SEASR-ENG

ROZPOZNÁVAČ ŘEČI PRO ÚČELY VYHLEDÁVÁNÍ V ANGLICKY MLUVENÉ ČÁSTI KORPUSU

uživatelská a instalační příručka

Obsah

1 Úvod

SEASR-ENG slouží ke zpracování anglicky mluvených nahrávek. Vytváří z nich databázi, se kterou MCLASS (WFBAS) dále pracuje.

Rozpoznávač řeči s modely pro angličtinu pro účely vyhledávání relevantních slov či krátkých frází v archivu přeživších holocaust, spravovaném USC (University of Southern California) Shoah Foundation Institute (http://dornsife.usc.edu/vhi/). Tento archiv obsahuje více než 110 tisíc hodin záznamů v 32 jazycích, přičemž přibližně polovina těchto rozhovorů je vedena v angličtině. Česká část archivu obnáší zhruba jeden tisíc hodin.

Standardní systém rozpoznávání řeči sestává z akustického modelu, modulu pro parametrizaci řeči a jazykového modelu. Akustické modely v našem systému jsou založeny na architektuře skrytých Markovových modelů (HMM), která představuje "lege artis" přístup v současném rozpoznávání mluvené řeči. Jsou použity standardní třístavové akustické modely s Gaussovskými směsmi. Model bere v úvahu trifónové závislosti včetně mezislovních. Řeč je parametrizována pomocí 15 PLP koeficientů a jejich delta a delta-delta derivací (tj. vektor příznaků má dimenzi 45). Příznaky jsou extrahovány 100x za vteřinu a je aplikována kepstrální normalizace na úrovni řečníka.

Systém obsahuje též zobecněný model ticha a při jeho tvorbě byly použity špičkové metody pro adaptivní a diskriminativní trénování.

Jednou z klíčových komponent systému pro rozpoznávání spontánních promluv uložených ve zpracovávaném archivu je také modul pro automatickou segmentaci akustického signálu. Nahraný stereo signál totiž teoreticky sice obsahuje řeč moderátora (zpovídajícího) v jednom kanálu a přeživšího (zpovídaného) v kanálu druhém, ale v praxi dochází k tzv. přeslechům, kdy oba kanály obsahují oba dva zvukové „proudy" s různou intenzitou. Pro dobré výsledky rozpoznávání je nezbytné správně vybrat ten kanál, ve kterém je signál právě hovořícího řečníka kvalitnější. Byl proto vyvinut modul, který ve vstupním signálu na základě výpočtu krátkodobé energie signálu a k-means shlukovací metody takovéto vhodné úseky označí.

Techniky pro extrakci příznaků a algoritmy pro tvorbu akustických modelů byly vybrány na základě zkušeností s rozpoznáváním v jiných úlohách a nebyly předmětem výzkumu v tomto projektu. Byla však věnována zvýšená pozornost technikám adaptace akustického modelu na konkrétního řečníka, neboť v archivu je od každého řečníka uloženo minimálně 30 minut řeči (v průměru ovšem více než 2 hodiny) a tak je adaptace systému na každého řečníka logickou volbou.

V „produkční“ verzi SEASR-ENG byl použit osvědčený jazykový model založený na lineární interpolaci trigramových pravděpodobností získaných z:

- přepisů části rozhovorů (tyto přepisy byly pořízeny primárně pro účely trénování akustického modelu) - slovník cca 30 tisíc (různých) slov, více než 2 miliony slov v textu
- databáze Google N-grams – slovník cca 230 tisíc slov Tento model byl zvolen pro svoji jednoduchost a robustnost.

Bližší informace o zpracování lze nalézt ve článku[1].

[1] *System for fast lexical and phonetic spoken term detection in a Czech cultural heritage archive.* [dokument ve formátu PDF] dostupný z: http://www.asmp.eurasipjournals.com/content/2011/1/10

2 Použití

Pomocí software SEASR-ENG je vytvářena databáze nahrávek. Vzhledem k tomu, že vytvoření takové databáze je jednorázová akce a během jejího používání již není potřeba software spouštět, je vhodné používat ho jako samostatný program.

Předkládaný vstupní soubor (v příkladu `example.wav`) musí mít jeden kanál, vzorkovací frekvenci 16kHz a rozlišení 16 bitů.

Příklad použití:

```
Recognition.exe -set AMALACH_EN.SET -file example.wav -result example.mlf
```

Software byl navržen speciálně pro zpracování anglické části archivu AMALACH a jeho použití pro jiná data bez dalších úprav se nepředpokládá.

3 Instalace

Software se pouze kopíruje na požadované umístění.