Technologická agentura
České republiky

Odborná příloha k průběžné zprávě popisující výsledek TE01020197DVV001

Název pracovního balíčku č. 5.
Hlasové systémy pro interakci člověka se stroji

Předkládá:

Název organizace: ZČU v Plzni
Jméno řešitele: prof. J. Psutka

Název organizace: SpeechTech, s.r.o.
Jméno řešitele: doc. L. Müller

Název výsledku pro vložení do RIV:
Korpus anotovaných akustických řečových dat z domény politických diskusních pořadů.

Identifikační číslo výsledku:
TE01020197DVV001

Pro výzkum a vývoj systému podporující elektronickou archivaci a vyhledávání v audiovisuálních archivech je vhodné stanovit vhodnou modelovou úlohu, pro kterou je dále možné zkoumat nejen metody pro vyhledávání v archívech, ale i metody rozpoznávání řeči, adaptace na řečníka a na kanál, případně i metody pro detekci různých neřečových segmentů a událostí. Z těchto důvodů je vhodné uvažovat především takové úlohy, ve kterých:

- je bohatá skladba mluvčích
- se vyskytují různé vložené audiovizuální předěly (jingly)
- se nachází segmenty, kde se mluvčí překrývají

Z těchto důvodů byla vybrána doména politických diskusních pořadů, které splňují výše uvedené podmínky.

Po analýze dostupných datových zdrojů byla získána sada obsahující 82.5 hodiny televizních záznamů politických diskusních pořadů. Zdrojová sada byla uložena na discích formátu DVD. Proto jedním z prvních kroků pro následující zpracování byla automatická konverze do následujících formátů:

- MP4 (MPEG-4 part 10) – pro následující prezentaci vybraných záznamů prostřednictvím webového rozhraní (prohlížeče Microsoft Internet Explorer, Google Chrome)
- OGV (OGG Vorbis audio / Theora video) – opět pro prezentaci prostřednictvím webového rozhraní (prohlížeče založené na Mozilla Firefox)

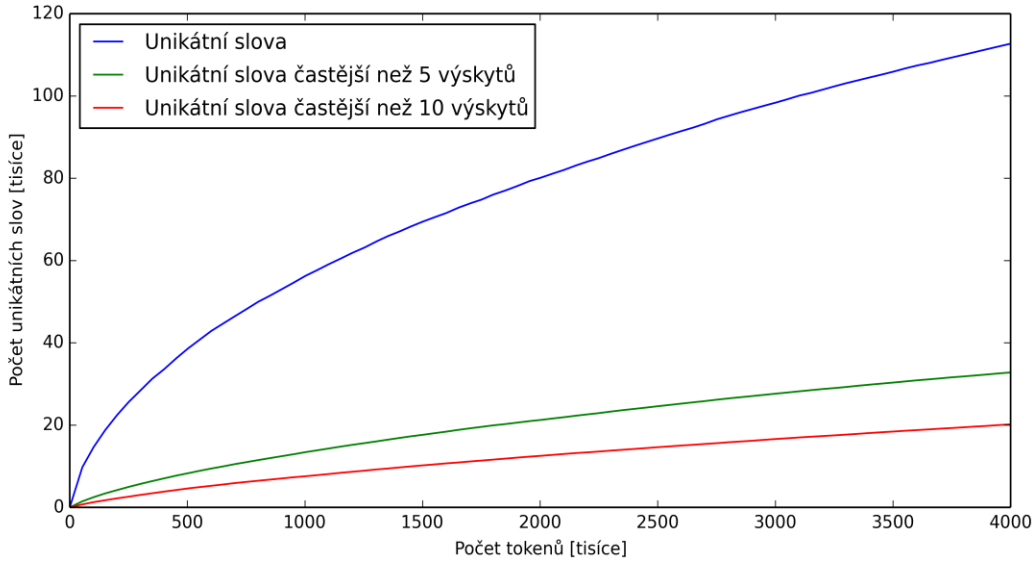
- WAV (PCM, 22.05 kHz 16 bit) – pouze audio kanál; pro použití v technologii automatického rozpoznávání řeči.

Požadavek na dva různé video formáty vzešel z potřeby podporovat širokou paletu webových prohlížečů, které k vyvíjenému archivačnímu a vyhledávacímu nástroji budou v budoucnu přistupovat. Zatímco část webových prohlížečů reprezentovaná Microsoft Internet Explorerem a Google Chrome zpracovává video data uložená v kontejneru MP4 s video kodekem H.264, druhá část prohlížečů – především Mozilla Firefox – pak vyžaduje uložení těchto dat do kontejneru s otevřenou specifikací OGV a s video kodekem Theora. Formát WAV je pak standardním formátem používaným napříč celou řadou nástrojů pro automatické rozpoznávání řeči a pro další analýzy řečového signálu. V celém konverzním procesu byl kladen maximální důraz na zachování časově synchronního zarovnání všech výsledných formátů.

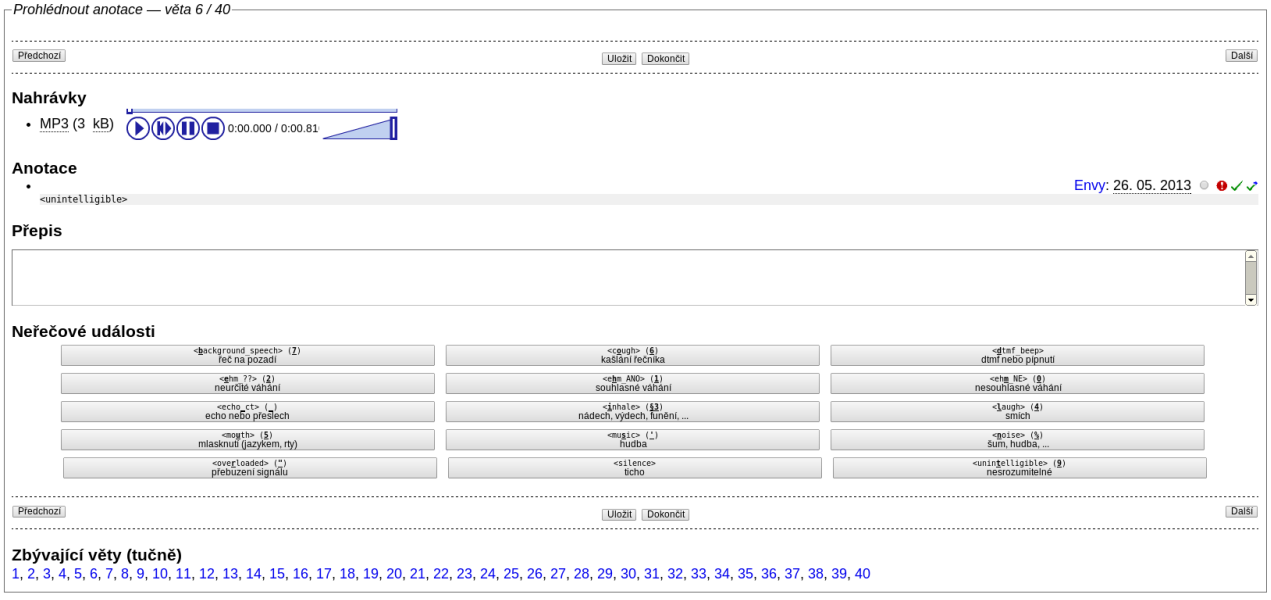
Po konverzi vstupních dat bylo přistoupeno k transkripci těchto dat do textového formátu a zhotovení výslovnostního slovníku. Byl použit standardní postup, při kterém je používám software WebTransc. Ten je výhodný pro zpracování audio dat týmem anotátorů, neboť umožňuje automatizovanou distribuci datových sad k anotaci a následně umožňuje jejich jednoduchou správu a kontrolu, to vše prostřednictvím internetového prohlížeče a účtů vytvořených v rámci tohoto prostředí. Druhým z nástrojů, který byl využit pro získání výslovnostního slovníku je program LMEdit, který opět umožňuje týmovou anotaci různých slov vyskytujících se v textovém přepisu. Při této anotaci je kladen důraz jednak na uvedení správné výslovnosti/výslovností, dále pak na opravu případných překlepů v rámci vybraného slova. Rovněž je možné jednotlivým slovům přiřadit tzv. příznaky, které slouží k bližšímu určení slova (např. jméno, geografické jméno, zkratka, číslo) a následně tato informace může být využita pro vybudování jazykového modelu založeného na třídách.

Výsledný korpus obsahuje 89 řečníků, textová anotace pak 254 298 vět a 4 061 382 tokenů (běžných slov). Součástí korpusu je výslovnostní slovník 113 567 slov vyskytujících se v korpusu, který obsahuje 566 238 výslovnostních variant a 11 903 multislov.

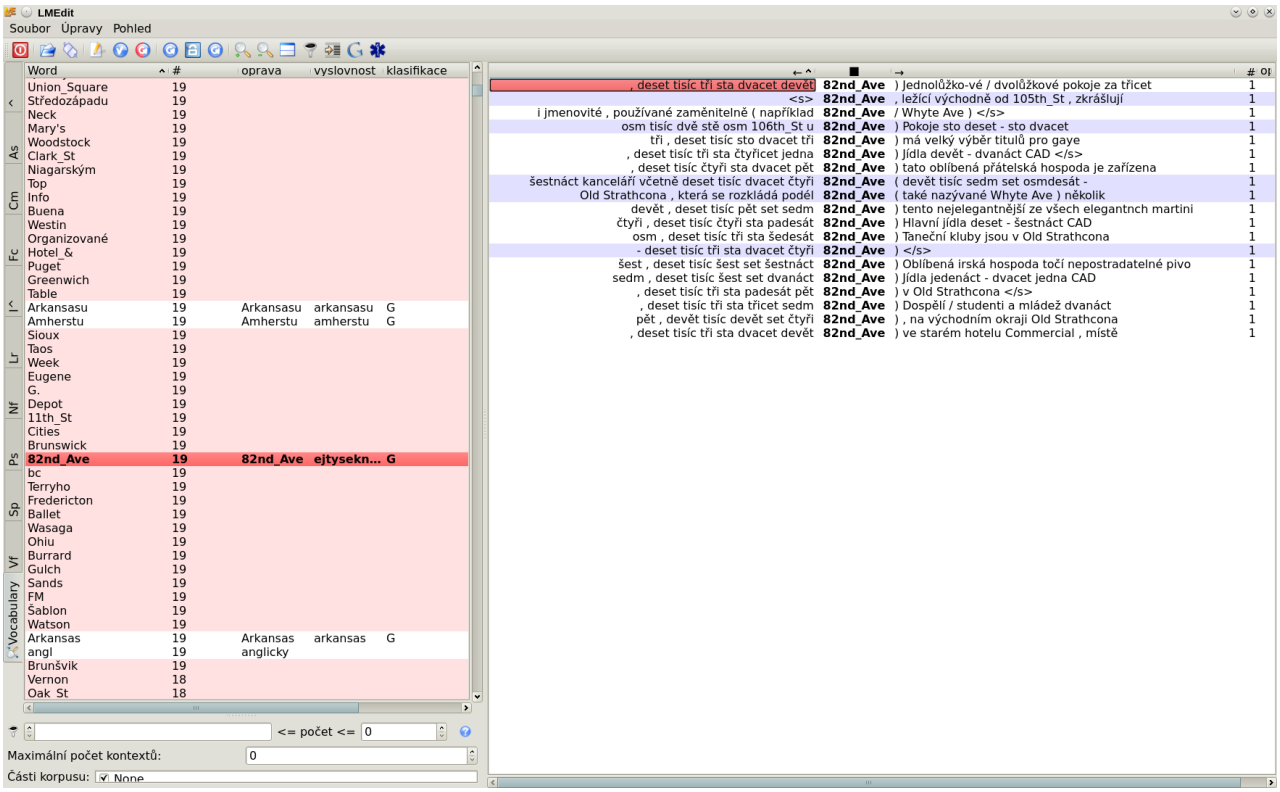
Na základě takto anotovaných dat je možné sestavit graf vykazující závislost velikosti slovníku (počtu unikátních slov) na celkovém počtu tokenů (běžných slov). Tato závislost je vynesena jako modrá křivka na grafu na Obrázku 1. V tomtéž grafu jsou pak zobrazeny i závislosti počtu slov častějších než 5, resp. 10 výskytů. Je vidět, že mezi 4 milióny tokenů je možné najít přibližně 110 tisíc unikátních slov (poznamenejme, že ne všechna tato slova jsou součástí výslovnostního slovníku, viz zpracování výslovnostního slovníku v programu LMEdit). Z těchto 110 tisíc unikátních slov je jich však pouze 30 tisíc častějších než 5 výskytů a přibližně 20 tisíc častějších než 10 výskytů. Tyto statistiky jsou zásadní pro další vývoj systému automatického rozpoznávání řeči, neboť je nutné použít metody, které zajišťují robustní odhady pravděpodobnosti výskytu i u slov, jejichž četnost je nízká – viz cca 80 tisíc slov s četností 4 a méně.



Obrázek 1: Závislost počtu unikátních slov na počtu tokenů. Zelená a červená křivka pak vyjadřuje počet unikátních slov častějších než 5, resp. 10 výskytů.



Obrázek 2: Snímek obrazovky webového anotačního nástroje WebTransc.



Obrázek 3: Snímek obrazovky nástroje pro anotaci rozpoznávacích slovníků LMEdit.

Výsledek „Korpus anotovaných akustických řečových dat z domény politických diskusních pořadů“ (TE01020197DVV001) je nutným mezikrokem k vývoji systému umožňujícího automatickou archivace a následné prohledávání audiovisuálních archívů. Data obsažená v této databázi umožňují výzkum a vývoj potřebných metod založených na rozpoznávání řeči a dalších příbuzných a nadstavbových technologií. Statistiky uvedené v této zprávě ukazují klíčové problémy související s výzkumem metod jazykového modelování (častý výskyt slov, pro které nejsou dostupné statisticky významné frekvence výskytu) stejně jako s metodami akustického modelování (množství různých mluvčích s vysokou variabilitou). Dále zvolené formáty sloužící k reprezentaci audiovisuálních dat umožní pozdější snadnou implementaci vyhledávacího nástroje pomocí moderních webových technologií dostupných ze současných webových prohlížečů.