Studentská Vědecká Konference 2010

DETEKCE CHYBNÝCH HRANIC V AUTOMATICKÉ FONETICKÉ SEGMENTACI ŘEČI

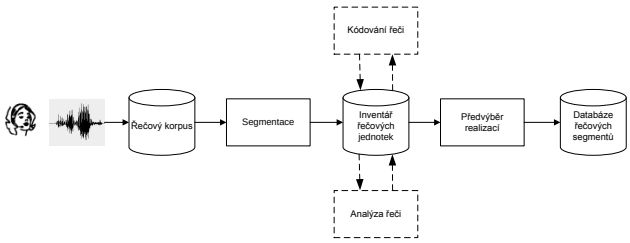
Ladislav KAŠPAR [1]

1 ÚVOD

Úkolem tohoto projektu bylo navržení algoritmu, který by detekoval chyby v automatické segmentaci řeči. Díky němu by bylo možné odstranit chybně segmentované jednotky z databáze řečových jednotek a tím vylepšit syntetizovanou řeč systému ARTIC, vyvíjeného v současné době na FAV KKY ZČU v Plzni. Tento projekt se tedy nezabývá přímo syntézou řeči, ale spíše přípravou databáze řečových jednotek. Jde tedy především o automatickou segmentaci řeči, díky které vzniká již zmiňovaná databáze.
TTS (text-to-speech) systém ARTIC je založen na konkatenační syntéze řeči. To v jednoduchosti znamená, že jednotlivé zvuky z databáze řečových jednotek (zpravidla zvuky odvozené z hlásek české fonetické abecedy - tzv. difony) jsou řetězeny za sebe a vznikají tak slova, následně věty a na konec celé syntetizované promluvy. Je tedy zřejmé, že přesnost automatické segmentace do značné míry ovlivňuje kvalitu syntetické řeči vytvářené konkatenačním systémem.

2 SEGMENTACE ŘEČI

Segmentace je proces, během kterého se hledají hranice akustických řečových jednotek v řečových promluvách. Vzhledem k množství, řádově desítky hodin, promluv nelze segmentaci provádět ručně. V dnešní době se nejčastěji využívají dva přístupy, kterými lze řeč segmentovat automaticky. Je to metoda skrytých Markovových modelů (HMM), nebo metoda dynamického borcení času (DTW). Jak fungují, se můžete dočíst např. v [1]. Systém ARTIC využívá prvně zmiňovanou metodu HMM. Na obrázku 1 je znázorněno blokové schéma procesu vytváření databáze řečových segmentů. Vstupem detekčního algoritmu (této práce) je však soubor, který je výstupem automatické segmentace, textový soubor ve formátu ASF (ARTIC Segmentation File), v němž lze nalézt informace o každé jednotce z řečového korpusu. Nás však budou v tuto chvíli zajímat časové údaje o každém segmentu, tedy čas jeho startu a konce, resp. jeho trvání.



Obrázek 1: Blokové schéma procesu vytváření databáze řečových jednotek

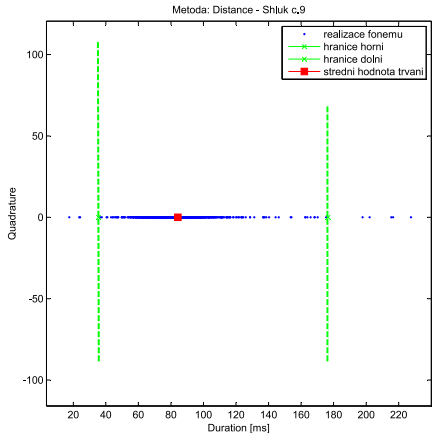
3 DETEKCE CHYBNÝCH HRANIC

Protože jedinou informací o automatické segmentaci jsou údaje o hranicích jednotek (uložené v ASF), snažili jsme se tuto informaci, tj. znalost délek, resp. trvání jednotek ve zdrojových datech, využít k detekci chybně segmentovaných jednotek. Délka trvání segmentu je totiž ovlivněna hned několika okolnostmi. Velice záleží na pozici realizace jednotky ve slově, ve větě, ale i na předcházejících a následujících realizacích [2]. Pokud bychom tedy vzali např. foném [a] (ten je reprezentován několika tisíci realizacemi "a"v souboru ASF) jako jednu skupinu všech jeho realizací, těžko bychom mohli najít chyby podle délky jejich trvání. Variabilita různých realizací jednoho fonému může být velká, zvláště pak u samohlásek. Použili jsme tedy shlukovací analýzu a foném rozdělili do několika skupin (shluků),u kterých jsme předpokládali

[1] Ladislav Kašpar, Bc. Ladislav Kašpar, student navazujícího studijního programu Aplikované vědy a informatika, obor Kybernetika a řídící technika, e-mail: kasparla@students.zcu.cz

stejné vlastnosti (např. u skupiny realizací fonému [a], které se vyskytují v koncovkách "la", v posledních slovech věty). Každý shluk pak reprezentuje určitou skupinu realizací, které mají podobné kontextové vlastnosti a měly by tedy mít i podobnou délku trvání. Nyní jsme byli schopni v rámci shluku označit ty realizace jednotky, které se příliš vzdalují od střední hodnoty trvání segmentů v příslušném shluku. Ve své bakalářské práci [3] jsem navrhl metodu vzdálenosti, která označila v každém shluku ty realizace fonému, jejichž trvání bylo extremální v porovnání se střední dobou trvání ostatních segmentů ve shluku. Na obrázku 2 je graf, který zobrazuje konkrétní shluk a jeho rozdělení metodou vzdálenosti na tzv. outliery a segmenty, které nejsou podezřelé. Outliery budeme považovat za potenciální chyby.

Smyslem práce je detekovat segmenty, jejichž hranice jsou určeny chybně, a tedy jejich řečový signál zasahuje do okolních jednotek, čímž vznikají chyby ve výsledné syntetizované řeči. Testováním a ruční kontrolou algoritmem označených realizací fonému [a] a [t] jsem zjistil, jak je metoda vzdálenosti efektivní. U fonému [a] jsem dosáhl 56% a u [t] 78% úspěšnosti. Tato procenta vypovídají o tom, že pouze 56 % a 78 % z algoritmem detekovaných segmentů je opravdu chybných. Pokud bychom tedy z výstupu segmentace tyto segmenty odstranili, přišli bychom i o mnoho bezchybných realizací fonému. Proto jsme se pokusili algoritmus ještě dále vylepšit. Přidali jsme statistické metody známé pod názvy "Five-number summary"a "Grubbův test na outliery"[4]. Průnikem metod jsme získali užší skupinu outlierů a dosáhli tak 81% efektivnosti u fonému [a]. V poslední řadě jsme se pokoušeli na místo trvání segmentu počítat jeho krátkodobou energii a tu pak analogicky k trvání porovnávat se střední hodnotou segmentů ve shluku. Bohužel se ukázalo, že energie je příliš variabilní a nedá se v tomto ohledu použít.



Obrázek 2: Vizualizace konkrétního shluku - metoda vzdálenosti, modre puntíky - realizace fonému, červený čtverec - střední hodnota trvání, zeleně čárkovaně - práh podezřelé/bezchybné

4 ZÁVĚR

Bylo dosaženo 81% úspešnosti detekce chybně segmentovaných hranic fonému [a]. K tomu však bylo zapotřebí na základě analýzy výsledků algoritmu ručně nastavit prahy pro určení outlierů. Vzhledem k velké pracnosti (nutno opakovat pro každý foném, protože prahy se mohou lišit, a je třeba zohlednit i velký počet shluků) se popsaný postup jeví jako neperspektivní a příliš zdlouhavý. Zdá se, že pouhá informace o trvání a energii segmentů není dostačující.

PODĚKOVÁNÍ: Příspěvek byl podpořen grantovým projektem SGC-2010-054 (Inteligentní metody strojového vnímání a porozumění)

REFERENCE

[1] J. Psutka, L. Müller, J. Matoušek, and V. Radová. *Mluvíme s počítačem česky.* Academia, Prague, 2006.

[2] J. Kominek and A. Black. Impact of durational outlier removal from unit selection catalogs In: Proceedings of the 5th speech synthesis workshop (SSW5), Pittsburgh, USA. pages 155 – 160, 2004.

[3] L. Kašpar. Detekce chybných hranic v automatické fonetické segmentaci řeči., Bakalářská práce. 2009.

[4] L. Kašpar. Detekce chybných hranic v automatické fonetické segmentaci řeči., Výzkumná zpráva, KKY FAV ZČU. 2009.