Syntéza řeči z audioknih

Jakub Vít[1]

1 Úvod

Audioknihy jsou velmi snadno dostupný zdroj kvalitních audio dat. Ty lze využít pro vytvoření syntetického hlasu řečníka, který danou knihu namluvil. Tento abstrakt popisuje postup jak takový hlas vytvořit a problémy s ním spojené. Zároveň představuje modifikace algoritmu syntézy řeči, které vylepšují kvalitu syntézy pro hlasy vytvořené tímto způsobem.

2 Jak se tvoří hlas pro syntézu

Syntéza řeči je převod textu na řeč. Hlas pro syntézu zpravidla nahrávají profesionální řečníci ve zvukovém studiu. Nahrávané texty jsou předem připravené tak, aby obsáhly všechny možné kombinace písmen a hlásek, které se v řeči mohou vyskytnout. Nahrávaný hlas je v neutrálním stylu. To znamená, že je klidný, konzistentní a dodržuje jednotná pravidla výslovnosti. Díky je tomu je možné vytvořit syntézu řeči s velmi vysokou kvalitou. Standardní metoda syntézy řeči totiž funguje na principu spojování malých úseků audio signálů z těchto nahraných vět. Tím, že jsou nahraná data konzistentní, lze snáze zajistit hladké napojení zvukových segmentů. Tento proces je však zdlouhavý a nákladný. Nabízí se úvaha zda nelze vytvořit syntetický hlas z nahrávek, které jsou již k dispozici.

3 Audioknihy jako zdroj dat

Jedním z takových zdrojů jsou audioknihy. Ty bývají obvykle také nahrávány ve zvukovém studiu pomocí řečníka profesionála. Velkou výhodou také je, že se k nim dá dohledat textová předloha a také to, že audio záznam bývá velmi dlouhý a obsahuje tak mnoho dat. Na druhou stranu však audioknihy nebývají nahrávány v neutrálním stylu a neobsahují emotivně zabarvené části. Například v přímé řeči, kde má každá postava svojí specifickou barvu hlasu. Řečník také zpravidla mění styl hlasu podle obsahu v textu. Řeč je více živější. Hlasitost a rychlost řeči se mění na základě vyprávěného příběhu.

Textová předloha sice existuje ale pro potřeby syntézy je nutné mít ještě tzv. *zarovnání na jednotky*. To znamená, že je nutné mít pro každé písmeno přesnou časovou značku v jakém čase se vyskytuje v audio signálu. Pro vytvoření takového zarovnání je lze použít například rozpoznávač řeči v tzv. *forced aligment* módu, který takové zarovnání vytvoří. Data je nutné pročištit, neboť takové zarovnání není vždy spolehlivé. Textová podoba musí odpovídat přesně vyslovené řeči. Pokud tomu tak není, může se stát že se do řečového inventáře dostanou špatně nasegmentované jednotky.

[1] student doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, specializace Umělá inteligence, e-mail: vit89@ntis.zcu.cz

4 Úprava algoritmu

Standardní algoritmus syntézy řeči nefunguje příliš dobře s takto vytvořeným hlasem, což je očekávané, neboť na taková data nebyl stavěn. Algoritmus byl proto modifikován tak, aby fungoval lépe. V této úloze byl použit algoritmus *unit selection*. Tento algoritmus pracuje tak, že vybírá vhodnou posloupnost kandidátů jednotek (fónů) z *řečového inventáře*. Každá jednotka mívá v inventáři stovky kandidátů. Lze tedy vytvořit velmi mnoho kombinací. Optimální posloupnost jednotek algoritmus vybírá tak, že minimalizuje dvě ceny: *cenu cíle* a *cenu napojení*. Pro výběr optimální sekvence se používá *Viterbiův algoritmus*.

Cena napojení udává to, jak dobře lze danou jednotku napojit na jednotku předcházející. Počítá se jako vzdálenost parametrů audio signálu v místě napojení. Mezi parametry patří výška hlasivkové frekvence, trvání, energie a spektrum. Pro syntézu z audioknih nebylo třeba dělat v této ceně žádné velké úpravy. Pouze byly změněny váhy jednotlivých parametrů tak, aby odrážely větší variabilitu řeči.

Cena cíle udává jak moc dobře pasuje daný kandidát do konkrétní jednotky v syntetizované větě. Obsahuje fonetický kontext, typ prosodému, poziční parametry či prosodické cíle. Hlavní úprava algoritmu spočívala v tom, že k těmto složkám byla přidána ještě jedna, která ohodnocuje každého kandidáta podle toho, jak moc dobře odpovídá statistickému modelu. Tento model byl vytvořen ze všech kandidátů pro danou jednotku. Model obsahuje akustické parametry: energie, trvání, hlas. frekvence a je reprezentován střední hodnotou a odchylkou těchto akustických parametrů. Odráží tak průměrného očekávaného kandidáta pro danou jednotku. Čím více se odlišují parametry kandidátů od daného modelu, tím je penalizace větší a snižuje se tak šance, že bude daný kandidát vybrán. Tato úprava tak téměř vylučuje to, aby byl pro danou jednotku vybrán kandidát který má extrémně odlišné parametry od ostatních, což je většinou způsobeno špatnou segmentací či úplně jinou akustickou podobou jednotky (např. řečník vyslovil slovo jinak). Rovněž jsou odfiltrovány kandidáti z atypicky vyslovených vět, což jsou většinou emotivně zabarvené úseky, přímá řeč, či jiná změna stylu hlasu.

Tato úprava rovněž umožňuje cíleně modifikovat cenu tak, aby preferovala určité prozodické styly. Lze tak například snadno vynutit rychlejší či pomalejší řeč. Rovněž lze měnit průměrnou výšku hlasivkové frekvence tedy to jak vysoko daný hlas zní.

5 Poslechové testy

Pro ověření přínosu úpravy algoritmu byly zorganizovány poslechové testy. V nich byly posluchačům předkládány dvě varianty téže věty. Jednou vysyntetizované původním systémem a podruhé vysyntetizované upraveným algoritmem. Posluchači měli za úkol vybrat, která z nich je lepší, či zda jsou stejně dobré / špatné. Téměř 70 % odpovědí preferovalo nový systém a 15 % bylo nerozhodných. To potvrzuje, že provedené úpravy vedly k lepší kvalitě syntetizované řeči, pokud se jako zdroj dat použijí audioknihy.

Poděkování

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

Literatura

Prahallad, K., Toth, A. R., Black, A. W., 2007. *Automatic building of synthetic voices from large multi-paragraph speech databases*. in Interspeech.