TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

SKFM 2020
Studentská konference Fakulty mechatroniky,
informatiky a mezioborových studií

Počítačová syntéza řeči pomocí umělých neuronových sítí

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

ABSTRAKT

V diplomové práci byly ověřeny různé druhy neuronových sítí a nejlepší z nich byly natrénovány a optimalizovány pro syntézu mužského a ženského hlasu v češtině. Výsledný systém byl porovnán pomocí poslechových testů s komerčními systémy pro syntézu českého jazyka a většinu z nich na daných datech překonal. Pro syntézu řeči byla vytvořena demonstrační webová aplikace. Nad rámec zadání byla řešena fonetická transkripce pro lepší výslovnost modelu a syntéza řeči pro více mluvčí.

CÍLE PRÁCE

- Prozkoumání a ověření současně používaných metod pro syntézu řeči využívajících neuronové sítě.
- Natrénování a optimalizace modelu pro syntézu češtiny mužským a ženským hlasem.
- Porovnání kvality vytvořeného systému s komerčně používanými systémy pro syntézu češtiny.
- Vytvoření demonstrační webové aplikace, která umožní generovat řečový signál ze zadaného textu.

POSTUP ŘEŠENÍ PRÁCE

Použitá data

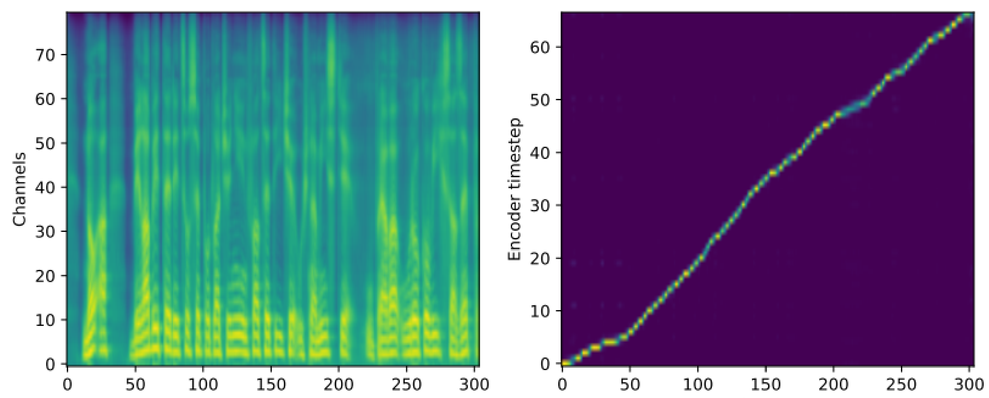
- Nahrávky z rozhlasu s textovým a fonetickým přepisem
- Data pro 4 osoby, zhruba 20 hodin pro každou z nich

Vybrané architektury

- Porovnávání pomocí metriky Mean Opinion Score
- Architektury většinou rozděleny na dvě části:
  1. Převádí vstupní text na mel spektogram
     - Vybrán DeepVoice3 a Tacotron 2 model
  2. Převádí daný spektogram zpět do časové oblasti
     - Vybrán WaveGlow model

Provedené experimenty a jejich výsledky

- Optimalizace hyperparametrů sítě
  - Změna rozměrů architektury má minimální vliv na kvalitu výstupní řeči
- Výrazné zlepšení po pročištění trénovacích dat
- Využití předtrénovaného NVIDIA modelu pro angličtinu
  - Dotrénování na českých datech



Obrázek 1: Výstupní mel spektogram a attention zarovnání Tacotron 2 modelu.

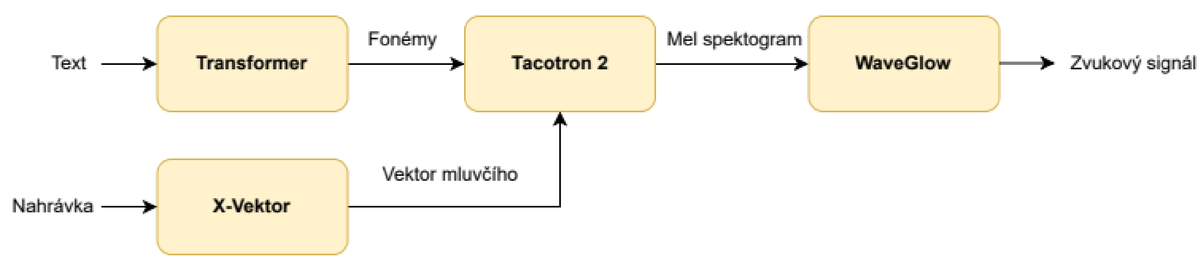
PROVEDENÉ PRÁCE NAD RÁMEC ZADÁNÍ

Fonetická transkripce

- Po použití fonetické transkripce se dosáhlo lepší výslovnosti modelu.
- K transkripci použita Transformer architektura.

Syntéza řeči pro více mluvčí

- Realizována rozšířením Tacotron 2 modelu o vektory mluvčího (tzv. X-Vektory).
- Cílem byla možnost změny výstupního hlasu v závislosti na přivedeném vektoru.



Obrázek 2: Struktura výsledného systému diplomové práce.

VÝSLEDKY

Poslechové testy

- Vytvořeno prostředí v demonstrační aplikaci
- Pro hodnocení vybrán mužský hlas Tacotron 2 a WaveGlow modelu
- Hodnocení se zúčastnilo 56 osob
- Celkem ohodnoceno 1 060 nahrávek od každého systému se stejným obsahem

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

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

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