Kvalita neurální syntézy řeči v závislosti na množství trénovacích dat

Lukáš Vladař[1]

1 Úvod

Počítačová syntéza řeči umožňuje z nahrávek řeči rekonstruovat hlas řečníka a následně jej použít pro „čtení“ libovolného textu.

Velmi dobrých výsledků dosahuje zejména syntéza založená na neuronových sítích. Syntetizér pracující tímto způsobem je, zjednodušeně řečeno, funkce závislá na velkém množství parametrů. Pro správné fungování syntetizéru je pak nutné najít optimální hodnoty těchto parametrů. Máme-li k dispozici nahrávky hlasu určitého člověka, můžeme pomocí metod strojového učení nalézt parametry, při jejichž použití syntetizér generuje řeč co možná nejpodobnější hlasu tohoto řečníka (tj. tzv. trénování neuronového modelu).

Může nastat situace, kdy máme k dispozici pouze malé množství nahrávek hlasu daného člověka, a přesto bychom chtěli jeho hlas rekonstruovat. Nabízí se tedy otázka, do jaké míry kvalita syntézy závisí na množství trénovacích dat. Bylo navrženo několik experimentů, které by měly (nejen) tuto otázku zodpovědět.

Pro experimenty byla vybrána architektura VITS, kterou představil Kim et al. (2021). VITS poskytuje velmi dobrou kvalitu syntetické řeči, navíc nefunguje zcela deterministicky, což zvyšuje přirozenost řeči, neboť ani člověk nevysloví stejnou větu pokaždé zcela stejně.

2 Vliv množství trénovacích dat na kvalitu syntézy

Syntetizér byl natrénován s využitím nahrávek hlasu laického řečníka, přičemž k trénování bylo použito nejprve 90 minut, následně 45 minut, a nakonec jen 22,5 minut řeči. Cílem experimentu bylo zjistit, jaký vliv má množství trénovacích dat na kvalitu syntézy.

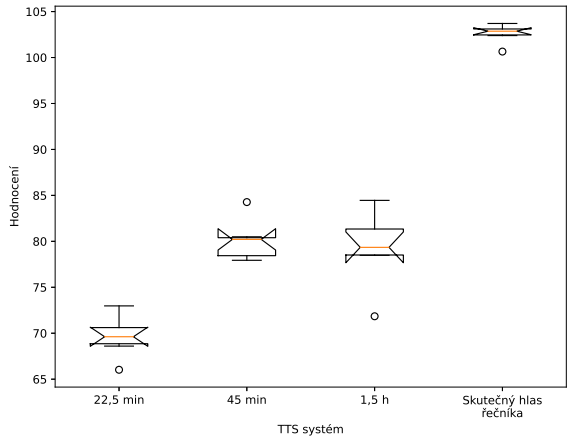
Syntetizéry natrénované z různého množství trénovacích dat byly ohodnoceny 7 respondenty v poslechovém testu typu MUSHRA. Během testu bylo každému respondentovi přehráno 20 promluv, přičemž každá promluva byla vyslovena všemi třemi syntetizéry, a navíc také skutečným řečníkem. Úkolem posluchače bylo ohodnotit každou z nahrávek číslem od 0 do 100 (vyšší hodnocení znamená lepší kvalitu řeči). Na základě těchto dílčích hodnocení bylo spočteno průměrné hodnocení každého syntetizéru daným respondentem, které bylo dále normalizováno, neboť se ukázalo, že někteří respondenti mají tendenci používat jen malý rozsah hodnocení, zatímco jiní využívají celou škálu od 0 do 100.

Chápeme-li hodnocení syntetizéru jako náhodnou veličinu, máme k dispozici 7 realizací této náhodné veličiny, na základě kterých lze odhadnout medián. Na Obrázku 1 jsou vykresleny výsledky poslechového testu. Výřezy v každém ze čtyř boxplotů odpovídají konfidenčnímu intervalu, v němž s pravděpodovností 95 % leží medián hodnocení daného syntetizéru.

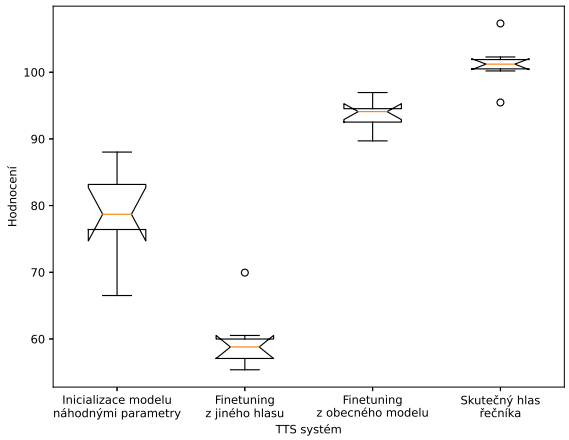
Dle očekávání se ukázalo, že jednoznačně nejlépe zní řeč skutečného řečníka. Mezi syn-

[1] student navazujícího studijního programu Aplikované vědy a informatika, obor Kybernetika a řídicí technika, specializace Umělá inteligence a biokybernetika, e-mail: vladarl@students.zcu.cz

tetizéry natrénovanými s využitím 45 min a 1,5 h řeči se neukázal statisticky významný rozdíl, avšak systém natrénovaný s využitím pouhých 22,5 min řeči dopadl výrazně hůře.



Obrázek 1: Srovnání kvality syntetizéru VITS při použití různého množství trénovacích dat



Obrázek 2: Srovnání kvality syntetizéru VITS natrénovaného z 22,5 minut řeči při různém způsobu inicializace modelu

3 Vliv použití předtrénovaných modelů na kvalitu syntézy

V předchozím experimentu se ukázalo, že při malém množství trénovacích dat dosahuje výsledná syntéza horší kvality. Je však možné, že by kvalitu syntetické řeči bylo možné zlepšit použitím předtrénovaného modelu.

Trénování syntetizéru je běžně zahajováno náhodnou inicializací parametrů, namísto toho však můžeme použít parametry modelu natrénovaného pro jiného řečníka (tzv. transfer learning a finetuning). Další možností je použít parametry obecného modelu, který byl natrénován z nahrávek různých řečníků, a měl by tedy vystihovat základní charakteristiky lidské řeči.

Trénování bylo inicializováno všemi výše uvedenými způsoby a následně byl syntetizér dotrénován s využitím 22,5 minut řeči daného řečníka. Kvalita syntézy při použití těchto přístupů je srovnána na Obrázku 2.

Můžeme pozorovat, že při finetuningu z hlasu jiného řečníka je kvalita syntézy horší než při náhodné inicializaci syntetizéru, což může být způsobeno mj. tím, že jsou hlasy daných řečníků hodně odlišné. Použitím obecného modelu se však kvalita syntézy zlepšila.

4 Závěr

Ukázali jsme, že při nedostatku trénovacích dat dosahuje syntéza řeči horší kvality. Kvalitu řeči lze nicméně zlepšit použitím předtrénovaného obecného modelu. Námětem pro další práci by mohlo být ověření, že vyvozené závěry platí i pro syntézu hlasu jiných řečníků, případně i pro jiné architektury nežli VITS. Při experimentech byl využíván obecný model natrénovaný z 6 hlasů, bylo by však zajímavé otestovat, zda se kvalita syntézy ještě více nezlepší, použijeme-li obecnější model natrénovaný s využitím nahrávek většího počtu řečníků.

Literatura

Kim, J., Kong. J., Son, J. (2021) Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *International Conference on Machine Learning*