ISOLATED WORD RECOGNITION

Pavel Hrdlička
Bachelor Degree Programme (3), FIT BUT
E-mail: xhrdli09@stud.fit.vutbr.cz

Supervised by: František Grézl
E-mail: grezl@fit.vutbr.cz

ABSTRACT

This work is concerned with creation of isolated word recognizer, testing its functionality on data sample and improvement by normalisation and speaker adaptation techniques. Phoneme recognizer is bulit on HTK (Hidden Markov Model Toolkit).

1. ÚVOD

Studium cizích jazyků v současné době nabývá na důležitosti. Tento trend se nevyhnutelně musí projevit i v moderních informačních technologiích. Elektronické slovníky umožňují vyhledávání velkého množství hesel za krátkou dobu.

Cílem této práce je doplnit klasické ovládání slovníku pomocí klávesnice a myši o ovládání pomocí lidské řeči. Tento přístup přináší větší komfort při vyhledávání jednotlivých slov. Uživatel pomocí mikrofonu zadá vyhledávané slovo a program mu během okamžiku nabídne několik možností, ze kterých si může vybrat. Toho se dá využít i v případě, kdy uživatel slovo někde zaslechl a neví přesně jak se píše. Uživatel si pak vybere nejvhodnější variantu a pracuje dále se slovníkem.

Hlavním úkolem je přizpůsobit uživatelská data, která byla dodána firmou Lingea, již existujícím modelům, natrénovaným na velkém množství řečových dat. Protože nemáme dostatek dat na natrénování nových modelů, přizpůsobujeme dodaná data tak, aby bylo možno použít existující modely natrénované v rozdílném prostředí.

2. TEORETICKÝ ROZBOR

Od firmy Lingea jsme obdrželi sady nahrávek 994 různých slov, každá sada byla nahrána jedním z 12 mluvčích v kancelářském prostředí s různými typy mikrofonů. Slova byla uložena ve formátu wav 44,1 kHz a pojmenována podle vysloveného slova. Z důvodu vysoké úrovně rušení byly tři sady nahrávek z testování vyřazeny. Kvůli přizpůsobení uživatelských dat natrénovaným modelům, byly nahrávky podvzorkovány programem sox metodou polyphase na 8 kHz.

Z řeči jsou vypočítány PLP (Perceptual linear prediction) parametry, které popisují řečový signál a jsou vhodné pro rozpoznávání řeči. Tyto parametry byly použity i při trénování modelů a je proto nezbytné použít stejný typ parametrů. Při vytváření PLP parametrů jsem použil delta a double delta.

Pro rozpoznávač izolovaných slov jsou kromě parametrů ještě potřeba modely, gramatika, výslovnostní slovník a dekodér.

Gramatika určuje posloupnost slov, které může rozpoznávač rozpoznat. Pro rozpoznávání izolovaných slov je ve tvaru: `ticho SLOVO ticho`, kde `SLOVO` je paralelní spojení všech slov ve slovníku.

Výslovnostní slovník s fonetickým přepisem rozpoznávaných slov dodala firma Lingea ve své vlastní notaci. Zápisy fonémů bylo potřeba převést do formy, se kterou pracuje rozpoznávač. Byla sestavena převodní tabulka mezi těmito notacemi. Obě sady mají jiný počet fonémů a jejich mapování na sebe není úplně jednoznačné. Byla vybrána varianta, která experimentálně vykazuje nejlepší výsledky. Pro převod výslovnostních slovníků z Lingea notace byl vytvořen perlový skript, který pracuje s regulárními výrazy. Převedený slovník obsahuje 994 slov, každé z nich má jednu výslovnostní variantu.

Příklad přepisu:

```
cognitive: ('kQgnItIv) => COGNITIVE k aa g n ih t ih v
```

Modely představují kontextově závislé fonémy, jejich trénování bylo provedeno na 270 hodinách telefonních dat (americká angličtina). Modely pak byly adaptovány na 70 hodinách meetingových dat nahraných na International Computer Science Institute (ICSI) v Berkeley.

Dekodér si podle gramatiky a slovníků sestaví síť modelů, kterou pak prochází pomocí Viterbiho algoritmu a hledá nejlepší možnou cestu. Nakonec zobrazí N nejlepších (N-best) variant.

3. VÝSLEDKY

Rozpoznávač byl testován na 9 sadách 8 kHz nahrávek. Úspěšnost byla vyhodnocována na třech úrovních:

- 1-best: rozpoznané slovo je přesně rovno referenčnímu.
- 5-best: referenční slovo se vyskytuje v seznamu 5-ti nejlepších variant.
- 10-best: referenční slovo se vyskytuje v seznamu 10-ti nejlepších variant.

Úspěšnost 1-best varianty se pohybuje kolem 50%. Pro tento projekt se však lépe hodí 5 best (a vyšší) varianta, uživatel totiž bude mít možnost vybrat z několika slov to správné. Výsledky jsou uvedeny v tabulce 1.

sada	*1-best*	*5-best*	*10-best*
1	41,9	64,3	73,2
2	34,6	54,4	61,1
3	66,4	84,5	89,1
4	65,4	80,7	86,1
5	38,2	56,4	63,9
6	61,9	80,6	84,8
7	73,2	86,7	90,0
8	54,1	76,1	82,2
9	51,7	69,9	78,4
průměr	54,2	72,6	78,8

Tab. 1: Úspěšnost (accuracy) měřená jako poměr počtu správně rozpoznaných slov vůči všem slovům udávaná v %.

4. ZÁVĚR

S 8 kHz modely bylo dosaženo dobrých výsledků, neboť jsme použili kvalitní modely, natrénované na velkém množství dat. Při 1-best dosahujeme v průměru úspěšnosti 54%, při 5-best je průměr úspěšnosti všech testovaných sad 72% a při 10-best 78%. Systém je proto vhodné používat s 5-best nebo vyššími variantami.

Další práce spočívá v použití 16 kHz modelů a technik pro normalizaci a adaptaci na mluvčího (VTLN – vocal tract length normalization), které by měly zvýšit úspěšnost rozpoznávání. Bude použit velký výslovnostní slovník, který má řádově desetitisíce slov s několika výslovnostními variantami. Tím dojde ke zvýšení náročnosti rozpoznávání a pravděpodobně však také k mírnému snížení úspěšnosti rozpoznávání. N-best variant však může obsahovat správné slovo i přes výskyt velmi podobných slov ve slovníku nebo při horší výslovnosti daného slova.

PODĚKOVÁNÍ

Tato práce byla částečně podporována Ministerstvem průmyslu a obchodu České republiky pod projektem č. FT-TA3/006, Grantovou agenturou České republiky pod projektem č. 102/05/0278 a Ministerstvem školství, mládeže a tělovýchovy České republiky pod projektem č. MSM0021630528. Hardware použitý pro tuto práci částečně poskytl CESNET pod projekty č. 119/2004, č. 162/2005 and č. 201/2006.

LITERATURA

[1] Psutka, J., Müller, L., Matoušek, J., Radová, V.: Mluvíme s počítačem česky, Academia Praha 2006, ISBN 80-200-1309-1

[2] Gold, B., Morgan, N.: Speech and audio signal processing, John Wiley & Sons, 2000, ISBN 0-471-35154-7

[3] Young, S., Jansen, J., Odell, J., Ollason, D., Woodland, P.: The HTK book, Entropics Cambridge Research Lab., 2002, Cambridge, UK