Počítačová syntéza řeči pomocí umělých neuronových sítí

Bc. František Kynych <frantisek.kynych@tul.cz>, Ing. Petr Červa Ph.D

Tato diplomová práce se zabývá syntézou řeči pomocí umělých neuronových sítí. Cílem bylo prozkoumání a ověření současných přístupů využívajících neuronových sítí a pomocí nejlepší architektury natrénování modelu pro syntézu mužského a ženského hlasu pro český jazyk. Dále porovnání s komerčními systémy a vytvoření demonstrační webové aplikace. Po natrénování vybraných modelů proběhlo porovnání nejlepšího mužského hlasu s komerčními systémy od společností Google, Microsoft a SpeechTech, kde výsledek této práce překonal standardní přístupy nevyužívající neuronových sítí. Nad rámec zadání je v práci řešena fonetická transkripce pro lepší kvalitu syntézy a dále byly provedeny experimenty se syntézou pro více mluvčí s využitím vektorů mluvčího.

Klíčová slova: syntéza řeči, neuronové sítě, syntéza řeči pro více mluvčí, Tacotron 2, WaveGlow

Úvod

Zadáním práce bylo seznámení s problematikou počítačové syntézy řeči, zejména s metodami využívajícími hluboké neuronové sítě. Dále po prozkoumání a ověření současných přístupů natrénování a optimalizace nejlepšího modelu pro syntézu češtiny mužským a ženským hlasem. Poté porovnání výsledného systému s dostupnými daty pro daný jazyk a vytvoření demonstrační aplikace, která umožní generovat řečový signál ze zadaného textu.

Motivací pro řešení práce bylo získání zkušeností s moderními metodami pro syntézu řeči. V Laboratoři počítačového zpracování řeči (SPEECHLAB) se zatím nikdo nezabýval využitím neuronových sítí pro tuto úlohu. Předchozí práce se věnovaly parametrické [1] a zřetězené [2] syntéze, ale výsledná řeč těchto prací nebyla přirozená.

Metodika

Teoretická část práce seznamuje se současně používanými architekturami neuronových sítí pro syntézu řeči. Většina z nich je rozdělena na dvě části, kde první vytváří mel spektogram z přivedené sekvence znaků a druhá část daný mel spektogram převede zpět do časové oblasti. Současné architektury jsou převážně tvořeny z konvolučních a rekurentních neuronových sítí využívajících attention mechanismu. Pro porovnání je využita metrika mean opinion score (MOS), u které jsou jednotlivé systémy hodnoceny pomocí poslechových testů a výsledný MOS je vypočten jako průměr těchto hodnocení.

Na základě výsledků jednotlivých publikací a veřejně dostupných zvukových ukázek byly pro další experimenty vybrány DeepVoice 3 [3], Tacotron 2 [4] a WaveGlow [5] architektury. Tacotron 2 i DeepVoice 3 převádí vstupní sekvenci znaků na mel spektogram. Obě architektury byly vybrány pro další experimenty z důvodu podobného MOS. WaveGlow model byl poté použit pro transformaci mel spektogramu do řečového signálu.

Při provádění experimentů bylo cílem natrénování modelu pro syntézu češtiny mužským a ženským hlasem. Pro trénování byla dostupná data pro 4 osoby, zhruba 20 hodin nahrávek pro každou z nich. Pro dané architektury se využily veřejně dostupné implementace. Experimentovalo se s filtrováním dostupných dat a nastavením hyperparametrů modelu, pro dosažení nejlepší kvality syntetizované řeči a také pro rychlé trénování. Po odstranění méně kvalitních dat a vyladění modelu se podařilo syntetizovat přirozený mužský i ženský hlas. Experimentovalo se i s využitím předtrénovaných modelů pro angličtinu, kdy byly přeučeny pro češtinu a dosahovaly kvalitnějšího výstupu.

Pro lepší výslovnost byla nad rámec zadání řešena fonetická transkripce vstupního textu s využitím Transformer [6] architektury. Dále byla nad rámec zadání rozšířena Tacotron 2 architektura o vektory

mluvčího (tzv X-Vektory), kde byla cílem možnost změny výstupního hlasu dle přivedeného vektoru.

Výsledky a diskuze

Po optimálním natrénování modelů byl vybrán mužský hlas Tacotron 2 a WaveGlow architektury pro porovnání s komerčními systémy pomocí poslechových testů. Pro tyto testy bylo vytvořeno prostředí v demonstrační webové aplikaci. Porovnáván byl výsledek této diplomové práce s komerčními systémy od společností Google, Microsoft a SpeechTech. Hodnocení se celkem zúčastnilo 56 osob a celkem bylo ohodnoceno 1060 nahrávek se stejným obsahem od každého systému.

Tabulka 1: MOS jednotlivých systémů s 95% intervalem spolehlivosti.

Systém	MOS
Google (Tacotron 2, WaveNet)	3.475 ± 0,312
Diplomová práce	3.171 ± 0,294
Microsoft	3.123 ± 0,290
SpeechTech	3.102 ± 0,275
Google (Standard)	2.693 ± 0,294

Natrénovaný systém z této práce je srovnatelný s komerčními systémy pro syntézu řeči v češtině a při stejném vlivu hodnotících se umístil na druhém místě. Výrazně ho překonává Google systém využívající téměř stejnou architekturu, ale vzhledem ke kvalitě použitých trénovacích dat dosáhla diplomová práce překvapivého výsledku.

Při provádění experimentů pro více mluvčí se systém podařilo natrénovat do stavu, ve kterém síť zvolila správné pohlaví dle přivedeného vektoru a mírně modifikovala hlas, který viděla u trénování. Trénování jednoho experimentu pro syntézu hlasu jednoho mluvčího trvalo zhruba týden a pro více mluvčí se časová náročnost zvýšila na dva týdny.

Závěr

Zadání diplomové práce bylo splněno a nad jeho rámec byla řešena fonetická transkripce pro lepší výslovnost systému a také byl upraven Tacotron 2 model o vektory mluvčího pro syntézu řeči pro více mluvčí.

Výsledek diplomové práce ukázal, že je srovnatelný se současně používanými komerčními systémy. V natrénovaném systému lze také např. modelovat nádechy, pauzy v řeči a hezitační zvuky.

Pokračování v práci by bylo možné např. vytvořením profesionálně zaznamenaných nahrávek s vyšší vzorkovací frekvencí pro kvalitnější natrénování modelu. Dále je možné zkoušet další architektury, které přinese výzkum, případně rozšíření současné architektury o možnost modifikace emocí ve výstupní řeči.

Poděkování

Tato práce byla podpořena z projektu Studentské grantové soutěže (SGS) na Technické univerzitě v Liberci v roce 2020.

Reference

[1] ŠILHÁN, Stanislav. Parametrická syntéza české řeči: Parametric synthesis of Czech speech. Liberec: Technická univerzita v Liberci, 2004. Diplomové práce.

[2] ŠKODA, Jan. Zřetězená syntéza řeči pracující s rozsáhlou databází promluv: Concatenation speech synthesis working with large speech databases. Liberec: Technická univerzita v Liberci, 2005. Diplomové práce.

[3] PING, Wei, et al. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654, 2017

[4] SHEN, Jonathan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018. p. 4779-4783.

[5] PRENGER, Ryan; VALLE, Rafael; CATANZARO, Bryan. Waveglow: A flow-based generative network for speech synthesis. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019. p. 3617-3621.

[6] VASWANI, Ashish, et al. Attention is all you need. In: Advances in neural information processing systems. 2017. p. 5998-6008.