FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2014

Optimalizace rychlosti výběru řečových jednotek v konkatenační syntéze řeči

Jiří Kala[1]

1 Úvod

Tato práce se zabývá urychlením procesu výběru jednotek v úloze syntézy řeči z textu, konkrétně v systémech na bázi konkatenační syntézy. Tento způsob generování syntetické řeči využívá dopředu připravenou databázi segmentů reálné řeči, které jsou podle požadavků na vstupu syntetizéru zřetězeny (spojeny – angl. concatenate) do výsledné promluvy. U obecného systému převádějícího libovolný text na řeč (TTS z angl. Text-to-Speech) se používají velmi krátké jednotky jako jsou difony, trifony nebo polofony (Psutka *et al.*, 2006, str. 555 až 557). Vysoce kvalitní a přirozeně znějící syntetická řeč vyžaduje velmi rozsáhlou databázi – např. systém ARTIC[2], na němž byly prováděny veškeré experimenty, používá databázi s více než 650 tisíci řečovými segmenty. V takto obsáhlé databázi se pro každou požadovanou jednotku dané promluvy vyskytují až tisíce segmentů (kandidátů) a TTS systém z nich v době syntézy vybírá optimální posloupnost těch nejvhodnějších. Pro každou promluvu je tak nutné provádět desítky až stovky milionů výpočtů cen řetězení (popsáno dále). Cílem práce bylo nalézt způsob jak urychlit proces výběru optimální posloupnosti kandidátů resp. snížit nároky na výpočetní výkon, a to při zachování kvality generované řeči.

2 Hledání optimální posloupnosti

Při hledání optimální posloupnosti kandidátů se pracuje se dvěma hodnotícími funkcemi. První z nich je tzv. cena cíle C^{t} (angl. target cost) a vyjadřuje jak vhodný či nevhodný je kandidát s ohledem na jeho umístění v promluvě, okolní jednotky apod. Druhou funkcí je cena řetězení C^{c} (angl. concatenation cost),která vyjadřuje jak dobře či špatně lze dva kandidáty dvou sousedících jednotek zřetězit. Celková (kumulativní) cena posloupnosti kandidátů je pak dána součtem jejich cen cíle a cen řetězení. Kandidáty lze uspořádat do pomyslného grafu, kde kandidáti tvoří uzly a hrany grafu reprezentují jejich spojení (viz obrázek 1). Optimální posloupnost kandidátů pak odpovídá cestě grafem kandidátů minimalizující celkovou cenu C.

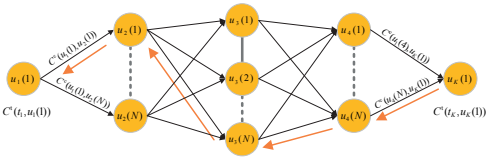
2.1 Základní Viterbiův algoritmus

K vyhledání optimální posloupnosti se typicky využívá Viterbiův algoritmus, který postupně prochází jednotlivé kandidáty syntetizovaných jednotek, kdy pro každého z nich vypočítá cenu řetězení se všemi předchůdci a uloží si odkaz na předchůdce dávajícího nejnižší kumulativní cenu. Po vyhodnocení všech kandidátů se vyhledá kandidát poslední jednotky s nejnižší kumulativní cenou a zpětným trasováním přes nejlepší předchůdce se naleznou kandidáti optimální sekvence. Výhodou Viterbiova algoritmu je to, že vždy zaručuje nalezení posloupnosti kandidátů s globálně minimální kumulativní cenou. Naopak zásadní nevýhodou je vysoká výpočetní náročnost způsobená tím, že se vyhodnotí ceny řetězení pro všechny přípustné spo-

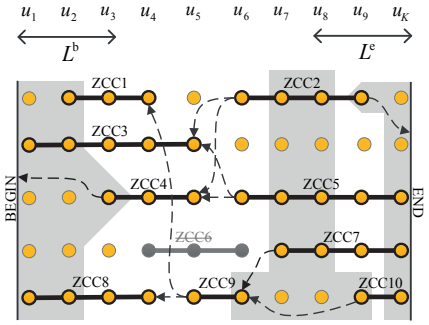
[1] student navazujícího studijního programu Aplikované vědy a informatika, obor Kybernetika a řídící technika, specializace Umělá inteligence, e-mail: jkala@kky.zcu.cz

[2] Systém vyvíjený na Katedře kybernetiky Západočeské univerzity v Plzni (Matoušek *et al.*, 2006)

jení grafu kandidátů. Výpočtů cen řetězení je u rozsáhlých databází obrovské množství, a jejich vyhodnocení zabírá přibližně 90 % výpočetního času celého procesu syntézy řeči.



Obrázek 1: Schéma základního Viterbiova algoritmu.



Obrázek 2: Schéma algoritmu hledání optimání posloupnosti s využitím ZCC řetězců.

2.2 Optimalizace využitím ZCC řetězců

Při hledání možností jak omezit množství výpočtů cen řetězení, a tím zároveň snížit výpočetní náročnost syntézy řeči, byla provedena analýza optimálních posloupností u 10.000 náhodně vybraných vět. Ta ukázala, že 95 % všech použitých segmentů bylo součástí řetězců tvořených dvěma nebo více řečovými segmenty, které spolu sousedily v původní nahrávce použité k vytvoření databáze řečových segmentů. Důležitou vlastností takových řetězců je to, že kumulativní cena celého řetězce je dána pouze součtem cen cíle v nich obsažených kandidátů, protože cena jejich řetězení je automaticky nulová a není ji nutné počítat. Tyto řetězce byly pojmenovány zkratkou ZCC (z angl. zero–concatenation–cost). Optimalizované hledání nejlepší posloupnosti kandidátů nejprve vyhledá v grafu kandidátů všechny ZCC řetězce. Poté se do množiny ZCC řetězců přidají i všechny jejich možné podřetězce, kvůli možným překryvům původních delších ZCC řetězců. Samotný algoritmus hledání je pak obdobný Viterbiovu algoritmu, pouze místo kandidátů se pracuje se ZCC řetězci (viz schéma na obrázku 2). Pokud na sebe ZCC řetězce přesně nenavazují, pak se vyhledá mezi nimi spojení z původních samostatných kandidátů.

3 Shrnutí

Použitím optimalizovaného algoritmu využívajícího ZCC řetězce bylo dosaženo snížení množství potřebných výpočtů cen řetězení u mužského hlasu 556.45× (438.30× pro ženský hlas). Algoritmus již sice nezaručuje vždy nalezení posloupnosti s globálně minimální kumulativní cenou, nicméně na testovací množině se většina promluv naprosto shodovala s Viterbiovým algoritmem a ostatní věty se lišily maximálně v počtu jednotek vybraných řečových segmentů. Důležitějším měřítkem jsou však poslechové testy, které ukázaly, že kvalita řeči nebyla ovlivněna.

Poděkování

Tato práce vznikla za podpory grantu Západočeské univerzity, projekt č. SGS-2013-032.

Literatura

MATOUŠEK, J., TIHELKA, D. & ROMPORTL, J. (2006). Current state of Czech text-to-speech system ARTIC. In: *Proceedings of 9th International Conference on Text, Speech and Dialogue*. Berlin, Germany: Springer.

PSUTKA, J., MÜLLER, L., MATOUŠEK, J. & RADOVÁ, V. (2006). *Mluvíme s počítačem česky*. Prague, Czech Republic: Academia.