Studentská Vědecká Konference 2011

ZAROVNÁVÁNÍ AUDIA A TEXTU PŘI VYUŽITÍ NOVÝCH ZDROJŮ DAT PRO AKUSTICKÉ MODELOVÁNÍ

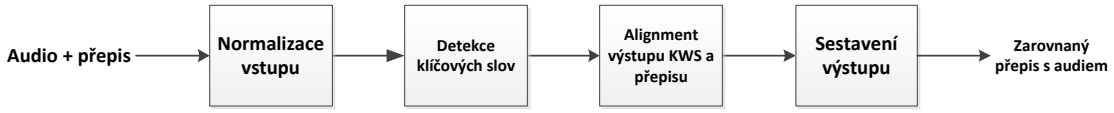
Petr STANISLAV [1]

1 ÚVOD

Pro trénování akustického modelu, viz Psutka (2006), využívaného k automatickému rozpoznávání řeči (ASR) je nezbytné mít k dispozici dostatečné množství správně anotovaných audio dat. Kontrola a oprava anotací je prováděna člověkem na základě obsahu jednotlivých nahrávek. Tento postup je ale velmi časově a finančně náročný, proto je množství zdrojů dat značně omezené. Z tohoto důvodu byl vytvořen systém, jehož funkce spočívá v přiřazení anotací odpovídajícím částem audio dat. Systém detekuje klíčová slova obsažená v náhrávkách a hledá nejdelší společnou podposloupnost anotace a výsledků detekce klíčových slov.

2 PRINCIP ZAROVNÁNÍ AUDIO DAT A PŘEPISU

Systém zajišťující zarovnání audia a přepisu je založen na detekci klíčových slov a hledání jejich nejdelší společné podposloupnosti. Obr. 1 znázorňuje jednotlivé části systému.



Obrázek 1: Blokové schéma systému zarovnávající audio a text

Základním předpokladem využití tohoto systému je nezávislost na podobě vstupních dat. Z toho důvodu je nejprve nutné provést normalizaci jejich formátu, která mimo jiné zahrnuje přepis všech arabských číslic do slovní podoby. Algoritmus přepisu číslic je velmi jednoduchý, jelikož všechny číslice převede do podoby číslovek základních v prvním pádu (např. 123 → *jedna dva tři*). Následuje hledání klíčových slov v nahrávce, jimiž jsou slova z přepisu. Výstupem detektoru klíčových slov je sekvence nalezených slov s příslušnými časovými značkami a mírou důvěry ve správnost detekce. Získaná množina klíčových slov zpravidla obsahuje značné množství redundantních výsledků a zároveň v ní některá klíčová slova chybí, proto je nutné provést zarovnání výstupu detektoru a přepisu. Tento krok je realizován pomocí metody *LCS* (**L**ongest **C**ommon **S**ubsequence), která hledá nejdelší společnou podposloupnost mezi dvěma řetězci při dodržení kontinuity znaků (znaky v obou řetězcích musí dodržovat pořadí). Konkrétně se jedná o Hirschbergův algoritmus uvedený v Hirsberg (1975), který je upraven tak, aby nalezl všechny možné nejdelší podposloupnosti (všechny mají stejnou délku) a z nich je vybrána ta, která má

[1] Petr STANISLAV, student doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, specializace Umělá inteligence, e-mail: pstanisl@kky.zcu.cz

nejvyšší míru důvěry ve správnost detekce. Na závěr je sestaven výstup, který vyhovuje požadavkům pro následné vytváření akustických modelů (např. XML formát obsahující časové značky výskytu slov a originální přepis zarovnaný s audiem).

3 VÝSLEDKY ZAROVNÁNÍ AUDIA A TEXTU

Funkčnost systému pro zarovnání audia a textu byla otestována na čtyřech skupinách vstupních dat, a to na telefonních hovorech (nesouvislá promluva jednoho řečníka mluvícího přirozenou řečí), záznamech rozhlasového vysílání s uveřejněným přepisem (téměř souvislá promluva obsahující minimum ruchů s velmi přesným až nadrámcovým přepisem), záznamech zpravodajské relace (více řečníků, přepisem jsou skryté titulky tvořené v reálném čase) a válečném dokumentu (na popředí zvukové stopy obsahující velké množství ruchů řečník, přepisem jsou velmi přesné skryté titulky).

V tab. 1 jsou uvedeny procentuální poměry správně a špatně zarovnaných slov ve zpracovaných skupinách vstupních dat. Za správně zarovnané slovo je považováno takové, jehož časové značky přesně odpovídají výskytu v audio datech (slovo obsažené v přepisu bylo správně zarovnáno s audiem). Naopak špatně zarovnané slovo je takové, které není nalezené nebo má špatně časové značky. Kontrola zarovnání byla provedena ručně pro každé slovo obsažené v přepisu.

Skupina	Správně zarovnaná slova [%]	Špatně zarovnaná slova [%]
Telefon	42, 3	57, 7
Rozhlas	74, 4	25, 6
TV zpravodajství	67, 1	32, 9
Dokument	83, 3	16, 7

Tabulka 1: Výsledky zarovnání audia a textu

4 ZÁVĚR

Jak je vidět v tab. 1, v případě záznamu telefonního hovoru bylo správně zarovnáno méně než poloviční množství slov. To bylo způsobeno zejména nezřetelnou artikulací mluvčího. U nahrávky rozhlasového pořadu lze poměrně vysoké procento správně zarovnaných slov přiřknout hlasité zřetelné promluvě řečníka, procento špatně zarovnaných slov pak nadbytečným slovům obsaženým v přepisu a špatnému zpracování internetových adres, zkratek a číslic v průběhu normalizace. V případě záznamu televizního zpravodajství hraje kladnou roli kvalitní promluva řečníka. Naproti tomu je výsledek negativně ovlivněn nekvalitním přepisem vytvářeným současně s promluvou. U válečného dokumentu je dosaženo velmi dobrých výsledků nejen díky kvalitní promluvě řečníka, ale i velmi kvalitnímu přepisu, přestože na pozadí nahrávky vystupuje množství ruchů a řeči.

Na základě získaných výsledků lze říci, že systém je schopen uspokojivě zarovnávat audionahrávku a přepis, pokud jsou mu předložena dostatečně kvalitní vstupní data.

REFERENCE

Psutka J., Müler L., Matoušek J., Radová V., 2006. *Mluvíme s počítačem česky.* Academia, Praha.

Šmíd L., 2005. *Metody rychlé detekce klíčových slov*, Vol. 3. pp 149-167.

Hirschberg, D. S., 1975 *A linear space algorithm for computing maximal common subsequences*, Communications of the ACM, Volume 18 Issue 6