koszule” niekiedy pojawia się słowo „spodnie”

Zdaniem autorki sprawozdania, duża ilość błędów wynika z niepoprawności automatycznej anotacji na poziomie fonemów. W celu sprawdzenia, czy faktycznie tak jest, zrealizowano w MATLAB-ie program, który na podstawie pliku z anotacją na fonemy (*fonet.mlf* – w załączniku) oraz plików audio odtwarza dźwięki, które są zawarte w danych przedziałach czasowych. Gdyby podział na fonemy był zrobiony poprawnie, rezultatem działania programu powinno być odtworzenie ciągu liter (np. „aaaaaaaa”), natomiast w przypadku prezentowanego systemu rozpoznawania mowy nie zawsze jest to czysta sekwencja danej litery. Szczególnie dobrze widać to w przypadku fonemu „e”, gdzie pojawia się nie tylko ta głoska, ale też fragmenty słów, a nawet całe wyrazy – próbka ta została dodana do listy załączników pod nazwą *blad.wav*.

7. Sprawdzenie działania systemu rozpoznawania mowy dla innego głosu

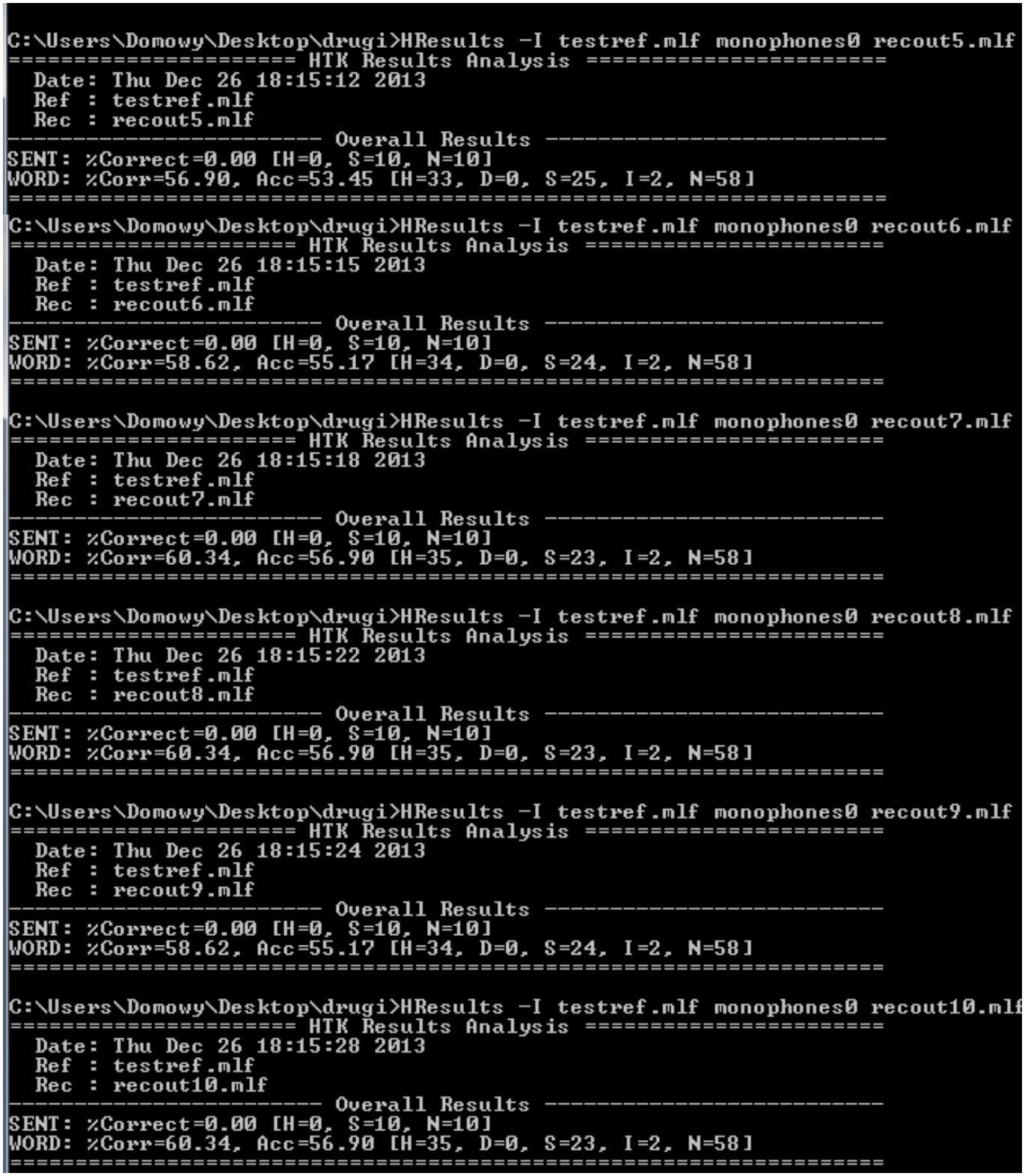
W tym celu nagrano kolejne dziesięć wypowiedzi (przy pomocy takiego samego sprzętu, jak wcześniej). Tym razem rejestrowany był głos dziecięcy.

```
C:\Users\Domowy\Desktop\drugi>HResults -I testref.mlf monophones0 recout1.mlf
====================== HTK Results Analysis =======================
  Date: Thu Dec 26 18:14:44 2013
  Ref : testref.mlf
  Rec : recout1.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=55.17, Acc=51.72 [H=32, D=0, S=26, I=2, N=58]
===================================================================

C:\Users\Domowy\Desktop\drugi>HResults -I testref.mlf monophones0 recout2.mlf
====================== HTK Results Analysis =======================
  Date: Thu Dec 26 18:14:58 2013
  Ref : testref.mlf
  Rec : recout2.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=63.79, Acc=60.34 [H=37, D=0, S=21, I=2, N=58]
===================================================================

C:\Users\Domowy\Desktop\drugi>HResults -I testref.mlf monophones0 recout3.mlf
====================== HTK Results Analysis =======================
  Date: Thu Dec 26 18:15:03 2013
  Ref : testref.mlf
  Rec : recout3.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=60.34, Acc=56.90 [H=35, D=0, S=23, I=2, N=58]
===================================================================

C:\Users\Domowy\Desktop\drugi>HResults -I testref.mlf monophones0 recout4.mlf
====================== HTK Results Analysis =======================
  Date: Thu Dec 26 18:15:06 2013
  Ref : testref.mlf
  Rec : recout4.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=48.28, Acc=44.83 [H=28, D=0, S=30, I=2, N=58]
===================================================================
```



Rys. 2. Wyniki rozpoznania zdań i słów dla głosu innej osoby

Rozpoznanie zdań wypowiadanych głosem innym niż nagrania treningowe, jest lepsze niż w przypadku tego samego głosu – dochodzi nawet do 63,8% dla drugiej reestymacji.

Do powyższego raportu załączono: nagranie treningowe *TMprobka.wav*, nagrania testowe *1-10.wav*, pliki tekstowe *gram.txt* oraz *dict.txt*, pliki z rozpoznaniami *recout.mlf*, plik dźwiękowy *blad.wav* oraz pozostałe pliki powstałe podczas realizowania automatycznego systemu rozpoznawania mowy.