FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA
KYBERNETIKY

Diplomová práce

Analýza velikosti dat pro neurální syntézu řeči

Lukáš Vladař

PLZEŇ

2023

FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA
KYBERNETIKY

Diplomová práce

Analýza velikosti dat pro neurální syntézu řeči

Bc. Lukáš Vladař

Vedoucí práce
doc. Ing. Jindřich Matoušek, Ph.D.

PLZEŇ 2023

Citace v seznamu literatury:
VLADAŘ, Lukáš. *Analýza velikosti dat pro neurální syntézu řeči*. Plzeň, 2023. Diplomová práce. Západočeská univerzita v Plzni, Fakulta aplikovaných věd, Katedra kybernetiky. Vedoucí práce doc. Ing. Jindřich Matoušek, Ph.D.

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2022/2023

ZADÁNÍ DIPLOMOVÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Bc. Lukáš VLADAŘ**
Osobní číslo:	**A21N0127P**
Studijní program:	**N3918 Aplikované vědy a informatika**
Studijní obor:	**Kybernetika a řídicí technika**
Téma práce:	**Analýza velikosti dat pro neurální syntézu řeči**
Zadávající katedra:	**Katedra kybernetiky**

Zásady pro vypracování

Seznamte se s problematikou syntézy řeči z textu (TTS), zaměřte se na metody neurální syntézy řeči.
Podrobně se seznamte s moderními frameworky pro trénování neurálních modelů syntézy řeči, např. s frameworkem Coqui-ai/TTS.
Navrhněte experimenty pro trénování neurálních modelů v závislosti na počtu a velikosti zdrojových dat (řečových nahrávek daného hlasu) a na zvolené strategii trénování (např. „single-speaker" trénování vs. využití předtrénovaných „multi-speaker", popř. „multi-language multi-speaker" modelů).
Navržené experimenty vyhodnoťte z hlediska kvality výsledné syntetické řeči a v závislosti na velikosti zdrojových dat.

Rozsah diplomové práce: **40-50 stránek A4**
Rozsah grafických prací:
Forma zpracování diplomové práce: **tištěná**

Seznam doporučené literatury:

Tan, X., Qin, T., Soong, F., Liu, T-Y. (2021). *A Survey on Neural Speech Synthesis*. Dostupné z https://arxiv.org/abs/2106.15561
Kim, J., Kong, J., Son, J. (2021). *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech*. Dostupné z https://arxiv.org/abs/2106.06103

Vedoucí diplomové práce: **Doc. Ing. Jindřich Matoušek, Ph.D.**
Katedra kybernetiky

Datum zadání diplomové práce: **1. října 2022**
Termín odevzdání diplomové práce: **22. května 2023**

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
L.S.

Doc. Ing. Miloš Železný, Ph.D.
děkan

Prof. Ing. Josef Psutka, CSc.
vedoucí katedry

V Plzni dne 1. října 2022

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného akademického titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Západočeská univerzita v Plzni má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Plzni dne 23. srpna 2023

. . . . . .

Lukáš Vladař

V textu jsou použity názvy produktů, technologií, služeb, aplikací, společností apod., které mohou být ochrannými známkami nebo registrovanými ochrannými známkami příslušných vlastníků.

Abstrakt

Hlavním cílem této práce je popsat, jaký vliv má množství použitých trénovacích dat na kvalitu syntetické řeči generované natrénovaným neurálním syntetizérem a jak lze kvalitu výsledné řeči ovlivnit použitím předtrénovaných neurálních modelů. Teoretická část práce popisuje základní přístupy použitelné pro umělé generování řeči, zvláštní pozornost je věnována především moderním metodám neurální syntézy řeči. Zmíněny jsou také možnosti hodnocení syntetické řeči. V praktické části jsou pak popsány experimenty, které byly provedeny s neurálním modelem VITS. V rámci těchto experimentů byly s využitím poslechových testů a objektivní míry MCD porovnávány modely, které se lišily jednak množstvím dat použitých k jejich natrénování, ale také způsobem inicializace parametrů při trénování.

Klíčová slova

syntéza řeči • TTS • VITS • trénovací data • transfer learning • poslechové testy

Abstract

The aim of this thesis is to describe the impact of the amount of used training data on the quality of the speech generated by a neural synthesizer. Another goal is to answer how the use of pretrained neural models can affect the quality of the final speech. The theoretical part of the thesis covers basic approaches applicable to artificial speech production, the main attention is given especially to modern methods of neural speech synthesis. Furthermore, some options of the synthetic speech evaluation are mentioned. The practical part describes experiments performed on the VITS neural model. During these experiments models trained using different amounts of data and different methods of parameter initialization were compared using listening tests and the MCD objective measure.

Keywords

Speech synthesis • TTS • VITS • Training data • Transfer learning • Listening tests

Poděkování

Chtěl bych tímto poděkovat vedoucímu práce, panu doc. Ing. Jindřichu Matouškovi, Ph.D., za odborné vedení, vstřícný přístup a množství věnovaného času. Děkuji také panu Ing. Zdeňku Hanzlíčkovi, Ph.D., za cenné rady ohledně přípravy a vyhodnocování poslechových testů.

Velké díky patří rovněž všem, kteří hodnotili kvalitu syntetické řeči v rámci poslechových testů, bez nich by totiž tato práce nemohla vzniknout. Seznam všech respondentů je v abecedním pořadí uveden níže:

Natálie Bauerová
Jakub Bořík
Klára Boříková
Ladislav Bořík
MUDr. Věra Boříková
Darja Brunová
Ing. Hana Grundmannová
Michaela Jendelová
Miriam Jendelová, DiS.
Ing. Barbora Jůzová
Ing. Marie Kunešová, Ph.D.
Bc. Michaela Kůsová
Ing. Markéta Řezáčková
Adéla Vladařová
Dominik Vladař
Jana Vladařová
Ing. Karel Vladař
Roman Vladař

Dále bych chtěl velice poděkovat svým rodičům za podporu během celé doby studia na vysoké škole.

Výpočetní zdroje byly poskytnuty projektem e-INFRA CZ (ID:90140), který je podporován Ministerstvem školství, mládeže a tělovýchovy.

Obsah

Úvod 1

Řeč je jednou z prvních dovedností, kterým se člověk hned po narození učí. Snad právě díky tomu mohou lidé ovládat mluvenou řeč s naprostou lehkostí a přirozeností. Mluvení nám umožňuje předávat si poměrně rychle a bez nadměrného soustředění informace a myšlenky.

V posledních desetiletích dochází k enormnímu rozvoji informačních technologií, což má za následek skutečnost, že lidstvo stále více času tráví „komunikací" s počítačem. Tato komunikace má obvykle podobu zadávání informací pomocí klávesnice, myši, dotykové obrazovky či jiného vstupního zařízení a čtení výstupních informací z displeje zařízení.

Uvedený způsob komunikace ovšem není pro člověka přirozený a vyžaduje vyšší soustředění. V některých úlohách by bylo vhodnější výměnu informací mezi počítačem a uživatelem realizovat formou přirozeného jazyka. Touto myšlenkou se zabývá výzkum tzv. hlasových dialogových systémů, které přijímají informace od uživatele pomocí automatického rozpoznávání řeči, tyto informace zpracují, a výstup uživateli prezentují počítačově syntetizovanou řečí.

V některých situacích je vhodnější prezentovat uživateli výstupní informace tradičním způsobem, tedy graficky s využitím obrazovky zařízení, existují ale úlohy, ve kterých lze komunikaci člověka se strojem plně nahradit hlasovým dialogovým systémem. Tento přístup mj. uživateli umožňuje, aby se během dialogu věnoval jiné činnosti, což je v některých situacích velmi důležité (např. při řízení motorového vozidla hlasová navigace neruší řidičovu pozornost potřebnou pro sledování silničního provozu). Hlasové dialogové systémy rovněž umožňují komunikaci s moderními technologiemi nevidomým, pro které je čtení informací z obrazovky nemožné.

Zaměříme-li se pouze na systémy pro generování umělé řeči, nalezneme množství dalších aplikací, ve kterých lze tyto metody uplatnit. Patrně největší přínos má syntéza řeči pro lidi, kteří kvůli zdravotním problémům nemohou používat vlastní hlas (např. pacienti s rakovinou hrtanu [1]). Těm počítačová syntéza řeči umožňuje komunikovat se svými blízkými, a dokonce, jsou-li k dispozici nahrávky pacientovy řeči pořízené před ztrátou hlasu, lze daný hlas rekonstruovat tak, že syntéza mluví hlasem, na který jsou pacientovi blízcí zvyklí.

Moderní metody počítačové syntézy řeči dosahují velmi dobré úrovně srozumitelnosti a přirozenosti, nicméně vyžadují poměrně velké množství trénovacích dat. Trénovacími daty rozumíme zpravidla nahrávky řeči a přepis vět, které řečník v nahrávkách říká.

Obstarání takovýchto dat je mnohdy časově či finančně náročné. Existují situace, kdy máme pro tvorbu syntetizéru k dispozici pouze omezený počet trénovacích dat a obstarání dalších dat již není možné. Uvažujme např. již zmíněné pacienty, kteří přišli o hlas – řečník již není schopen nahrát další trénovací data, je tedy nutné použít pouze nahrávky, které byly vytvořeny před ztrátou hlasu (a těch nemusí být mnoho).

Vyvstávají tedy mj. následující otázky: Jak ovlivňuje počet trénovacích dat kvalitu výsledné syntetické řeči? Pokud je k dispozici málo trénovacích dat, nezlepší kvalitu syntézy použití předtrénovaných neuronových modelů? Právě zodpovězení těchto otázek je hlavním cílem této práce.

Metody generování umělé řeči

2

V této kapitole bude popsán základní princip různých metod použitelných pro generování syntetické řeči. Budou zde uvedeny historické pokusy o vytvoření mluvicích strojů, ale též metody počítačové syntézy, které byly běžně používány ještě před několika lety. Moderním přístupům založeným na neuronových modelech je věnována kapitola 3.

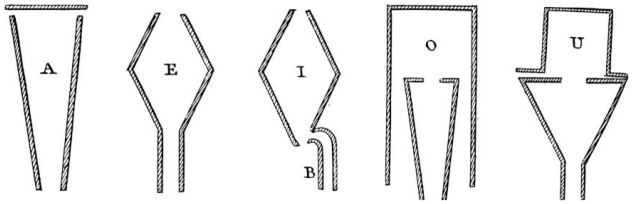
2.1 Mechanické syntetizéry

První pokusy o umělou tvorbu řeči sahají až do 18. století. Poznatky získané ohledně anatomie řečového traktu odstartovaly diskuze o tom, jakým způsobem vzniká lidská řeč a jak by bylo možno tento proces uměle napodobit [2].

2.1.1 Kratzensteinovy „samohláskové varhany“

V roce 1780 petrohradská akademie věd vypsala soutěž, jejíž účastníci měli za úkol popsat charakteristiky pěti samohlásek ([a], [e], [i], [o], [u]) a navrhnout zařízení, které bude vyslovování těchto samohlásek napodobovat pomocí varhanních píšťal. Vítězem soutěže se stal původem německý profesor Christian Gottlieb Kratzenstein, který představil jakési varhany, které zvuk samohlásek imitovaly pomocí pěti píšťal [3].

Samotné zařízení ani jeho nákres se bohužel nedochovaly, víme však, že se skládalo z pěti píšťal ovládaných malou klaviaturou a z měchu, který „hráč“ ovládal nohou. Nástroj obsahoval čtyři tzv. jazykové píšťaly, které vytvářejí zvuk kmitáním kovového jazýčku, a jednu tzv. retnou píšťalu, která funguje na podobném principu jako zobcová flétna [3]. Píšťaly byly vytvarovány tak, aby jejich zvuk co nejvíce připomínal vyslovování samohlásek [2]. Tvary píšťal určených pro jednotlivé samohlásky jsou znázorněny na obrázku 2.1.



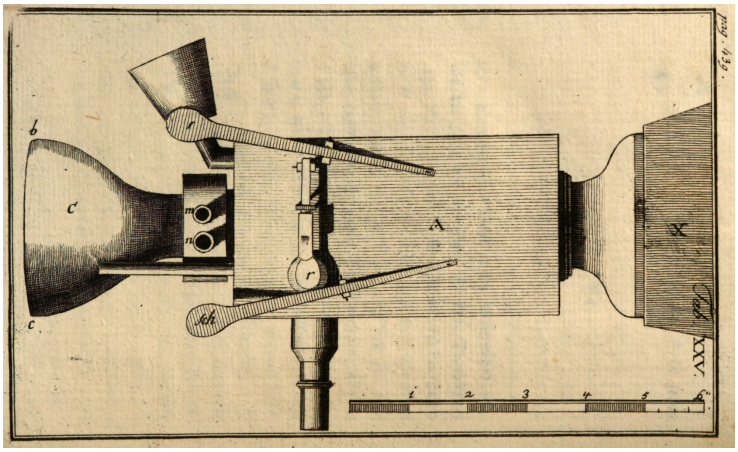
Obrázek 2.1: Tvary píšťal používané Kratzensteinovými samohláskovými varhanami pro vyslovení jednotlivých samohlásek. Převzato z [4]

2.1.2 Kempelenův mluvicí stroj

Dalším známým příkladem mechanického syntetizéru je tzv. Kempelenův mluvicí stroj. Wolfgang von Kempelen byl vynálezce původem z města Preßburg (dnešní Bratislava) [3].

Kempelenovým cílem bylo sestrojit zařízení, které by umožnilo hluchoněmým lidem mluvit. Aby tento stroj dokázal plně nahradit lidskou řeč, bylo třeba, aby dokázal vyslovit jakoukoliv hlásku, ne jen samohlásky, které imitovalo zařízení Ch. G. Kratzensteina. Kempelen se snažil zařízení navrhnout tak, aby jeho konstrukce co nejvíce odpovídala anatomii hlasového traktu člověka [3].

Mluvicí stroj byl manuálně ovládán oběma rukama uživatele. Pravým loktem bylo nutné stlačovat měch, který do celého zařízení vháněl vzduch (analogie s lidskými plícemi), levá dlaň vytvářela artikulaci zakrýváním gumové trubice představující ústa a pravá ruka ovládala uzavíratelné nosní dírky a páky pro vyslovování hlásek [s] a [š] [2, 3]. Kempelenův mluvicí stroj je zobrazen na obrázku 2.2.



Obrázek 2.2: Nákres Kempelenova mluvicího stroje. Převzato z [5]

Článek [2] uvádí, že pomocí Kempelenova mluvicího stroje bylo možné srozumitelně syntetizovat libovolnou větu, avšak řeč zněla velmi nepřirozeně a ovládání zařízení navíc vyžadovalo hodně zkušeností.

2.2 Elektronické syntetizéry

V 1. polovině 20. století došlo k masivnímu rozvoji elektrických zařízení, což se odrazilo mj. také v oblasti generování umělé řeči. Vědci zabývající se touto problematikou se oprostili od myšlenky, že syntetizér musí pracovat na mechanickém principu stejně jako hlasový trakt člověka, namísto toho se pokoušeli fungování hlasového ústrojí imitovat pomocí vhodného elektrického přístroje.

V této práci představíme jeden z těchto elektronických syntetizérů, který nesl označení Voder.

2.2.1 Voder[1]

Voder (Voice Operation Demonstrator) byl elektronický syntetizér řeči, který v roce 1937 představil Homer Dudley. Dudleyovým cílem bylo vytvořit elektrický přístroj, který by svou funkcí napodoboval procesy probíhající v lidském hlasovém traktu.

Obrázek 2.3: Helen Harperová obsluhující elektronický syntetizér Voder[1]

Lidská řeč se skládá ze znělých zvuků, při jejichž vyslovování kmitají hlasivky, čímž vzniká tón, a z neznělých zvuků, na jejichž vytvoření se hlasivky nepodílejí a žádný tón nevzniká. Tento fenomén byl vystižen i ve Voderu – znělé zvuky byly

[1] Informace o systému Voder, stejně jako použitý obrázek, byly čerpány z brožury *What Is Voder*. Dostupné z: `https://www.specialtyansweringservice.net/wp-content/uploads/resources_papers/what-is-the-voder/The-Voder.pdf`

vytvářeny pomocí elektrického oscilátoru, zatímco neznělé vznikaly pomocí náhodného šumu.

Řeč je dále u člověka obohacena o další charakteristiky tím, že se pohybem čelistí, rtů, jazyka apod. mění tvar rezonanční dutiny, což způsobuje úpravu frekvenčního spektra řeči. Voder se tento jev snažil napodobit pomocí deseti paralelně zapojených analogových filtrů, které byly manuálně ovládány a upravovaly charakteristiky elektrického signálu.

Uživatel obsluhující Voder přepínal zápěstím mezi generátorem znělých a neznělých zvuků, pomocí deseti kláves ovládal filtry modifikující frekvenční spektrum signálu, nohou měnil frekvenci kmitání elektrického oscilátoru (a tedy výšku hlasu), a navíc měl k dispozici tři klávesy pro generování tzv. ploziv (hlásek [p], [t], [b] apod.) a afrikátů (např. hlásek [č] nebo [dž]). Fotografii ženy obsluhující Voder můžeme vidět na obrázku 2.3.

Ačkoliv nebyla kvalita syntetické řeči produkované Voderem příliš vysoká, ve své době sklidil tento syntetizér velký úspěch. Nutno však podotknout, že ovládání přístroje bylo velmi složité – jednalo se o činnost podobnou hře na hudební nástroj. Uvádí se, že trvalo téměř rok, než se operátor naučil ovládat přístroj tak, aby vytvořené zvuky zněly srozumitelně.

2.3 Digitální syntetizéry

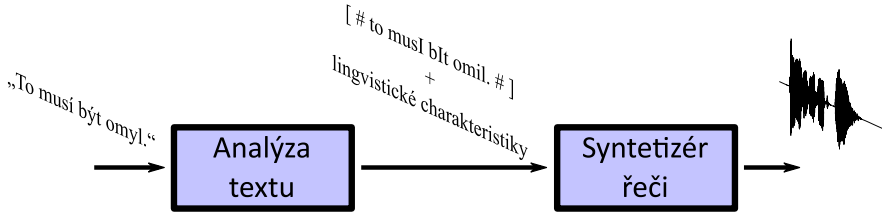
Rozšíření výpočetní techniky umožnilo provádět syntézu řeči s využitím počítače. Číslicový přístup k vytváření řeči je velmi vzdálen skutečnému procesu mluvení, ukazuje se však, že tímto způsobem lze dosáhnout velmi dobrých výsledků, a navíc je syntetizér v podobě počítačového programu na rozdíl od mechanických či elektronických syntetizérů snadno implementovatelný.

Oproti dříve popsaným metodám syntézy řeči není digitální syntetizér pouze přístroj, který při správném ovládání vytváří srozumitelnou řeč. Jedná se o zcela autonomní algoritmus, který dokáže umělou řeč generovat bez přičinění člověka, potřebuje k tomu pouze vstupní text, který má být syntetizován. Z toho důvodu se pro tyto typy syntetizérů často používá termín *systém TTS* pocházející z anglického slovního spojení ***T**ext **t**o **S**peech*, čili systém převádějící text na řeč.

Systém TTS se zpravidla skládá ze dvou komponent. První částí je analyzátor vstupního textu, někdy též nazývaný „frontend systému TTS“. Tento modul extrahuje ze vstupního textu lingvistické příznaky, které obsahují bohaté informace o výslovnosti a prozodii syntetizované promluvy, čímž usnadňují následnou syntézu řeči [6]. Modul analýzy textu provádí konkrétně normalizaci textu, tedy převod nestandardních výrazů do vyslovitelné podoby (tj. rozepsání zkratek, číslic apod.), fonetickou transkripci, což je proces převodu textu do fonetické reprezentace (tj. do posloupnosti fonémů), či predikci prozodie, tj. určování intonace, rytmu

řeči a přízvuků [6]. V některých případech bývají prováděny ještě další podúlohy, např. tzv.*POS tagování*, tj. určování větných členů pro všechna slova vstupního textu [6].

Druhým blokem systému TTS je samotný syntetizér řeči, který z výslovnostní reprezentace syntetizované věty, případně z dalších lingvistických charakteristik, generuje výsledný zvuk. Obecné schéma typického systému TTS je znázorněno na obrázku 2.4.



Obrázek 2.4: Schéma obecného systému TTS. Fonetická reprezentace věty je uvedena ve fonetické abecedě EPA, viz tabulka 5.2

V této podkapitole shrneme základní princip různých přístupů k digitální syntéze řeči.

2.3.1 Artikulační syntéza řeči

Přirozeným přístupem ke generování syntetické řeči je imitovat proces vytváření řeči člověkem, o což se ostatně pokoušeli již návrháři mechanických a elektronických syntetizérů.

Artikulační syntéza se pokouší stejnou myšlenku aplikovat v oblasti číslicové syntézy řeči. Řeč je generována pomocí modelu, který simuluje funkci artikulátorů, jako jsou rty, jazyk apod. [6]. Takovýto artikulační model popisuje např. článek [7].

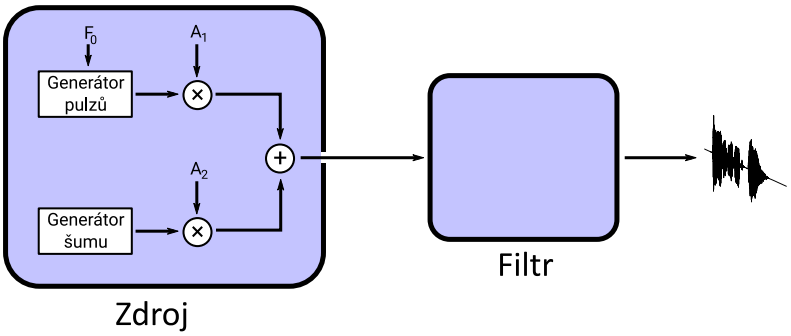
Chceme-li napodobit lidskou řeč, pak bychom intuitivně měli simulací řečového ústrojí obdržet nejlepší výsledky. Je ovšem velmi složité dostatečně přesně popsat funkci artikulátorů, a proto artikulační syntéza zpravidla nedosahuje tak dobrých výsledků jako jiné metody [6].

2.3.2 Formantová syntéza řeči

Pojmem *formant* rozumíme lokální maximum frekvenčního spektra zvuku [8]. Cílem formantové syntézy je napodobit spektrální vlastnosti řeči, zejména pak frekvence a amplitudy formantů [6, 9].

Schéma systému je znázorněno na obrázku 2.5. Řeč je generována pomocí zjednodušeného modelu „zdroj-filtr“ [6]. Filtr je realizován jako sada rezonátorů, které zesilují určité frekvence a zajišťují tak vznik formantů. Navíc je filtr doplněn o anti-

rezonátory, které umožňují generování dalších zvuků, např. frikativ (hlásek [f], [š], [j], [ch] apod.) či ploziv (hlásek [p], [t], [b] apod.) [9].



Obrázek 2.5: Schéma systému pro formantovou syntézu řeči – podle [10]. F_0 označuje frekvenci základního hlasivkového tónu, A_1 a A_2 jsou amplitudy generovaných pulzů, resp. šumu.

Filtr je buzen periodickým signálem pro generování znělých zvuků či šumem v případě vyslovování hlásek neznělých. Parametry modelu jsou nastavovány sadou expertně navržených pravidel [6, 9].

Sestavování těchto pravidel je bohužel velmi náročné a výsledná syntetická řeč zpravidla nezní příliš přirozeně. Přesto však lze pomocí formantové syntézy generovat velmi dobře srozumitelnou řeč [6, 9].

2.3.3 Konkatenační syntéza řeči

Všechny doposud zmíněné metody syntézy (s výjimkou Krantzensteinových varhan) vycházely z myšlenky, že bychom se pro kvalitní syntézu řeči měli co nejvíce přiblížit způsobu, kterým vytváří řeč člověk. Bohužel však popsané syntetizéry nikdy nedosáhly příliš velké přirozenosti, neboť se ukázalo, že lidské řečové ústrojí je velmi složité a není jednoduché ho dostatečně přesně modelovat.

Oproti tomu konkatenační syntéza je prvním z přístupů založených na datech, u nichž návrhář syntetizéru nepotřebuje prakticky žádné apriorní znalosti o fungování lidského hlasového traktu.

Umělá řeč je poskládána z částí dříve nahraných promluv, tzv. řečových jednotek. Řečovými jednotkami mohou být např. slova, slabiky, fonémy[2] či difony[3] [9]. Při generování syntetické promluvy je tedy nejprve analyzováno, ze kterých řečových

[2] Foném je „nejmenší lingvistická jednotka schopná rozlišovat významové jednotky (např. slova). […] Například hlásky `[a]` a `[á]` (resp. `[t]` a `[d]`) jsou pro češtinu dva různé fonémy `/a/` a `/á/` (resp. `/t/` a `/d/`), protože odlišují například slova *hrabě* a *hrábě* (resp. *ten* a *den*)“ [10].

[3] Difon lze definovat jako „segment řeči začínající v polovině předchozí hlásky a končící v polovině následující hlásky…“ [10]. Difon tedy umožňuje plynulejší řetězení řečových jednotek nežli foném, neboť řetězení je prováděno uprostřed hlásky, kde nedochází ke koartikulaci, tj. k přechodu mezi sousedními hláskami.

jednotek se bude výsledná promluva skládat, tyto jednotky jsou následně vyhledány v databázi a jsou pospojovány do souvislé promluvy. Před použitím každé řečové jednotky je však ještě nutné upravit její trvání a frekvenci základního hlasivkového tónu tak, aby tyto vlastnosti odpovídaly požadovaným prozodickým charakteristikám syntetizované promluvy [9].

Čím delší řečové jednotky konkatenační syntéza používá, tím přirozeněji syntetická řeč zní, neboť obsahuje méně bodů, ve kterých muselo dojít ke spojování řečových jednotek. Na druhou stranu však použití delších řečových jednotek vede na větší paměťové nároky, protože existuje větší počet řečových jednotek, které je třeba uchovávat v databázi [9].

2.3.3.1 Konkatenační syntéza řeči s výběrem řečových jednotek

Nevýhodou konceptu konkatenační syntézy je především nutnost každou řečovou jednotku před použitím modifikovat, tyto modifikace totiž zhoršují přirozenost syntetické promluvy [9]. Řešením tohoto problému je vytvoření rozsáhlejší databáze, v níž bude každá řečová jednotka zastoupena větším počtem realizací. Při generování promluvy pak bude možné vybrat realizaci, která co nejlépe odpovídá požadavkům na hledanou řečovou jednotku, a tu již tedy nebude nutné jakkoliv modifikovat. Tento přístup se nazývá syntéza výběrem jednotek neboli anglicky *unit selection*.

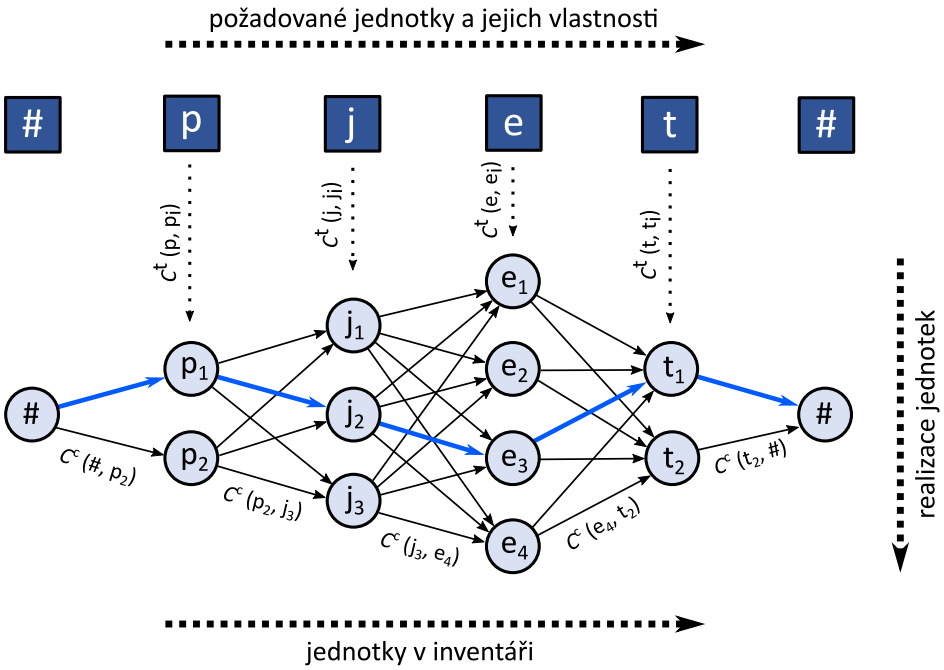
Realizace řečových jednotek, které budou použity pro syntetizování požadované promluvy, jsou vybírány na základě dvou hodnoticích funkcí. Cena cíle $C^{\mathrm{t}}(t_i, u_i)$ odhaduje, do jaké míry se liší realizace řečové jednotky v databázi u_i od požadované řečové jednotky t_i, zatímco cena konkatenace $C^{\mathrm{c}}(u_{i-1}, u_i)$ hodnotí, jak vhodné je zřetězit realizaci řečové jednotky u_i s předchozí řečovou jednotkou u_{i-1} [11].

Sečteme-li ceny cíle všech použitých řečových jednotek a jejich ceny konkatenace, obdržíme celkovou cenu výběru daných jednotek. Vybereme-li pro syntézu řečové jednotky u_1 až u_N, spočteme tuto cenu podle předpisu

$$C\left(t_1^N, u_1^N\right) = \sum_{i=1}^{N} C^{\mathrm{t}}(t_i, u_i) + \sum_{i=2}^{N} C^{\mathrm{c}}(u_{i-1}, u_i). \tag{2.1}$$

Cílem syntézy je pak vybrat takovou posloupnost realizací řečových jednotek u_1^N, která celkovou cenu $C(t_1^N, u_1^N)$ minimalizuje. K tomuto účelu se zpravidla využívá Viterbiův algoritmus [10]. Problém výběru řečových jednotek je ilustrován na obrázku 2.6.

Konkatenační syntéza řeči založená na výběru řečových jednotek dosahuje dobré kvality, její nevýhodou je však velká paměťová náročnost, neboť je třeba v paměti uchovávat rozsáhlou databázi řečových jednotek [9].



Obrázek 2.6: Ilustrace hledání optimální posloupnosti řečových jednotek pro slovo *pět* – podle [10]. Symbol # označuje pauzu. Algoritmus hledá nejvhodnější posloupnost fonémů, tj. cestu v orientovaném grafu, které odpovídá nejmenší cena $C(t_1^N, u_1^N)$.

Uveďme dále, že konkatenační syntéza je první z metod syntézy řeči uvedených v této práci, která nejenže umožňuje „vyslovit" požadovanou promluvu, ale zároveň automaticky napodobuje hlas řečníka, z jehož promluv byly segmentovány řečové jednotky. Tato skutečnost může být v určitých situacích velmi užitečná, má však také své nevýhody. V případě konkatenační syntézy je totiž velmi problematické generovat řeč jiným hlasem nežli tím, z nějž pocházejí řečové jednotky v databázi [10]. Komplikovaná bývá dokonce i pouhá změna stylu řeči.

2.3.4 Statistická parametrická syntéza řeči

Statistická parametrická syntéza řeči je další z metod generování umělé řeči založených na datech. Zatímco však konkatenační syntéza využívá řečová data jako stavební kameny, ze kterých je následně přímo poskládána požadovaná promluva, v případě statistické parametrické syntézy jsou data používána jen pro natrénování modelu, který následně generuje řeč v podobě čistě syntetického signálu.

Systém statistické parametrické syntézy se obvykle skládá ze tří částí. První část obstarává analýzu vstupního textu (viz výše). Druhou komponentou je akustický model, který generuje akustické charakteristiky řeči (např. průběh frekvence základního hlasivkového tónu, frekvenční spektrum řeči či kepstrum). Proces syntézy řeči

je dokončen ve vokodéru, který na základě vygenerovaných akustických parametrů vytváří výsledný zvuk [6].

Úlohu generování akustických příznaků pomocí akustického modelu formalizuje článek [12]. Akustický model na základě vstupního textu W generuje akustické příznaky $\mathbf{O}$, přičemž je závislý na parametrech $\boldsymbol{\lambda}$. Máme-li k dispozici trénovací data obsahující vstupní text W i požadované akustické příznaky $\mathbf{O}$, můžeme parametry modelu odhadnout na základě kritéria maximální věrohodnosti:

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} p(\mathbf{O}|W, \boldsymbol{\lambda}) \tag{2.2}$$

Následně již můžeme generovat akustické příznaky $\hat{\mathbf{o}}$ pro libovolný vstupní text w maximalizováním výstupní pravděpodobnosti

$$\hat{\mathbf{o}} = \arg\max_{\mathbf{o}} p(\mathbf{o}|w, \hat{\boldsymbol{\lambda}}). \tag{2.3}$$

Akustický model bývá v praxi realizován např. sadou skrytých Markovových modelů (HMM) [6, 12].

Statistická parametrická syntéza řeči nevyžaduje k správnému fungování tak velké množství trénovacích dat jako konkatenační syntéza. Další výhodou je, že lze změnou parametrů ovlivňovat charakteristiky výsledné řeči. Syntetická řeč získaná touto metodou však obvykle obsahuje artefakty, jako je šum či bzučení, a je tedy snadno rozeznatelná od přirozené řeči člověka [6].

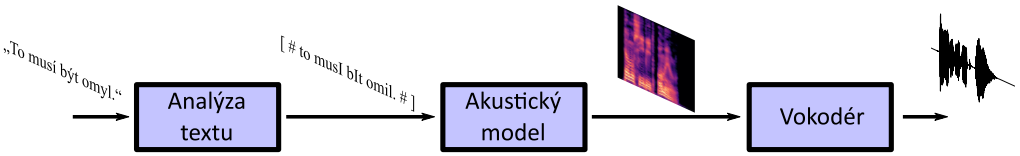
Neurální syntéza řeči 3

V poslední době se pro řešení nejrůznějších úloh v oblasti umělé inteligence stále častěji používají neuronové sítě. Jejich cílem je napodobit fungování lidského mozku modelováním neuronů, tj. nervových buněk, a interakcí mezi nimi.

Neuronová síť je závislá na velkém množství parametrů, jejichž hodnoty je nutno určit metodami strojového učení. Nevýhodou neuronových sítí je, že mohou na určitý vstup zareagovat nesprávně a v takovém případě bývá složité určit, proč k tomuto neobvyklému chování došlo a jak daný problém ošetřit.

I přes určité nevýhody však neuronové sítě nabízejí velký potenciál a ve všemožných úlohách dosahují velmi dobrých výsledků. Proto je také na principu neuronových sítí založena většina moderních systémů pro syntézu řeči.

Systém TTS založený na neuronových sítích má zpravidla stejnou strukturu jako systémy pro statistickou parametrickou syntézu – skládá se z bloku provádějícího analýzu textu, z akustického modelu a z vokodéru [6]. Schéma typického TTS systému je ilustrováno na obrázku 3.1. Poznamenejme však, že třífázová struktura systému nemusí být vždy striktně zachována, některé komponenty systému mohou být sloučeny. Jsou-li všechny tři bloky spojeny do jednoho systému, mluvíme o tzv. *end-to-end* syntéze řeči, které bude věnována jedna z následujících podkapitol.



Obrázek 3.1: Obecné schéma neurálního systému TTS. Fonetická reprezentace věty je uvedena ve fonetické abecedě EPA, viz tabulka 5.2

3.1 Modul analýzy textu

Analyzátor textu používaný v systémech pro statistickou parametrickou syntézu řeči prováděl velké množství operací, v případě neurálních TTS systémů bývá však tento

modul značně zjednodušen, neboť neuronové modely dokáží generovat kvalitní syntetickou řeč i bez bohatých lingvistických informací na vstupu [6]. Analýza textu je v této situaci zpravidla omezena na normalizaci vstupního textu a na fonetickou transkripci. Některé end-to-end systémy dokáží generovat řeč dokonce přímo na základě textu, takže není nutné ani provádění fonetické transkripce [6].

3.2 Akustický model

Akustický model je systém, který na základě lingvistických příznaků získaných analýzou vstupního textu, případně přímo na základě vstupního textu či jeho fonetické reprezentace, generuje akustické charakteristiky, které jsou následně pomocí vokodéru převáděny na výslednou řeč.

Akustickými charakteristikami, z nichž je následně generován zvuk, si lze představit mnohé. Článek [6] uvádí mj. možnost použití melovských kepstrálních koeficientů, frekvence základního hlasivkového tónu nebo informace o znělosti či neznělosti daného řečového úseku. Systémy TTS založené na neuronových sítích však zpravidla používají vysokodimenzionální akustické charakteristiky, jako jsou melovské či lineární spektrogramy [6].

V následujících podkapitolách budou popsány některé běžně používané akustické modely.

3.2.1 Modely založené na RNN – Tacotron 2

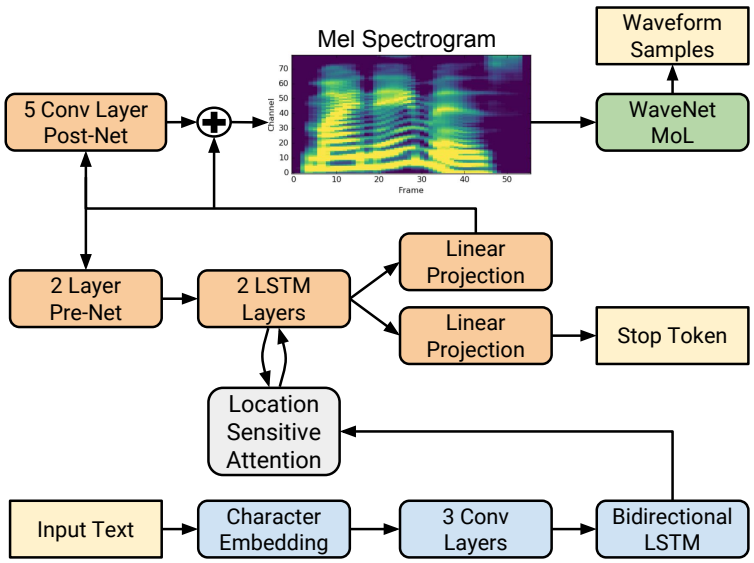
Jednou z možností realizace akustického modelu je využití rekurentních neuronových sítí. Jako příklad uveďme akustický model využívaný systémem Tacotron 2. Akustickými charakteristikami jsou v případě tohoto akustického modelu melovské spektrogramy, které jsou následně převáděny na zvuk vokodérem WaveNet [6]. Schéma celého systému je znázorněno na obrázku 3.2.

Systém Tacotron 2 byl představen v článku [13]. Akustický model se skládá z enkodéru, který převádí vstupní text do vnitřní reprezentace, a z dekodéru, který na základě této reprezentace generuje melovský spektrogram.

Enkodér vytváří *embedding* vstupních znaků, který následně vstupuje do konvoluční sítě a do vrstvy LSTM. Pomocí mechanismu *attention* je k enkodéru připojen dekodér. Dekodér je autoregresní rekurentní síť, což znamená, že je k vygenerování každého vzorku spektrogramu využit předchozí vzorek, který je nejprve zpracován *fully-connected* sítí nazývanou *pre-net*.

3.2.2 Modely založené na CNN – Deep Voice 3

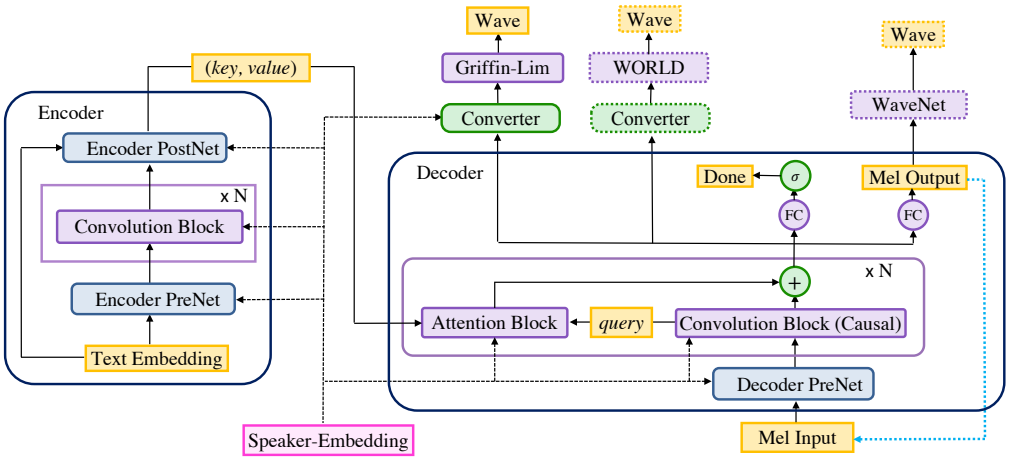
Uveďme dále Deep Voice 3 jako příklad akustického modelu založeného na konvolučních neuronových sítích. Tento model je schopen z textové či fonetické repre-



Obrázek 3.2: Schéma neurálního modelu Tacotron 2. Převzato z [13]

zentace syntetizované promluvy vytvářet různé akustické parametry, např. lineární či melovské spektrogramy, ale též průběh frekvence základního hlasivkového tónu apod. [14]. Díky tomu je Deep Voice 3 kompatibilní s různými typy vokodérů.

Akustický model se skládá z enkodéru a dekodéru, které jsou navrženy na principu konvolučních neuronových sítí. Enkodér převádí vstupní text do vnitřní reprezentace. Autoregresní dekodér tuto reprezentaci postupně dekóduje a vytváří melovský spektrogram řeči, který je následně konvertorem převeden na požadované akustické charakteristiky [14]. Schéma akustického modelu Deep Voice 3 je vykresleno na obrázku. 3.3.



Obrázek 3.3: Schéma akustického modelu Deep Voice 3. Převzato z [14]

Poznamenejme, že na principu konvolučních neuronových sítí jsou založeny i další akustické modely, např. ParaNet či DCTTS [6].

3.2.3 Modely založené na principu transformerů – Transformer TTS, FastSpeech

Transformer TTS je akustický model představený v článku [15]. Vstupem tohoto systému je posloupnost fonémů, na jejíž základě je generován melovský spektrogram [6, 15].

Transformer TTS je inspirován modelem Tacotron 2 (viz podkapitola 3.2.1), snaží se však odstranit jeho nedostatky. Článek [15] upozorňuje na skutečnost, že rekurentní neuronové sítě generují svůj výstup sekvenčně, což modelu Tacotron 2 znemožňuje provádět paralelní trénování a inferenci, proto je tento systém poměrně pomalý. Navíc je pro modely založené na RNN problematické modelování dlouhodobých závislostí mezi vstupní a výstupní posloupností [6].

Transformer TTS využívá princip transformerů pro odstranění zmíněných problémů a ukazuje se, že tím překonává výsledky akustického modelu Tacotron 2 [15]. Na obrázku 3.4 je znázorněno schéma systému Transformer TTS.

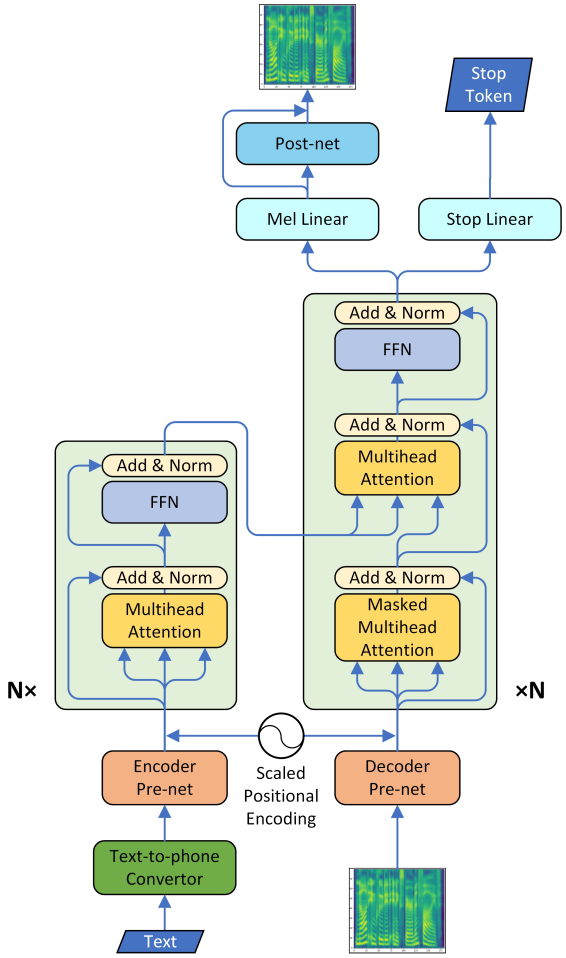
Dalším akustickým modelem fungujícím na principu transformerů je model FastSpeech. Tento model generuje melovský spektrogram řeči, který je ovšem vytvářen paralelně, což oproti autoregresním akustickým modelům poskytuje velmi dobrou rychlost inference [6, 16]. Navíc autoři modelu uvádějí, že FastSpeech téměř eliminuje problémy s opakováním či vynecháváním slov, kterými např. akustický model Tacotron 2 trpí [16].

3.2.4 Další typy akustických modelů

Rekurentní a konvoluční neuronové sítě či transformery nejsou jedinými architekturami použitelnými pro realizaci akustického modelu. V této sekci budou zmíněny některé další principy, na nichž může být akustický model založen.

Diff-TTS je akustický model založený na principu difúze (angl. *diffusion*). Rozumíme-li difúzním procesem sérii transformací, které převedou melovský spektrogram na Gaussovský šum, pak můžeme sérií opačných transformací naopak získat z Gaussovského šumu melovský spektrogram [17]. Na stejné myšlence jsou založeny také akustické modely Grad-TTS či PriorGrad [6].

Akustický model Flow-TTS využívá princip *flow*, který spočívá v modelování složité hustotní funkce dat sérií invertovatelných transformací, které jsou aplikovány na náhodnou proměnnou s jednoduchou hustotní funkcí [18]. Dalšími akustickými modely fungujícími na principu flow jsou např. Flowtron či Glow-TTS [6].



Obrázek 3.4: Schéma akustického modelu Transformer TTS. Převzato z [15]

Některé další akustické modely využívají pro generování akustických příznaků kupř. architekturu GAN (např. Multi-SpectroGAN, TTS-Stylization a GAN exposure) či princip VAE (GMVAE-Tacotron, VAE-TTS apod.) [6].

3.3 Vokodér

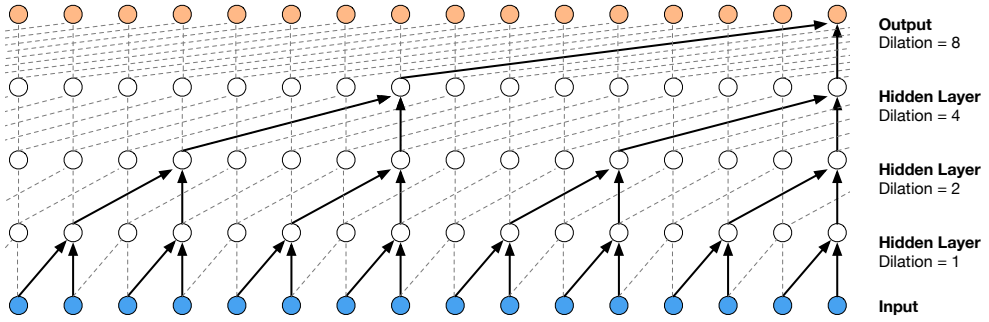
Jak již bylo zmíněno, vokodér je systém, který převádí akustické příznaky vygenerované akustickým modelem do podoby výsledného zvukového signálu. V této sekci budou uvedeny některé přístupy používané k realizaci vokodéru v neurálních systémech TTS.

3.3.1 Autoregresní vokodéry – WaveNet

WaveNet byl původně zamýšlen jako systém generující řeč přímo z lingvistických příznaků (tedy bez použití akustického modelu), lze jej však použít též jako vokodér

zpracovávající lineární či melovské spektrogramy [6].

WaveNet funguje autoregresně, tj. k vygenerování aktuálního zvukového vzorku využívá vzorky předchozí. Predikce je prováděna s využitím konvoluční sítě. Konkrétně je využita dilatovaná neboli *à trous* konvoluce, která umožňuje zpracovávat velký počet předchozích vzorků v kratším čase, než jaký by byl nutný v případě klasické konvoluce [19]. Princip dilatované konvoluce je ilustrován na obrázku 3.5.



Obrázek 3.5: Ilustrace principu dilatované konvoluce. Převzato z [19]. Oproti klasické konvoluci je každá vstupní i spočtená hodnota zpracovávána pouze jednou, což snižuje výpočetní náročnost

Řeč generovaná vokodérem WaveNet dosahuje vysoké kvality, problémem je však autoregresní povaha systému, která způsobuje dlouhou dobu inference [6].

3.3.2 Vokodéry založené na principu flow – WaveGlow

Některé vokodéry využívají ke generování audia princip *flow*, který již byl zmíněn v podkapitole 3.2.4. Mezi tyto systémy se řadí např. vokodér WaveGlow. Zvuk je zde vytvářen na základě melovského spektrogramu [20]. Při generování zvuku je náhodně vybrán vektor z normálního rozdělení s nulovou střední hodnotou a jednotkovou kovarianční maticí a na ten je následně aplikována série transformací, které jej převedou do požadovaného rozdělení [20].

3.3.3 Vokodéry založené na architektuře GAN – MelGAN, HiFi-GAN

Dalším typem vokodérů jsou neuronové modely typu GAN (***G**enerative **A**dversarial **N**etworks*). Vokodéry založené na tomto principu se při trénování skládají ze dvou komponent – z generátoru, který na základě vstupních charakteristik generuje audio, a z diskriminátoru, jehož cílem je rozlišovat skutečnou nahrávku od zvuku vytvořeného generátorem [6]. Při inferenci je pak využíván pouze generátor.

Na architektuře GAN je založen např. vokodér MelGAN, který byl představen v článku [21]. Generátor je v případě tohoto modelu plně konvoluční síť, která generuje zvuk na základě vstupního melovského spektrogramu. MelGAN obsahuje tři dílčí diskriminátory, které mají stejnou strukturu, avšak každý hodnotí vstupní audio na jiné škále. Zatímco jeden z diskriminátorů pracuje s originálním zvukem, vstup zbylých dvou diskriminátorů je podvzorkován, díky čemuž jsou před nimi skryty vysoké frekvence obsažené v původním signálu, takže se tyto diskriminátory musejí naučit klasifikovat vstupní nahrávku pouze na základě nízkých frekvencí [21]. Modifikacemi vokodéru MelGAN vznikl model Multi-band MelGAN, který generuje kvalitnější řeč a zároveň umožňuje stabilnější proces trénování [22].

Další vokodér využívající architekturu GAN se nazývá HiFi-GAN. Stejně jako v případě vokodéru MelGAN i zde vytváří zvuk z melovského spektrogramu generátor založený na konvoluční neuronové síti [23]. Model HiFi-GAN obsahuje dva diskriminátory. Tzv. *Multi-Scale Discriminator* funguje na stejném principu jako diskriminátor používaný v případě modelu MelGAN – skládá se ze tří částí, které zpracovávají vstupní signál v různém měřítku [23]. Navíc však HiFi-GAN představuje tzv. *Multi-Period Discriminator*, který vyhodnocuje periodické jevy v nahrávce. Tento diskriminátor se rovněž skládá z několika částí, přičemž každá část zkoumá periodické jevy o jiné periodě. Vstupní signál je v rámci tohoto diskriminátoru přeskládán do dvourozměrného pole, na nějž jsou následně aplikovány 2D konvoluce [23].

3.3.4 Vokodéry založené na principu difúze

V sekci 3.2.4 byly zmíněny akustické modely fungující na principu difúze. Na stejné myšlence jsou založeny rovněž některé vokodéry. Takovýto vokodér vytváří řeč postupným transformováním náhodného šumu [6]. Mezi vokodéry využívající princip difúze patří např. DiffWave, WaveGrad či PriorGrad [6]. V publikaci [6] se uvádí, že řeč generovaná touto metodou může dosáhnout velmi dobré kvality, problémem však bývá pomalý proces inference.

3.4 Neurální systémy TTS typu end-to-end

Termínem *end-to-end* označujeme systémy TTS, které se neskládají z akustického modelu a vokodéru, ale vytvářejí syntetickou řeč přímo na základě textu či posloupnosti fonémů [6].

Tento přístup redukuje chyby způsobené nekompatibilitou akustického modelu a vokodéru, navíc současné trénování celého systému TTS jako celku může snížit čas trénování [6]. Za zmínku stojí rovněž skutečnost, že v případě dvoufázových systémů TTS návrhář systému určuje akustické charakteristiky, kterými bude řeč vnitřně reprezentována (např. melovské spektrogramy), avšak tato volba nemusí být

zcela optimální. Výhodou end-to-end systémů tedy je, že nejsou omezeny apriorní volbou akustických příznaků.

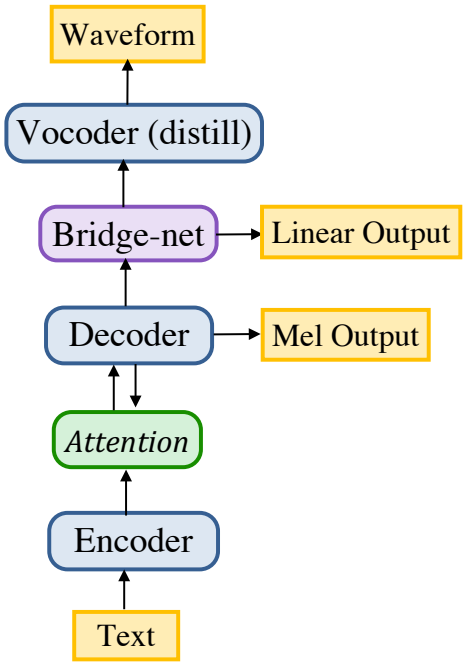
Na druhou stranu však bývá trénování end-to-end TTS systémů komplikovanější. Jedním z hlavních problémů je obrovský rozdíl mezi dimenzionalitou vstupu a výstupu – zatímco vstup systému obsahuje třeba jen řádově desítky znaků či fonémů, výstupní řeč se může skládat i z desítek tisíc vzorků [6].

V následujících podkapitolách budou popsány dva end-to-end systémy TTS: ClariNet a VITS.

3.4.1 ClariNet

Systém ClariNet byl představen v článku [24]. Jedná se o plně konvoluční model založený na akustickém modelu Deep Voice 3, který se skládá z enkodéru a dekodéru (viz podkapitola 3.2.2). Skrytá reprezentace řeči generovaná modelem Deep Voice 3 je zpracována konvoluční sítí nazvanou *Bridge-net* a následně vstupuje do vokodéru WaveNet, který vytváří finální řeč. Celý model je trénován současně jako end-to-end systém [24].

Schéma systému ClariNet je vykresleno na obrázku 3.6.



Obrázek 3.6: Schéma end-to-end systému ClariNet. Převzato z [24]

3.4.2 VITS

VITS neboli *Variational Inference with adversarial learning for end-to-end Text-to-Speech* je moderní end-to-end model představený v článku [25].

Cílem systému je s využitím skryté reprezentace $\mathbf{z}$ generovat výstupní řeč $\mathbf{x}$. Jednou z možností, jak získat „nejpravděpodobnější“ realizaci syntetické řeči, je maximalizovat hustotní funkci generovaných dat $p_\theta(\mathbf{x})$. K tomu lze využít princip variančního enkodéru (VAE) [26], podle něj totiž platí

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x})\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})\, q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})\, q_\phi(\mathbf{z}|\mathbf{x})}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right],
\end{aligned} \tag{3.1}$$

kde $p_\theta(\mathbf{x},\mathbf{z})$ je sdružená hustotní funkce dat $\mathbf{x}$ a jejich skryté reprezentace $\mathbf{z}$, $p_\theta(\mathbf{z}|\mathbf{x})$ je hustotní funkce skryté reprezentace $\mathbf{z}$ podmíněná veličinou $\mathbf{x}$ a $q_\phi(\mathbf{z}|\mathbf{x})$ je aproximace tohoto rozdělení.

Jelikož je druhý člen v odvozeném vztahu vždy nezáporný [26], můžeme napsat

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \\
\log p_\theta(\mathbf{x}) &\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}) - \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})}\right].
\end{aligned} \tag{3.2}$$

Cílem systému TTS však není syntetizovat libovolnou řeč, nýbrž řeč odpovídající požadavkům uživatele, proto se systém VITS snaží maximalizovat hustotu vygenerovaných dat $\mathbf{x}$ podmíněnou vstupními informacemi $\mathbf{c}$. Vztah (3.2) je pak nutné modifikovat takto:

$$\log p_\theta(\mathbf{x}|\mathbf{c}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}) - \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{c})}\right]. \tag{3.3}$$

Abychom tedy dosáhli co možná největší hodnoty logaritmu podmíněné hustoty $\log p_\theta(\mathbf{x}|\mathbf{c})$, můžeme maximalizovat pravou stranu nerovnice (3.3), která se nazývá ELBO (***e**vidence **l**ower **bo**und*). Ztrátová funkce používaná při trénování systému VITS je proto definována jako záporná hodnota ELBO a je možné ji rozdělit na dvě dílčí ztrátové funkce – na chybu rekonstrukce $-\log p_\theta(\mathbf{x}|\mathbf{z})$ a na tzv. *KL divergenci* $\log q_\phi(\mathbf{z}|\mathbf{x}) - \log p_\theta(\mathbf{z}|\mathbf{c})$ [25].

Chyba rekonstrukce (angl. *reconstruction loss*) hodnotí správnost dekódování řeči ze skryté reprezentace $\mathbf{z}$. Systém VITS tuto hodnotu aproximuje L_1 vzdáleností

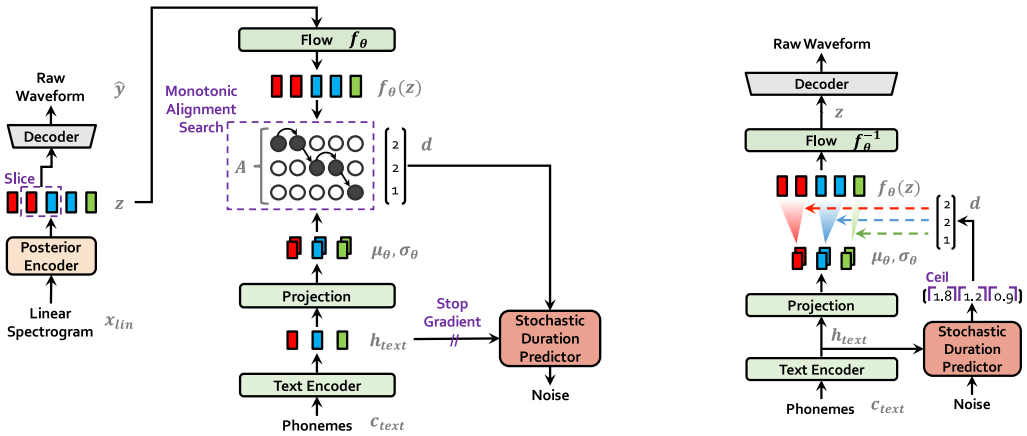
melovských spektrogramů vygenerované řeči a řeči skutečného řečníka:

$$L_{\text{recon}} = \|\mathbf{x}_{\text{mel}} - \hat{\mathbf{x}}_{\text{mel}}\|_1. \tag{3.4}$$

Druhá část ztrátové funkce nazývaná *Kullback-Leiblerova divergence* či zkráceně *KL-divergence* hodnotí kódování vstupních informací do skryté reprezentace. Za vstupní informace je považována posloupnost fonémů určených k syntéze, ale dále také zarovnání v podobě matice **A**, která obsahuje informace o trvání jednotlivých fonémů. Tato matice je odhadována pomocí algoritmu *Monotonic Alignment Search* [25].

Hlavními komponentami modelu VITS je apriorní enkodér a dekodér. Apriorní enkodér, který je založen na architektuře transformerů a na principu *normalizing flow*, modeluje hustotní funkci $p_\theta(\mathbf{z}|\mathbf{c})$, tj. převádí vstup systému do latentní reprezentace. Dekodér, realizovaný generátorem HiFi-GAN V1, tuto reprezentaci převádí do zvukové podoby [25].

Při trénování systému je dále využíván aposteriorní enkodér, který reprezentuje hustotní funkci $q_\Phi(\mathbf{z}|\mathbf{x}_{\text{lin}})$, generuje tedy skrytou reprezentaci $\mathbf{z}$ na základě lineárních spektrogramů $\mathbf{x}_{\text{lin}}$ získaných z řečového korpusu [25]. Důležitou součástí systému VITS je též tzv. *stochastic duration predictor*, což je modul, který generuje pravděpodobnostní rozdělení délek jednotlivých fonémů [25]. Schéma celého systému VITS je vykresleno na obrázku 3.7.



Obrázek 3.7: Schéma end-to-end systému VITS používané při trénování (vlevo) a při inferenci (vpravo). Převzato z [25]

End-to-end model VITS umožňuje generovat velmi kvalitní syntetickou řeč. Velkou výhodou tohoto systému je rovněž fakt, že díky své stochastické povaze vygeneruje při stejném vstupu pokaždé jiný řečový signál – jednotlivé realizace stejné promluvy se liší trváním i intonací [25]. Tato skutečnost zvyšuje přirozenost syntetické řeči, neboť ani skutečný řečník nevysloví stejnou větu vždy zcela stejně.

Hodnocení syntetické řeči

4

Uvažujme situaci, kdy máme k dispozici několik systémů pro syntézu řeči a chceme rozhodnout, který z nich generuje umělou řeč nejlépe. Zodpovězení této otázky není zdaleka jednoduché.

V mnoha úlohách, pro jejichž řešení je používána umělá inteligence, existuje jednoznačný správný výsledek. Pokud tento správný výsledek známe např. od experta v daném oboru, můžeme jej porovnat s výsledkem, který vygeneroval systém umělé inteligence, a na základě tohoto srovnání můžeme fungování systému ohodnotit. Tímto způsobem lze poměrně jednoduše ověřit správnost klasifikace obrázků, rozpoznávání řeči apod.

V úloze syntézy řeči je však situace komplikovanější, neboť pro daný vstupní text neexistuje pouze jediná správná řečová realizace. Dokonce i stejný řečník může při vyslovování téže promluvy použít pokaždé odlišnou intonaci, hlasitost, rychlost řeči apod. Ze skutečnosti, že se syntetizovaná promluva liší od nahrávky lidského řečníka, kterou máme k dispozici, tedy zdaleka nelze usuzovat, že je syntetická řeč nekvalitní.

Hodnocení syntetické řeči je proto poměrně složitá úloha. Dosud nebyl vyvinut systém, který by spolehlivě dokázal kvalitu syntézy určit, proto je umělá řeč zpravidla posuzována lidskými hodnotiteli v rámci tzv. poslechových testů. Přesto však existují objektivní míry hodnotící syntézu na základě porovnání s nahrávkami pocházejícími od skutečného řečníka. Oba přístupy budou stručně popsány v následujících podkapitolách.

4.1 Subjektivní hodnocení řeči – poslechové testy

Zřejmě nejlépe dokáží kvalitu syntetické řeči posoudit lidé, protože jsou řečí obklopeni celý život a znají tak dokonale její charakteristiky. Mnohdy stačí v řečové nahrávce provést byť i nepatrnou změnu a lidský posluchač je schopen nepřirozený artefakt v řeči rozeznat.

Z tohoto důvodu umělou řeč velmi často hodnotí právě lidé v rámci tzv. poslechových testů. Takovéto hodnocení je do značné míry subjektivní, ohodnotí-li však řeč dostatečný počet respondentů, lze z výsledků testu vyvodit objektivní závěr.

Existuje mnoho způsobů, jak poslechový test realizovat. V knize [10] jsou poslechové testy rozděleny na *testy srozumitelnosti* a *testy přirozenosti*.

Cílem testů srozumitelnosti je ohodnotit, do jaké míry posluchači syntetické řeči rozumějí. Mezi takové testy patří mj. testy DRT (***D**iagnostic **R**hyme **T**est*) či MRT (***M**odified **R**hyme **T**est*), při nichž je respondentům předloženo několik slov lišících se pouze jednou hláskou a jejich úkolem je rozhodnout, které z těchto slov bylo systémem TTS vysloveno [10].

Moderní systémy TTS již obvykle netrpí problémy se srozumitelností, proto je větší pozornost věnována testům přirozenosti, které hodnotí celkovou kvalitu syntetizované řeči. Takovým poslechovým testem je např. test MOS (***M**ean **O**pinion **S**core*), při němž posluchači hodnotí kvalitu řeči známkou v rozsahu $1 - 5$. Průměrováním známky napříč respondenty lze získat celkové hodnocení systému [10].

Často používaný je též tzv. párový test, při němž jsou posluchači vždy přehrány dvě realizace stejné promluvy a jeho úkolem je vybrat, která z realizací zní přirozeněji [10].

Dalším testem přirozenosti je test MUSHRA (***Mu**lti **S**timulus Test with **H**idden **R**eference and **A**nchor*) založený na doporučení Mezinárodní telekomunikační unie[1]. Tento typ testu slouží k porovnání syntetické řeči generované několika systémy TTS. Posluchači je vždy předloženo několik realizací stejné promluvy a jeho úkolem je každou z nich ohodnotit číslem v intervalu $0 - 100$. Zároveň je posluchači zpřístupněna referenční nahrávka pořízená skutečným řečníkem, která udává maximální dosažitelnou kvalitu dané promluvy.

Poslechové testy představují při dostatečném počtu respondentů velice přesnou a spolehlivou metodu hodnocení syntetické řeči. Jejich nevýhodou však je, že vyžadují zaměstnání značného počtu respondentů a jsou poměrně časově náročné. Objevují se tedy snahy nahradit lidské hodnotitele objektivním hodnocením řeči, které lze spočítat automaticky. Této problematice bude věnována následující podkapitola.

4.2 Objektivní hodnocení řeči – MCD

Jednou z metod objektivního hodnocení řeči je hodnocení na základě vzdálenosti MCD (***M**el **C**epstral **D**istortion*), jejíž výpočet je popsán v článku [27].

[1] International Telecommunication Union: *Method for the subjective assessment of intermediate quality level of audio systems*. Recommendation ITU-R BS.1534-3 (10/2015). Dostupné také z: `https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-3-201510-I!!PDF-E.pdf`

Cílem této metody je porovnat syntetizovanou promluvu s referenční promluvou nahranou řečníkem. Obě porovnávané nahrávky jsou nejprve reprezentovány posloupností melovských kepstrálních koeficientů $\mathbf{v}^{\text{targ}}$, resp. $\mathbf{v}^{\text{ref}}$. Vzdálenost MCD daných nahrávek je potom definována jako

$$\text{MCD}(\mathbf{v}^{\text{targ}}, \mathbf{v}^{\text{ref}}) = \frac{\alpha}{T} \sum_{t=0}^{T-1} \sqrt{\sum_{d=s}^{D-1} \left(v_d^{\text{targ}}(t) - v_d^{\text{ref}}(t) \right)^2}, \tag{4.1}$$

kde $v_d^{\text{targ}}(t)$, resp. $v_d^{\text{ref}}(t)$ představuje d-tý prvek t-tého vektoru melovských kepstrálních koeficientů spočteného pro syntetizovanou, resp. referenční nahrávku, T je délka kratší z posloupností vektorů, tj. $T = \min(|\mathbf{v}^{\text{targ}}|, |\mathbf{v}^{\text{ref}}|)$, D je dimenze těchto vektorů a α je konstanta používaná z historických důvodů, pro niž platí $\alpha = \frac{10\sqrt{2}}{\ln 10}$.

Hodnota s určuje, počínaje kterou dimenzí budou melovské kepstrální koeficienty využívány, při volbě $s = 0$ jsou k výpočtu použity všechny koeficienty. Je však známo, že nultý melovský kepstrální koeficient odpovídá energii signálu. Jelikož obvykle nechceme penalizovat rozdíly v hlasitosti promluv, bývá zvykem nultý koeficient z výpočtu vyřadit, čehož dosáhneme volbou $s = 1$ [27].

Uvedený postup výpočtu vzdálenosti MCD předpokládá, že t-tý vektor melovských kepstrálních koeficientů spočtený pro syntetickou řeč odpovídá stejné části promluvy jako t-tý vektor spočtený pro referenční nahrávku. Tento předpoklad však nemusí být obecně splněn, neboť v každé z nahrávek může být promluva vyslovena jiným tempem.

V takovém případě je vhodné namísto výpočtu (4.1) určit vzdálenost MCD pomocí algoritmu dynamického programování (DTW) [27]. Tento algoritmus spočívá v konstruování matice $\mathbf{M}$ o rozměrech $|\mathbf{v}^{\text{targ}}| \times |\mathbf{v}^{\text{ref}}|$, jejíž prvek $M_{i,j}$ spočteme jako

$$M_{i,j} = \frac{\alpha}{T} \sum_{t=0}^{T-1} \sqrt{\sum_{d=s}^{D-1} \left(v_d^{\text{targ}}(i) - v_d^{\text{ref}}(j) \right)^2} + \min\left(M_{i-1,j}, M_{i,j-1}, M_{i-1,j-1} \right). \tag{4.2}$$

Vzdálenost mezi dvěma nahrávkami je poté dána posledním prvkem v matici $\mathbf{M}$, tj.

$$\text{MCD}(\mathbf{v}^{\text{targ}}, \mathbf{v}^{\text{ref}}) = M_{|\mathbf{v}^{\text{targ}}|, |\mathbf{v}^{\text{ref}}|}. \tag{4.3}$$

Kromě vzdálenosti MCD existují i jiné metriky použitelné pro objektivní hodnocení syntetické řeči, uveďme např. metodu PESQ (***P**erceptual **E**valuation of **S**peech **Q**uality*) [28].

Výhodou objektivního hodnocení kvality syntetické řeči je, že může probíhat automaticky bez přičinění člověka. Výsledky těchto metod však nejsou zcela spolehlivé, proto je stále obvykle preferováno subjektivní hodnocení řeči pomocí poslechových testů.

Popis experimentů 5

Moderní metody syntézy řeči založené na neuronových modelech, ale také např. dříve populární konkatenační syntézu či statistickou parametrickou syntézu řeči, lze řadit mezi tzv. *korpusově orientované* přístupy k syntéze řeči. Společnou vlastností těchto metod je, že jsou pro jejich správné fungování třeba anotovaná trénovací data [10].

Trénovacími daty rozumíme nahrávky řeči a jejich textový přepis, např. pro konkatenační syntézu jsou však potřebné také údaje určující hranice řečových jednotek v signálu či prozodické informace o jednotlivých řečových jednotkách.

Pro vytvoření kvalitního systému TTS se běžně používají řečové korpusy obsahující desítky až stovky hodin řeči [10], obstarání takto rozsáhlého korpusu je však značně nákladné.

V některých případech je získání většího množství trénovacích dat dokonce nemožné. Uvažujme např. situaci, kdy chceme syntetizovat řeč hlasem člověka, který kvůli zdravotním problémům o hlas přišel. V takovémto případě jsme odkázáni na nahrávky, které pacient pořídil před ztrátou hlasu, a žádné další již není možné vytvořit. Stejný problém nastává, chceme-li rekonstruovat hlas osobnosti, která již nežije. I v tomto případě si musíme vystačit s nahrávkami hlasu, které máme k dispozici, byť jich nemusí být mnoho, žádné další nahrávky již totiž nezískáme.

Nabízí se tedy otázka, jak se na kvalitě syntetické řeči projeví, využijeme-li k vytvoření systému TTS menší množství trénovacích dat.

Vytváříme-li systém TTS založený na neuronových sítích, můžeme se dále ptát, zda by v případě nedostatku trénovacích dat nebylo možné zlepšit kvalitu syntetické řeči použitím principu *transfer learning*, který popisuje např. článek [29]. Tento přístup lze v případě syntézy řeči aplikovat tak, že natrénujeme systém TTS s využitím nahrávek hlasu jiného řečníka, kterých máme dostatek, neboť předpokládáme, že se v lidské řeči nacházejí charakteristiky nezávislé na řečníkovi. Ve chvíli, kdy již systém TTS „umí mluvit“, jej dotrénujeme pomocí nahrávek cílového řečníka, aby výsledná syntetická řeč mluvila hlasem, který požadujeme (tj. tzv. *fine-tuning*).

Další možností je provést transfer learning s využitím tzv. *multi-speaker modelu*. Tímto pojmem označujeme syntetizér, který byl natrénován na základě nahrávek

většího počtu řečníků a měl by tedy vystihovat základní charakteristiky lidské řeči. Fine-tuningem multi-speaker modelu opět získáme systém syntetizující řeč požadovaným hlasem. Je otázkou, zda má využití multi-speaker modelu nějaké výhody oproti situaci, kdy provádíme fine-tuning z modelu natrénovaného s využitím nahrávek pouze jediného řečníka.

Praktická část této práce bude věnována zodpovězení všech zmíněných otázek. Byla navržena sada experimentů, které by měly odhalit vliv množství trénovacích dat na kvalitu syntetické řeči, stejně jako vliv použití předtrénovaných neurálních modelů.

Návrh a průběh experimentů je podrobně popsán v této kapitole. Podkapitola 5.1 uvádí, jaké modely byly pro potřebu experimentů natrénovány, v podkapitole 5.2 jsou pak uvedeny detaily ohledně trénování těchto modelů. Podkapitola 5.3 je věnována přípravě syntetických nahrávek a v podkapitole 5.4 je popsán postup hodnocení natrénovaných modelů.

5.1 Návrh experimentů

Jedním z cílů prováděných experimentů bylo zjistit, jaký vliv má množství trénovacích dat na kvalitu syntetické řeči.

Je otázkou, jakým způsobem množství trénovacích dat kvantifikovat. Udávat toto množství celkovým časem nahrávek v řečovém korpusu patrně není zcela ideální, neboť tento údaj nezohledňuje rychlost řeči daného řečníka či např. délku ticha na začátcích a koncích nahrávek. Lze tedy polemizovat o tom, zda by nebylo vhodnější reprezentovat velikost řečového korpusu počtem vyslovených fonémů. Je však běžnou praxí vyjadřovat velikost korpusu právě časem řečových nahrávek, proto se tohoto zvyku budeme i v této práci držet.

Pro experimenty byl k dispozici poměrně rozsáhlý řečový korpus složený z nahrávek českého profesionálního řečníka, které byly pořízeny pomocí kvalitní techniky v nahrávacím studiu. Tato trénovací data byla využita k natrénování několika neurálních syntetizérů, přičemž pro trénování každého z nich byla použita pouze část řečového korpusu, aby následně bylo možné porovnat, jak množství trénovacích dat ovlivnilo kvalitu generované řeči.

Bylo vytvořeno celkem šest menších korpusů, které obsahovaly postupně 12 hodin, 6 hodin, 3 hodiny, 1,5 hodiny, 45 minut a 22,5 minut nahrávek řeči. Poznamenejme, že se původní korpus skládal z různých typů promluv (krátké věty, souvětí, tázací věty apod.), bylo tedy nutné dbát na to, aby i menší korpusy obsahovaly všechny tyto typy promluv, a to přibližně ve stejném poměru, v jakém byly zastoupeny v původním korpusu.

Cílem systému TTS není vždy jen napodobit hlas řečníka, který má profesionální projev a dokonalou výslovnost, a navíc tvůrci systému poskytne množství vysoce

kvalitních nahrávek z nahrávacího studia. Mnohdy se setkáváme také s úlohou, kdy je třeba rekonstruovat hlas amatérského řečníka, od nějž navíc máme k dispozici často pouze méně kvalitní nahrávky. I v tomto případě nás může zajímat, jak kvalitu syntetické řeči ovlivní množství trénovacích dat.

Pro zodpovězení této otázky byl použit ještě jeden, tentokrát méně rozsáhlý, korpus obsahující nahrávky neškoleného českého řečníka. Tento korpus byl rovněž rozdělen na menší korpusy, které obsahovaly postupně 1,5 hodiny, 45 minut a 22,5 minut nahrávek řeči.

Dalším cílem experimentální části práce bylo vypozorovat vliv přístupu transfer learning na kvalitu výsledné syntetické řeči. K tomuto účelu byly použity dva předtrénované modely. Jeden z modelů byl předtrénován na nahrávkách hlasu jiného profesionálního řečníka, druhým pak byl multi-speaker model, při jehož trénování byly použity nahrávky celkem šesti hlasů (tří ženských a tří mužských).

Každý z vytvořených řečových korpusů byl tedy použit k natrénování syntetizéru, přičemž byly vždy provedeny tři možnosti inicializace trénovaného modelu. Nejprve byly parametry modelu inicializovány náhodně, poté byly použity parametry modelu natrénovaného na základě nahrávek jiného řečníka, a nakonec byly využity hodnoty parametrů multi-speaker modelu.

Celkem tedy bylo natrénováno 27 modelů (18 modelů pro hlas profesionálního řečníka a 9 pro řečníka amatérského). Následující podkapitola je věnována detailům ohledně trénování zmíněných modelů.

5.2 Trénování neurálních modelů

Pro experimenty byl zvolen neurální model VITS (viz podkapitola 3.4.2), který je v současnosti hojně používán. Trénování bylo provedeno pomocí optimizéru *AdamW* s parametry $\beta_1 = 0{,}8$, $\beta_2 = 0{,}99$ a $\lambda = 0{,}01$. Pro konstantu učení (tzv. *learning rate*) byla zvolena hodnota 0,001, v případech, kdy byl model dotrénováván s využitím předtrénovaného modelu, byl learning rate snížen na hodnotu 0,0005. Konstanta učení používaná pro trénování generátoru i diskriminátoru vokodéru HiFi-GAN měla hodnotu 0,0002. Modely byly trénovány pomocí dávek (angl. *batches*) obsahujících 32 nahrávek.

Doba trénování byla pro každý model určena individuálně – o ukončení trénování bylo rozhodnuto na základě průběhu ztrátové funkce počítané pro evaluační data a především na základě subjektivního dojmu, že se již kvalita syntetizované řeči trénováním nezlepšuje. Obecně však lze říci, že při použití většího množství trénovacích dat probíhalo trénování déle než v případě modelů trénovaných z menších řečových korpusů.

V tabulce 5.1 je uveden přibližný počet provedených kroků trénování[1] v závislosti na velikosti použitého korpusu.

Tabulka 5.1: Přibližný počet kroků použitých pro trénování neurálních modelů v závislosti na velikosti použitého řečového korpusu. V uvedených hodnotách nejsou započítány trénovací kroky, které byly provedeny pro natrénování modelu použitého pro fine-tuning

Množství trénovacích dat	Počet kroků trénování
22,5 min	120 000
45 min	230 000
1,5 h	360 000
3 h	480 000
6 h	550 000
12 h	650 000

Trénování modelů bylo realizováno pomocí frameworku Coqui TTS[2], který implementuje mnoho neurálních modelů pro syntézu řeči z textu, mj. právě také end-to-end model VITS.

Vstupem trénovaných modelů byla fonetická reprezentace syntetizované promluvy. Pro zápis fonetické reprezentace textu bylo nejprve nutné zvolit tzv. fonetickou abecedu, tj. množinu symbolů, které budou k reprezentaci výslovnosti používány.

Fonetických abeced existuje velké množství. Jednou z nejznámějších je mezinárodní fonetická abeceda IPA (***I**nternational **P**honetic **A**lphabet*), kterou lze použít pro zápis výslovnosti v libovolném jazyce [10]. Problémem této abecedy je, že používá znaky těžko reprezentovatelné v počítači, proto byla vytvořena fonetická abeceda SAMPA (***S**peech **A**ssesment **M**ethods **P**honetic **A**lphabet*), která se skládá pouze ze znaků ASCII [10].

Speciálně pro český jazyk byla navržena fonetická abeceda ČFA (***Č**eská **f**onetická **a**beceda*), která používá rovněž znaky ASCII [10]. Každý symbol této abecedy je však obecně reprezentován několika znaky ASCII, takže je nutné ve fonetickém zápisu jednotlivé symboly oddělovat mezerami, což není příliš praktické.

V rámci popisovaných experimentů byl vstupní text trénovaných modelů vyjádřen s využitím znaků fonetické abecedy EPA. Tato abeceda pokrývá všechny fonémy používané v českém jazyce a její výhodou je, že každý symbol reprezentuje pouze jedním znakem ASCII, fonetický zápis je tedy paměťově nenáročný a jednotlivé

[1] Trénovacím krokem rozumíme jednu iteraci trénování, při níž jsou parametry trénovaného modelu aktualizovány na základě podmnožiny trénovacích dat zvané dávka (angl. *batch*). V případě popisovaných experimentů byly používány dávky obsahující 32 trénovacích nahrávek, jeden krok trénování tedy odpovídá aktualizaci parametrů modelu na základě 32 promluv.

[2] Dostupné z: `https://github.com/coqui-ai/TTS`

symboly není nutné oddělovat mezerami. Navíc je zápis pomocí této abecedy dobře čitelný pro člověka.

Fonetická abeceda EPA byla pro účely experimentů mírně zjednodušena – nebyly používány speciální znaky pro diftongy (dvojhlásky), rovněž nebylo odlišováno slabikotvorné „r“ a „l“ apod. Seznam všech symbolů, které byly používány pro fonetickou reprezentaci syntetizovaných promluv, je uveden v tabulce 5.2.

Tabulka 5.2: Seznam symbolů fonetické abecedy EPA používaných pro zápis syntetizovaných promluv

	Foném	Používaný symbol	Příklad použití
Vokály	/a/	a	klavír
	/e/	e	dřevo
	/i/	i	zisk
	/o/	o	emoce
	/u/	u	koruna
	/á/	A	sál
	/é/	E	téma
	/í/	I	cíl
	/ó/	O	gól
	/ú/	U	stůl
Frikativy	/f/	f	fyzika
	/v/	v	vlast
	/s/	s	láska
	/z/	z	význam
	/š/	S	škola
	/ž/	Z	žíně
	/ch/	x	pochod
	/h/	h	hotel
	/l/	l	bolest
	/r/	r	koroze
	/ř/	R	křik
	/j/	j	jeviště

Foném	Používaný symbol	Příklad použití	
/p/	p	pýcha	Plozivy
/b/	b	borovice	
/t/	t	otec	
/d/	d	dobrodružství	
/ť/	T	kotě	
/ď/	D	dívka	
/k/	k	křídlo	
/g/	g	gorila	
/m/	m	romantismus	Nazály
/n/	n	násobek	
/ň/	J	oni	
/c/	c	cíl	Afrikáty
/č/	C	činitel	
Dlouhá pauza	$		
Krátká pauza	#		
Nádech	%		

5.3 Příprava nahrávek pro hodnocení natrénovaných modelů

Jak již bylo uvedeno v podkapitole 5.1, v rámci experimentů bylo natrénováno celkem 27 modelů, které se navzájem lišily množstvím dat použitých pro trénování, ale také způsobem inicializace parametrů. Aby bylo možné tyto modely vzájemně porovnat, bylo třeba pomocí nich vygenerovat testovací syntetické nahrávky.

Pomocí každého modelu bylo syntetizováno 100 promluv, přičemž byly voleny věty, které byly rovněž namluveny skutečným řečníkem, což později umožnilo syntetickou řeč porovnat i s referenčními nahrávkami. Poznamenejme však, že tyto referenční nahrávky byly předem odděleny jako testovací data a nebyly tedy použity k trénování neurálních modelů.

Ukázalo se, že v syntetizovaných nahrávkách občas dochází k redukci fonému /r/, tj. k tzv. ráčkování. Tento jev se vyskytoval víceméně náhodně, a to u všech natrénovaných modelů. Zdálo se, že se tedy jedná spíše o obecný problém spojený s modelem VITS, nežli o chybu natrénovaných modelů.

Při poslechových testech, které byly plánovány pro porovnání natrénovaných modelů, by však tento jev mohl způsobovat problémy. Uvažujme situaci, kdy by posluchači měli porovnat kvalitu syntetické řeči vygenerované několika modely, přičemž v jedné z nahrávek by došlo k ráčkování. Posluchači by pak jistě měli tendenci hodnotit tuto nahrávku negativně, ačkoliv je v podstatě náhoda, že se redukce fonému /r/ projevila právě u tohoto modelu, v případě jiné promluvy může stejný problém postihnout jiný z modelů. Z tohoto důvodu bylo rozhodnuto nahrávky, ve kterých k redukci fonému /r/ dochází, pro hodnocení syntetizérů nepoužít.

Otázkou bylo, jak nahrávky, v nichž k tomuto problému dochází, odhalit, aniž by bylo nutné všechny nahrávky poslouchat. Nabízela se možnost použít k tomuto účelu objektivní míru MCD (viz podkapitola 4.2).

Díky stochastické povaze modelu VITS, který byl při experimentech používán, je i při stejném vstupu systému vygenerován pokaždé jiný řečový signál. Provedeme-li tedy opakovaně syntézu stejného vstupního textu, může se stát, že k ráčkování dojde pouze v některých případech. Vybereme-li ze všech realizací stejné promluvy tu, která minimalizuje vzdálenost MCD od nahrávky skutečného řečníka, pak by se mohlo jednat o nahrávku, ve které k redukci fonému /r/ nedochází.

Bohužel se však ukázalo, že eliminace ráčkování tímto způsobem nefunguje – redukce fonému /r/ na vzdálenost MCD prakticky nemá vliv. Bylo tedy nutné si všechny vygenerované nahrávky poslechnout a o správnosti výslovnosti fonému /r/ rozhodnout osobně. Tento proces byl poměrně časově náročný, neboť bylo třeba posoudit 100 nahrávek pro každý z 27 natrénovaných modelů, přičemž docházelo-li v nahrávce k ráčkování, byla daná promluva syntetizována opakovaně a znovu sluchem kontrolována. Jistě tedy není přehnané odhadovat, že pro vyřazení nahrávek obsahujících ráčkování bylo třeba osobně poslechnout přibližně 5 000 řečových nahrávek.

Dále bylo zjištěno, že je mezi syntetizovanými promluvami a referenčními nahrávkami lidským uchem slyšitelný rozdíl v energii signálu – nahrávky pořízené skutečným řečníkem byly tišší než nahrávky syntetizované. Kvůli tomuto jevu by respondenti při poslechových testech velice snadno odhalili, která z nahrávek je referenční. Cílem poslechových testů však není hodnotit hlasitost řeči, proto byly všechny nahrávky před hodnocením normalizovány.

5.4 Hodnocení natrénovaných modelů

Připravené syntetické nahrávky byly posléze použity k porovnání natrénovaných modelů. Syntetická řeč byla hodnocena dvěma způsoby – objektivně pomocí vzdálenosti MCD a subjektivně v rámci poslechových testů. Oba způsoby hodnocení jsou popsány v následujících podkapitolách.

5.4.1 Hodnocení pomocí vzdálenosti MCD

Vzdálenost MCD je hodnota, která na základě melovských kepstrálních koeficientů odhaduje, do jaké míry se liší syntetizovaná promluva od referenční nahrávky pořízené skutečným řečníkem (viz podkapitola 4.2).

V programovacím jazyce Python byla vytvořena vlastní implementace výpočtu vzdálenosti MCD s využitím dynamického programování (DTW). Hodnota vzdálenosti MCD od referenční nahrávky byla spočtena celkem pro 90 syntetických promluv vygenerovaných pomocí každého natrénovaného modelu, výsledky tedy bylo možné následně statisticky vyhodnotit, viz kapitola 6.

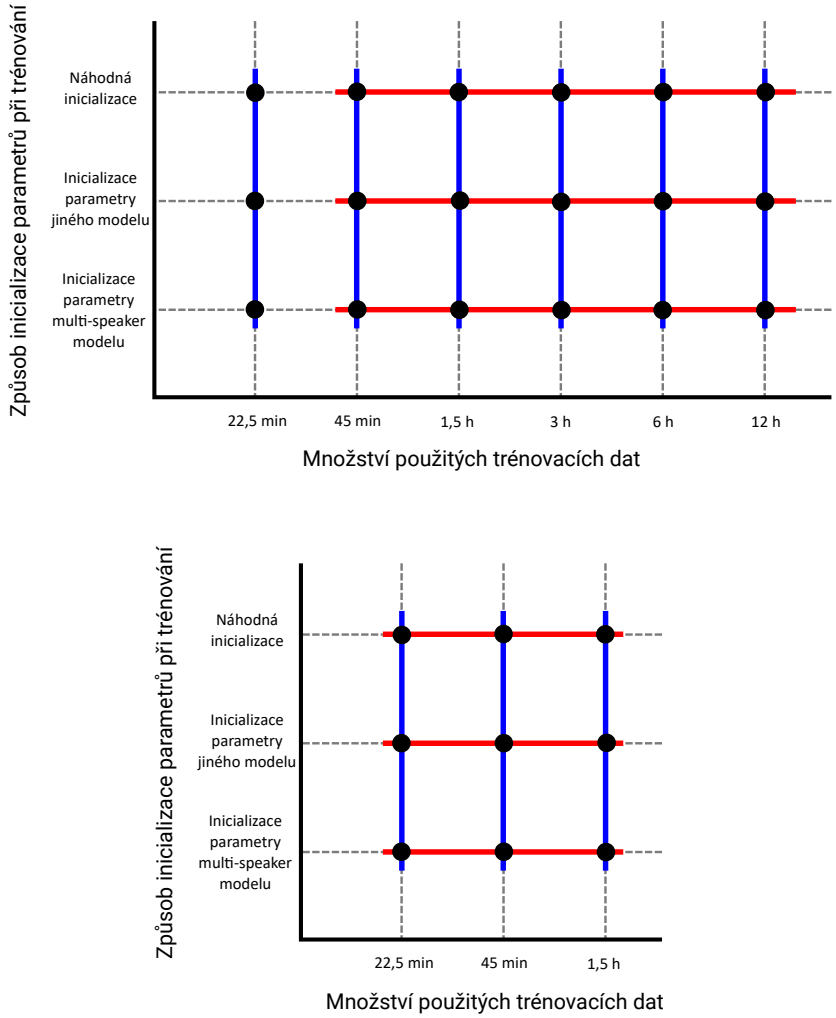
5.4.2 Hodnocení pomocí poslechových testů

Kromě hodnocení syntetické řeči pomocí vzdálenosti MCD byla též pro porovnání kvality jednotlivých syntetizérů realizována sada poslechových testů. Cílem testů bylo určit, jaký vliv má na kvalitu syntetické řeči množství použitých trénovacích dat, resp. způsob inicializace modelu při trénování.

Představme si každý natrénovaný model jako bod v rovině dané velikostí použitého řečového korpusu a způsobem inicializace parametrů při trénování. Potom můžeme provedené poslechové testy reprezentovat úsečkami, které procházejí právě těmi body, jimž odpovídají modely srovnávané v příslušném testu, viz obrázek 5.1.

Jak je naznačeno na obrázku 5.1, byly provedeny dva druhy testů. Cílem testů reprezentovaných červenými horizontálními úsečkami bylo zjistit, jaký vliv má na kvalitu syntetické řeči velikost trénovacího korpusu při daném způsobu inicializace parametrů modelu. Testy odpovídající modrým vertikálním úsečkám měly naopak porovnat kvalitu generované řeči při různém způsobu inicializace parametrů, máme-li k dispozici pevně dané množství trénovacích dat.

V případě hlasu profesionálního řečníka byly z testů pro zjištění vlivu množství trénovacích dat vyřazeny modely natrénované s využitím 22,5 minut řeči. Cílem tohoto rozhodnutí bylo snížit složitost těchto testů. Pokud by totiž byly modely natrénované z 22,5 minut nahrávek použity, museli by respondenti hodnotit celkem 7 nahrávek (6 syntetických a 1 referenční), což by bylo poměrně komplikované. Navíc bylo zcela evidentní, že syntetická řeč generovaná modely natrénovanými



Obrázek 5.1: Ilustrace provedených poslechových testů pro profesionální (nahoře), resp. amatérský (dole) hlas

pomocí 22,5 minut řeči dosahuje nejnižší kvality, a zdálo se tedy zbytečné tento fakt ověřovat pomocí poslechových testů.

Z obrázku 5.1 je zřejmé, že bylo celkem provedeno 15 poslechových testů.

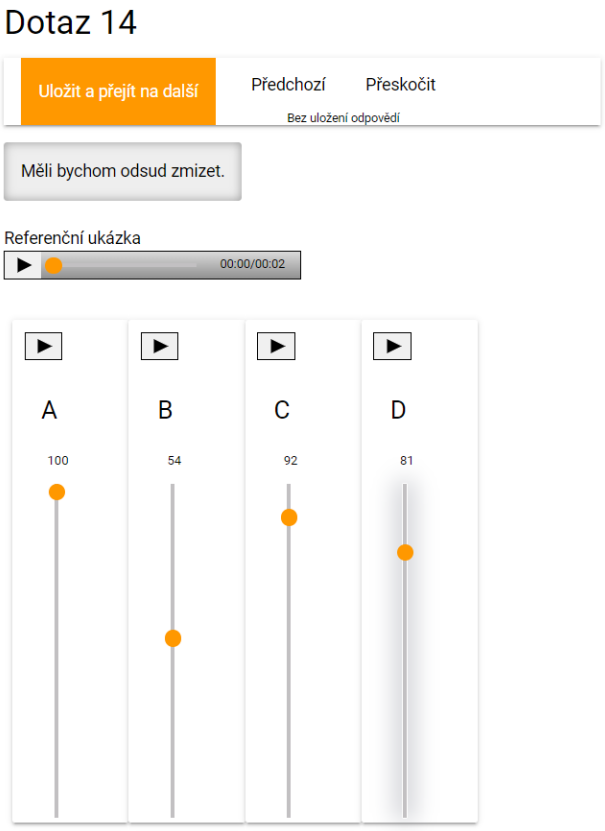
5.4.2.1 Realizace poslechových testů

Jelikož bylo nutné, aby v rámci každého poslechového testu respondenti hodnotili několik modelů současně, byly zvoleny testy typu MUSHRA (viz podkapitola 4.1). Testy byly realizovány s využitím existujícího webového frameworku [30].

Během každého testu bylo posluchačům postupně předloženo 20 promluv, přičemž každá promluva byla reprezentována několika nahrávkami, mezi nimiž byla skryta také nahrávka pořízená skutečným řečníkem. Úkolem respondentů bylo ohodnotit každou nahrávku číslem v rozsahu 0 – 100, přičemž hodnota 100 od-

povídá kvalitě skutečné lidské řeči, zatímco hodnota 0 by měla být použita pouze v případě, je-li kvalita příslušné nahrávky zcela nepřijatelná.

Respondenti měli možnost přehrát si každou nahrávku opakovaně, mohli si též pro srovnání poslechnout referenční nahrávku vytvořenou skutečným řečníkem a měli k dispozici také textový přepis posuzované promluvy. Grafické rozhraní používané při poslechových testech je zobrazeno na obrázku 5.2.



Obrázek 5.2: Grafické rozhraní webového frameworku používaného při poslechových testech

Aby respondenti provádějící více poslechových testů nebyli nuceni poslouchat stále dokola tytéž věty, což by v nich mohlo vzbudit pocit jednotvárnosti, a mohlo by tak dojít ke snížení soustředěnosti, byly do každého z poslechových testů zařazeny jiné promluvy.

5.4.2.2 Distribuce poslechových testů

Připravené poslechové testy byly následně distribuovány posluchačům k vyhodnocení. Bylo vybráno celkem 20 respondentů. Jednalo se výhradně o posluchače, pro něž je čeština, ve které byla řeč syntetizována, rodným jazykem. Mezi respondenty

byli muži i ženy z různých věkových kategorií, přičemž 30 % respondentů uvedlo, že má zkušenosti se syntézou řeči.

Poslechové testy byly respondentům rozesílány e-mailem. Jelikož bylo třeba vyhodnotit celkem 15 testů a doba strávená zodpovídáním jednoho testu byla dle počtu porovnávaných modelů odhadnuta na 20 až 30 minut, zdálo se nepřiměřené požadovat po každém posluchači provedení všech testů. Každému posluchači byla tedy zpravidla odeslána pouze část poslechových testů, v případě potřeby bylo později požádáno o provedení dalších testů.

Dále bylo předpokládáno, že budou respondenti provádět poslechové testy v pořadí, v jakém byly v e-mailu uvedeny. To znamená, že pokud se oslovený jedinec rozhodne zodpovědět pouze některé testy, pak to budou pravděpodobně testy uvedené na začátku seznamu, který obdržel. Aby byl tedy počet respondentů v jednotlivých testech přibližně vyrovnaný, byly testy každému posluchači odeslány v jiném pořadí.

5.4.2.3 Vyhodnocení poslechových testů

Uveďme nyní postup, jakým byly poslechové testy vyhodnocovány. V každém testu posluchači $p = 1, 2, 3, \ldots, P$ hodnotili několik modelů $m = 1, 2, 3, \ldots, M$ (ve většině testů byl počet hodnocených modelů $M = 4$, v některých případech $M = 6$). Každý posluchač ohodnotil každý model celkem dvacetkrát, neboť mu bylo předloženo celkem 20 nahrávek vygenerovaných každým modelem. Označme hodnocení modelu m posluchačem p na základě k-té nahrávky jako $h_m^p(k)$. Potom můžeme spočítat průměrné hodnocení modelu m posluchačem p jako

$$h_m^p = \frac{1}{20} \sum_{k=1}^{20} h_m^p(k). \tag{5.1}$$

Ukázalo se, že každý posluchač přistupuje k hodnocení syntetické řeči odlišně. Zatímco někteří respondenti využívali pouze malý rozsah hodnocení, jiní udělovali hodnocení z celé škály 0 – 100. Díky tomu by hodnocení některých posluchačů mohlo mít větší váhu než hodnocení jiných, což nebylo žádoucí. Proto byla hodnocení každého posluchače normalizována.

Každý posluchač byl charakterizován odhadem střední hodnoty μ_p a rozptylu σ_p^2 hodnocení, které používal. Tyto hodnoty byly spočteny podle vztahů

$$\begin{aligned} \mu_p &= \frac{1}{M} \sum_{m=1}^{M} h_m^p \\ \sigma_p^2 &= \frac{1}{M} \sum_{m=1}^{M} (h_m^p - \mu_p)^2. \end{aligned} \tag{5.2}$$

Následně byla odhadnuta střední hodnota a rozptyl hodnocení „průměrného posluchače“ jako aritmetický průměr těchto hodnot, tedy

$$\begin{aligned} \mu &= \frac{1}{P}\sum_{p=1}^{P}\mu_p \\ \sigma^2 &= \frac{1}{P}\sum_{p=1}^{P}\sigma_p^2. \end{aligned} \tag{5.3}$$

Pak již bylo možné průměrné hodnocení každého modelu m posluchačem p normalizovat tak, aby odhad střední hodnoty a rozptylu hodnocení tohoto posluchače odpovídal odhadu střední hodnoty a rozptylu „průměrného posluchače“. Normalizovaná hodnota $\tilde{h}_m^p$ byla získána podle předpisu

$$\tilde{h}_m^p = \frac{h_m^p - \mu_p}{\sqrt{\sigma_p^2}} \cdot \sqrt{\sigma^2} + \mu. \tag{5.4}$$

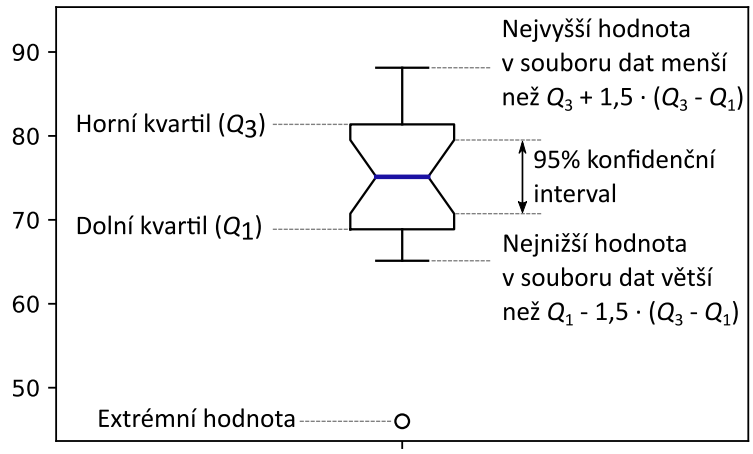
Poznamenejme, že při použití takovéto normalizace není zaručeno, že bude výsledné hodnocení $\tilde{h}_m^p$ ležet v intervalu $\langle 0; 100\rangle$, ve kterém leželo původní hodnocení h_m^p před normalizací.

Pomocí výše popsaného postupu byla pro každý hodnocený model m získána množina hodnocení $\left\{\tilde{h}_m^p; p \in \{1, 2, 3, \ldots, P\}\right\}$, na základě které byl vykreslen tzv. *krabicový diagram* (angl. *boxplot*).

Boxplot je diagram, který graficky znázorňuje hned několik statistických charakteristik zkoumaného souboru dat. Příklad tohoto diagramu je s vysvětlujícími popisky znázorněn na obrázku 5.3. Boxplot vykresluje mj. dolní kvartil Q_1 a horní kvartil Q_3, tedy hodnoty, mezi kterými se nachází 50 % naměřených dat. Součástí diagramu jsou dále tzv. *vousy* (angl. *whiskers*). Dolní vous představuje nejnižší hodnotu obsaženou ve zkoumaném souboru, která je vyšší než hodnota výrazu $Q_1 - 1{,}5 \cdot (Q_3 - Q_1)$. Horní vous analogicky odpovídá nejvyšší hodnotě z uvažované množiny, která je menší než hodnota výrazu $Q_3 + 1{,}5 \cdot (Q_3 - Q_1)$. Obsahuje-li soubor data, která neleží mezi těmito vousy, pak jsou tyto extrémní hodnoty (angl. *outliers*) vykreslovány individuálně v podobě kroužků.

Pro vzájemné porovnání natrénovaných modelů je podstatný především tzv. 95% *konfidenční interval*, který odpovídá výšce výřezu v diagramu (viz obrázek 5.3). Chápeme-li normalizované hodnocení modelu $\tilde{h}_m^p$ jako náhodnou veličinu, pro níž bylo právě spočteno několik realizací, pak s pravděpodobností 95 % leží medián této náhodné veličiny právě v 95% konfidenčním intervalu.

Srovnávání kvality syntetické řeči na základě konfidenčních intervalů je velice praktické, neboť tato metoda umožňuje určit, zda je rozdíl mezi dvěma syntetizéry statisticky významný. Uvažujme např. situaci znázorněnou na obrázku 5.4.



Obrázek 5.3: Krabicový diagram s vyznačenými statistickými charakteristikami

Je zřejmé, že 95% konfidenční interval spočtený pro syntetizér A obsahuje vyšší hodnocení nežli interval odpovídající syntetizéru B a oba intervaly se vzájemně nepřekrývají. Označíme-li tedy medián hodnocení syntetizéru A, resp. B symbolem $\tilde{h}_1^{\text{med}}$, resp. $\tilde{h}_2^{\text{med}}$ a příslušné konfidenční intervaly $I_1^{0,95}$, resp. $I_2^{0,95}$, můžeme prohlásit, že je medián hodnocení syntetizéru A vyšší nežli medián hodnocení syntetizéru B s pravděpodobností

$$\begin{aligned} P\left(\tilde{h}_1^{\text{med}} > \tilde{h}_2^{\text{med}}\right) &\geq P\left(\tilde{h}_1^{\text{med}} \in I_1^{0,95} \ \wedge \ \tilde{h}_2^{\text{med}} \in I_2^{0,95}\right) = \\ &= P\left(\tilde{h}_1^{\text{med}} \in I_1^{0,95}\right) \cdot P\left(\tilde{h}_2^{\text{med}} \in I_2^{0,95}\right) = 0{,}95 \cdot 0{,}95 \doteq 0{,}90. \end{aligned} \tag{5.5}$$

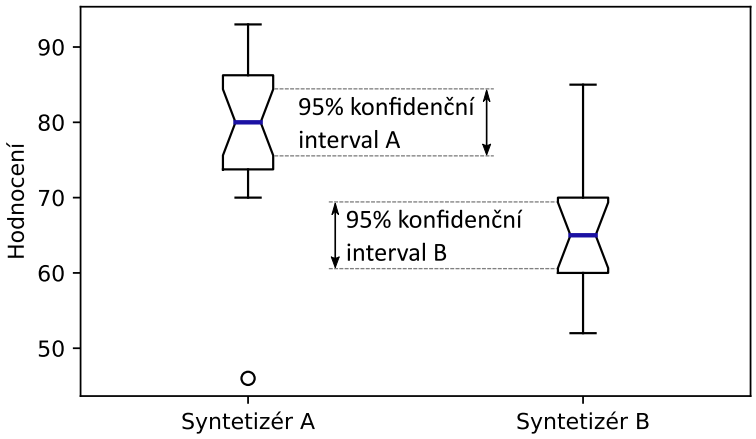
Tímto způsobem byly natrénované modely porovnávány, viz kapitola 6.

Dále byl navržen mechanismus, který měl z výsledků poslechových testů vyřadit odpovědi „nedůvěryhodných“ respondentů. V každém testu byla jedna z promluv zopakována dvakrát a bylo ověřováno, zda v obou případech hodnotí posluchači předložené nahrávky „podobně“.

Bez újmy na obecnosti předpokládejme, že byla stejná promluva posluchači p předložena jako první a druhá nahrávka, tj. pro $k = 1$ a $k = 2$. Potom bylo hodnocení posluchače zachováno pouze v případě, že nahrávku, kterou respondent v první iteraci označil jako nejlepší, označil i v druhé iterace jako lepší nežli nahrávku, kterou v první iteraci označil jako nejhorší, a naopak. Jinak řečeno, musely být splněny dvě podmínky, a sice

$$\begin{aligned} h^p_{\arg\max_m h^p_m(1)}(2) &\geq h^p_{\arg\min h^p_m(1)}(2) \\ h^p_{\arg\max_m h^p_m(2)}(1) &\geq h^p_{\arg\min h^p_m(2)}(1). \end{aligned} \tag{5.6}$$

Pokud tyto podmínky nebyly splněny, indikovalo to, že daný posluchač nehodnotil nahrávky příliš pozorně, a jeho odpovědi byly z výsledků vyřazeny.



Obrázek 5.4: Krabicový diagram vykreslený pro dva syntetizéry, mezi jejichž kvalitou je statisticky významný rozdíl

Nakonec však mechanismus kontroly věrohodnosti posluchačů nebyl použit, neboť respondentů, kteří poslechové testy provedli, nebylo příliš mnoho (každý test byl proveden deseti lidmi) a nezdálo se tedy rozumné jejich počet ještě dále redukovat. Navíc se ukázalo, že vyřazení nevěrohodných posluchačů nezpůsobí žádné výrazné změny ve výsledcích experimentů, proto byly nakonec pro hodnocení natrénovaných modelů použity odpovědi všech respondentů.

Výsledky experimentů

6

Tato část práce je věnována vyhodnocení experimentů popsaných v kapitole 5. Kapitola je rozdělena na dvě části, přičemž jedna z částí je věnována experimentům prováděným pro hlas profesionálního řečníka a druhá část experimentům s hlasem amatérského řečníka. V každé části jsou uvedeny nejprve výsledky experimentů hodnotících vliv množství trénovacích dat na kvalitu syntetické řeči a následují experimenty zkoumající vliv použití předtrénovaných modelů.

Výsledky každého experimentu jsou hodnoceny pomocí vzdálenosti MCD a pomocí poslechových testů typu MUSHRA. V případě každého experimentu byla vzdálenost MCD od referenční nahrávky spočtena celkem pro 90 promluv, pro každý model byl tedy získán soubor 90 hodnot. Na základě tohoto souboru byl vykreslen krabicový diagram, který vykresluje mj. 95% konfidenční interval, ve kterém s pravděpodobností 95 % leží medián vzdálenosti MCD. Pomocí krabicových diagramů lze tedy modely snadno vizuálně porovnat.

Každý z poslechových testů byl proveden právě deseti respondenty, v rámci každého testu byl tedy každý model reprezentován souborem deseti hodnocení, na základě kterých byl stejně jako v případě vzdálenosti MCD vykreslen krabicový diagram.

Při poslechových testech byly kromě syntetických promluv hodnoceny také originální nahrávky řečníka, což čtenáři umožňuje posoudit, jak zásadní je rozdíl mezi syntetickou řečí a referenčními nahrávkami. Vzdálenost MCD nebyla pro hodnocení skutečného hlasu počítána, neboť z její definice je zřejmé, že vzdálenost referenční nahrávky od sebe samé je nulová, viz podkapitola 4.2.

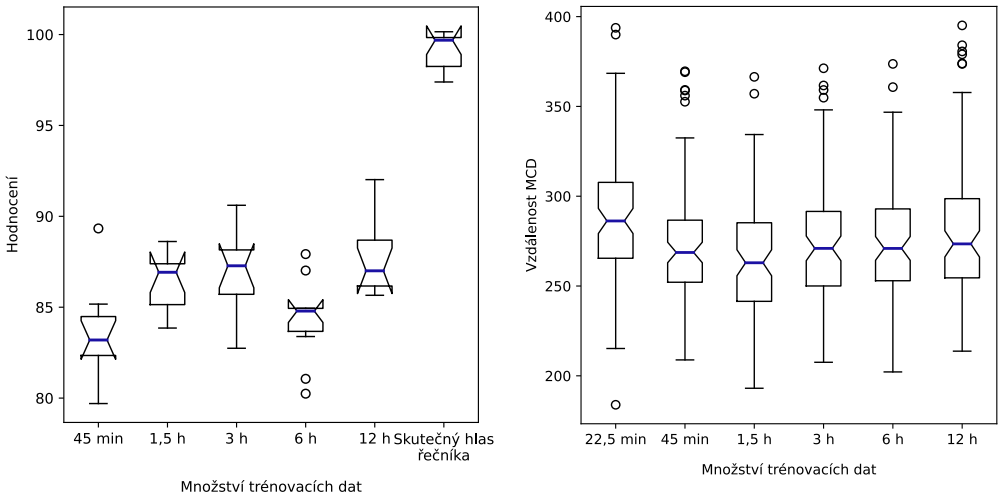
Při vyhodnocování experimentů byl kladen větší důraz na výsledky poslechových testů, neboť se zřejmě jedná o spolehlivější metodu hodnocení syntetické řeči nežli výpočet vzdálenosti MCD. V mnoha případech došlo k tomu, že se hodnocení pomocí objektivní míry MCD s hodnocením pomocí poslechových testů shodovalo, velmi častá byla však též situace, kdy na základě vzdálenosti MCD nebylo možné vyvodit statisticky významný závěr, viz dále.

6.1 Hlas profesionálního řečníka

6.1.1 Vliv množství trénovacích dat na kvalitu syntetické řeči

V této sekci budou představeny výsledky experimentů, které byly provedeny za účelem vypozorování vlivu množství trénovacích dat na kvalitu syntetické řeči, používáme-li k natrénování modelu kvalitní nahrávky pocházející od zkušeného řečníka.

Uvažujme nejprve situaci, kdy model před trénováním inicializujeme náhodně zvolenými parametry. Modely natrénované s využitím různého množství trénovacích dat jsou na základě poslechových testů a objektivní míry MCD porovnány na obrázku 6.1.



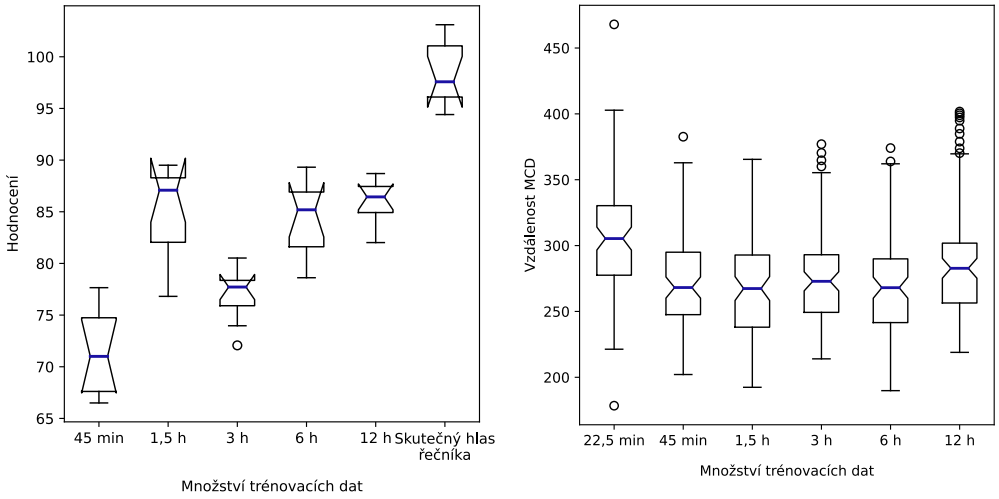
Obrázek 6.1: Porovnání modelů syntetizujících hlas profesionálního řečníka při náhodné inicializaci parametrů a různém množství trénovacích dat. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD

Ze srovnání na základě vzdálenosti MCD je patrné, že nejhorší kvality dosahoval model natrénovaný s využitím pouze 22,5 minut řeči, což ostatně bylo i sluchem zcela zřejmé. Jiné závěry však na základě vzdálenosti MCD není možné vyvodit. Ze srovnání pomocí poslechového testu lze vypozorovat, že ani model využívající 45 minut trénovacích dat nedosáhl příliš dobrých výsledků, počínaje modelem natrénovaným z 1,5 hodiny řeči se však zdá, že již velikost řečového korpusu nehraje zásadní roli. Překvapivé je nízké hodnocení modelu natrénovaného na základě 6 hodin nahrávek, z obrázku 6.1 je však patrné, že soubor hodnocení tohoto modelu obsahoval poměrně vysoký počet extrémních hodnocení (tzv. *outlierů*), vnímání kvality daného syntetizéru se tedy patrně napříč respondenty dosti lišilo.

Obecně však lze na základě výsledků tohoto testu doporučit při náhodné inicializaci modelu použít alespoň 1,5 hodiny trénovacích nahrávek. Jednoznačně nejlepší kvality však stále dosahuje originální hlas řečníka, což odpovídá očekáváním.

Pozastavme se nad nepřirozeným tvarem krabicových diagramů vykreslených na základě poslechových testů. V mnoha případech byl obdržen natolik široký 95% konfidenční interval, že příslušný výřez v boxplotu přesahuje hranice diagramu dané dolním a horním kvartilem. Široký konfidenční interval znamená méně přesný odhad pozice mediánu hodnocení, což je velmi pravděpodobně způsobeno nízkým počtem dat ve zkoumaném souboru, tj. nízkým počtem respondentů hodnotících každý model.

Porovnejme dále vliv velikosti použitého řečového korpusu na kvalitu syntetické řeči, použijeme-li princip transfer learning, tj. inicializujeme-li před trénováním parametry modelu pomocí modelu natrénovaného pro hlas jiného profesionálního řečníka. Výsledky experimentu jsou vykresleny na obrázku 6.2.

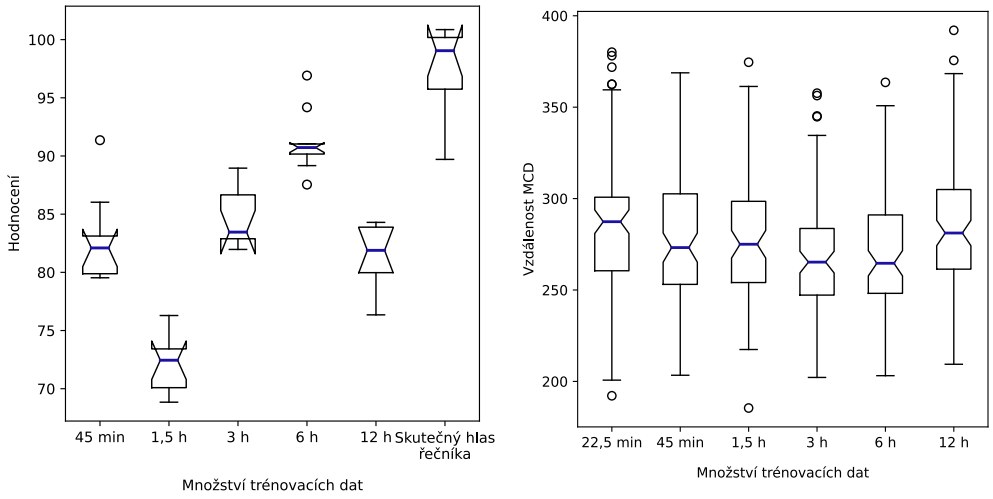


Obrázek 6.2: Porovnání modelů syntetizujících hlas profesionálního řečníka při inicializaci parametrů pomocí modelu natrénovaného pro jiného řečníka a při různém množství trénovacích dat. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD

Stejně jako v případě náhodné inicializace parametrů lze ze srovnání na základě vzdálenosti MCD konstatovat, že nejhoršího výsledku dosahuje model natrénovaný s využitím 22,5 minut řečových nahrávek. Výsledek poslechových testů je rovněž velmi podobný výsledkům předchozího experimentu – při použití 45 minut trénovacích dat není kvalita syntetické řeči příliš vysoká, použijeme-li však 1,5 hodiny nahrávek či více, kvalita řeči se již příliš nemění.

Stejně jako v případě předchozího experimentu dosáhl jeden z modelů využívajících větší množství trénovacích dat nepříliš uspokojivého výsledku, tentokrát

se jedná o model natrénovaný na základě 3 hodin řeči. Lze polemizovat o tom, zda je tento jev způsoben skutečností, že při trénování daného modelu dospěla kvalita generované řeči pouze do jakéhosi lokálního optima, či např. tím, že byly do poslechových testů náhodou vybrány promluvy, jejichž výslovnost byla právě pro tento model komplikovaná. Jednoznačnou odpověď na tuto otázku bohužel zřejmě nelze poskytnout.



Obrázek 6.3: Porovnání modelů syntetizujících hlas profesionálního řečníka při inicializaci parametrů pomocí multi-speaker modelu a při různém množství trénovacích dat. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD

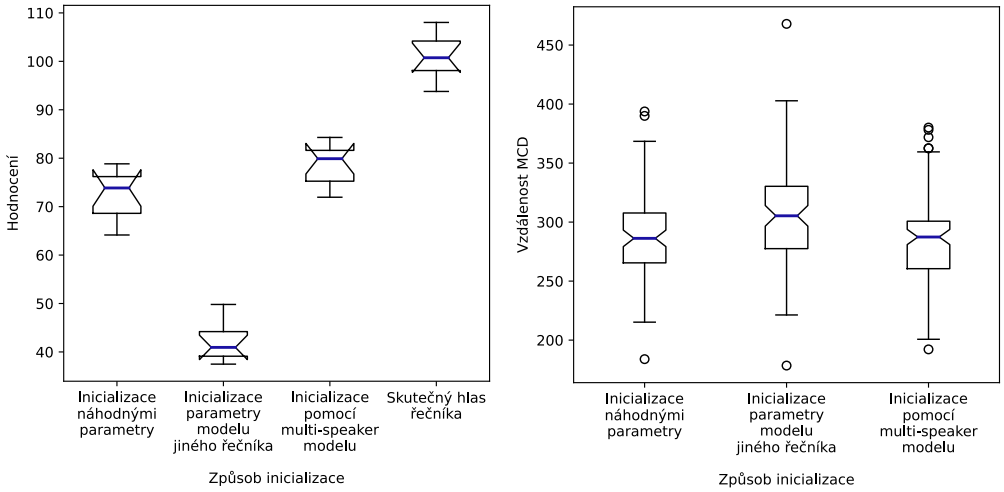
V případě modelů inicializovaných parametry multi-speaker modelu se nepodařilo najít jednoznačný trend kvality syntetické řeči v závislosti na množství použitých trénovacích dat, viz obrázek 6.3. Zdá se, že multi-speaker model má v sobě již natolik dobře zakódované charakteristiky lidské řeči, že stačí poměrně malé množství trénovacích dat, aby výsledná syntetická řeč zněla kvalitně. Mezi modely využívajícími 45 minut, 3 hodiny a 12 hodin řeči poslechové testy neodhalily zásadní rozdíl, pouze modely založené na 1,5 a 6 hodinách trénovacích nahrávek dosahují výrazně odlišné kvality, což však lze pravděpodobně považovat spíše za anomálie konkrétních modelů nežli za důsledek daného množství trénovacích dat.

6.1.2 Vliv způsobu inicializace parametrů modelu na kvalitu syntetické řeči

Uvažujme nyní situaci, kdy máme k dispozici pouze 22,5 minut nahrávek hlasu profesionálního řečníka a chceme pomocí nich natrénovat co možná nejkvalitnější

syntetizér. Na obrázku 6.4 jsou porovnány modely, při jejichž trénování byly použity různé způsoby inicializace parametrů.

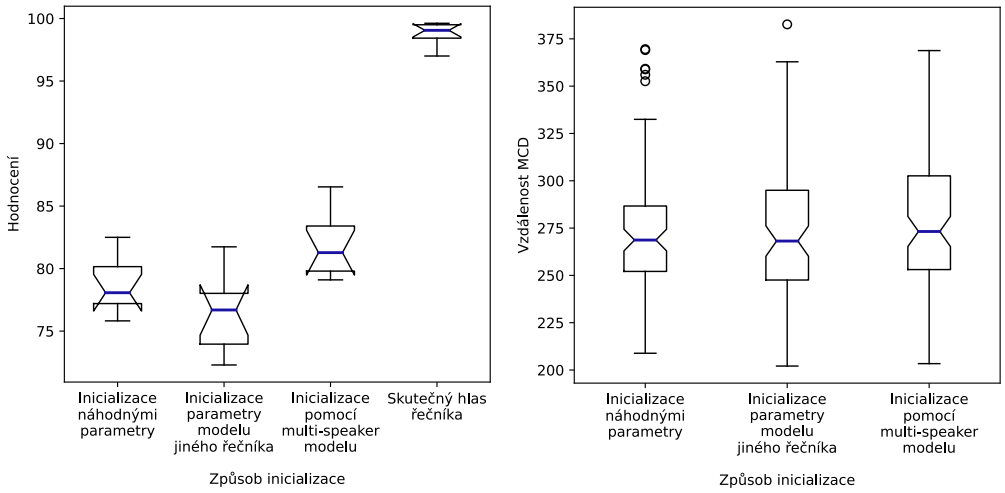
Výsledky poskytnuté poslechovými testy i objektivní mírou MCD se v tomto případě zcela shodují – mezi náhodnou inicializací modelu a použitím parametrů multi-speaker modelu není statisticky významný rozdíl, při fine-tuningu modelu natrénovaného pro hlas jiného řečníka však syntetická řeč dosahuje jednoznačně nejhorších výsledků. Tato skutečnost je poměrně překvapivá, může však být způsobena mj. tím, že si hlas cílového řečníka a hlas syntetizovaný modelem použitým pro fine-tuning „nejsou příliš podobné“.



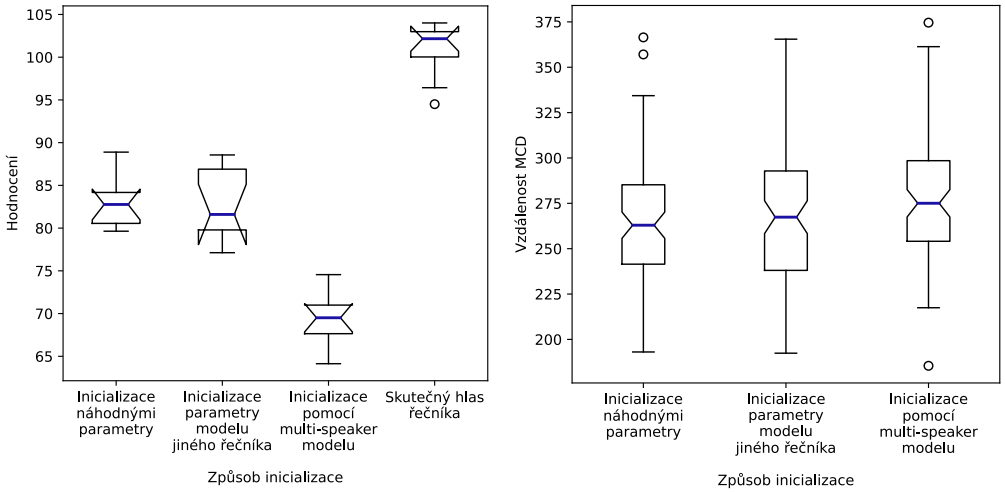
Obrázek 6.4: Porovnání modelů syntetizujících hlas profesionálního řečníka při použití 22,5 min řeči a různém způsobu inicializace parametrů modelu. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD

Máme-li k dispozici 45 minut řeči, není již rozdíl mezi způsoby inicializace trénování tak výrazný, poslechové testy však stále poukazují na to, že lepší kvality dosáhneme použitím multi-speaker modelu nežli použitím modelu natrénovaného pro hlas jediného řečníka, viz obrázek 6.5.

V případě použití řečového korpusu o rozsahu 1,5 hodiny nahrávek náhle došlo k obrácení trendu – kvalita syntetické řeči při fine-tuningu modelu natrénovaného pro hlas jednoho řečníka v poslechovém testu předčila kvalitu syntetizéru natrénovaného s využitím parametrů multi-speaker modelu, viz obrázek 6.6. Je však otázkou, zda se nejedná pouze o jakýsi ojedinělý úkaz spojený s konkrétními modely, neboť při použití 3 hodin trénovacích nahrávek opět dosahuje lepších výsledků syntetizér inicializovaný parametry multi-speaker modelu, viz obrázek 6.7.

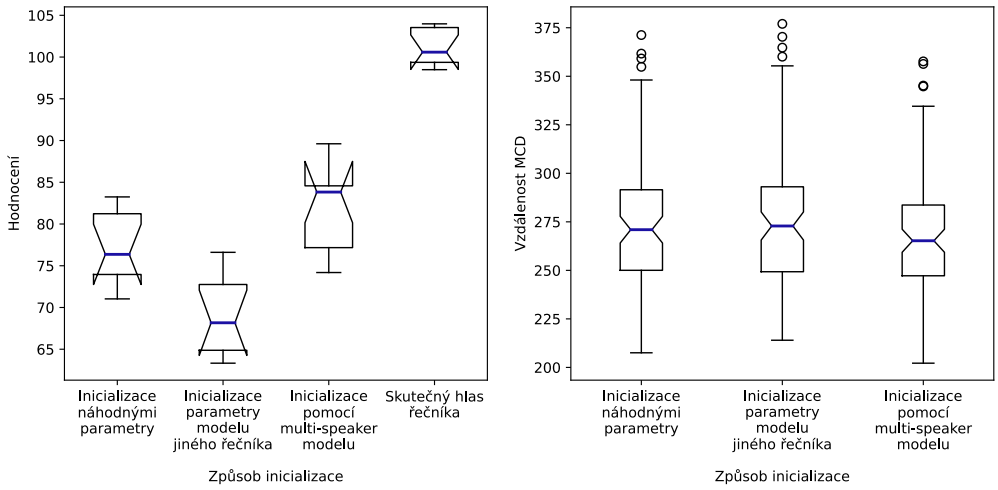


Obrázek 6.5: Porovnání modelů syntetizujících hlas profesionálního řečníka při použití 45 min řeči a různém způsobu inicializace parametrů modelu. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD

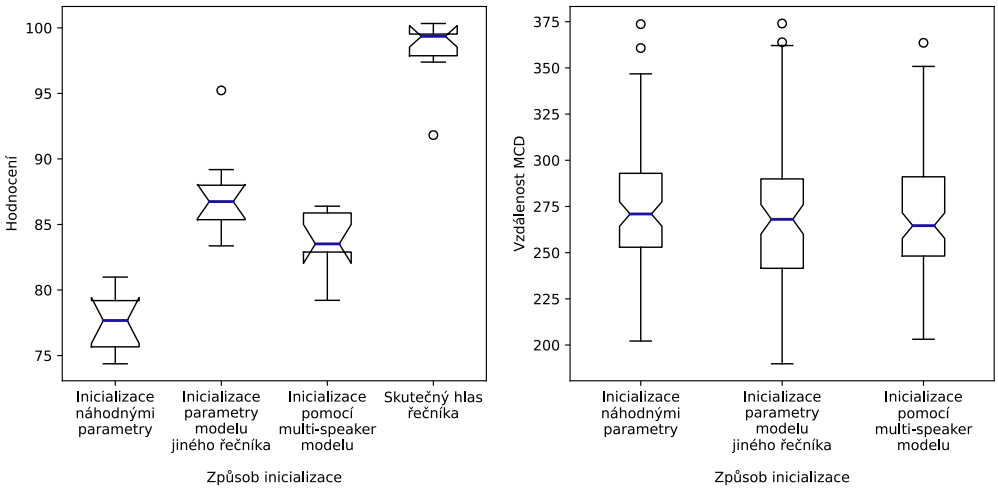


Obrázek 6.6: Porovnání modelů syntetizujících hlas profesionálního řečníka při použití 1,5 h řeči a různém způsobu inicializace parametrů modelu. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD

V případě použití 6 hodin trénovacích dat poslechový test zcela jasně ukázal, že je nejvýhodnější inicializace modelu parametry modelu jiného řečníka, následuje využití multi-speaker modelu a nejhorších výsledků dosáhl model inicializovaný náhodně, viz obrázek 6.8. K podobnému výsledku pravděpodobně dochází i v případě modelů natrénovaných s využitím 12 hodin trénovacích dat, v tomto případě však již není rozdíl statisticky významný, protože se konfidenční intervaly spočtené pro jednotlivé modely vzájemně překrývají, viz obrázek 6.9.



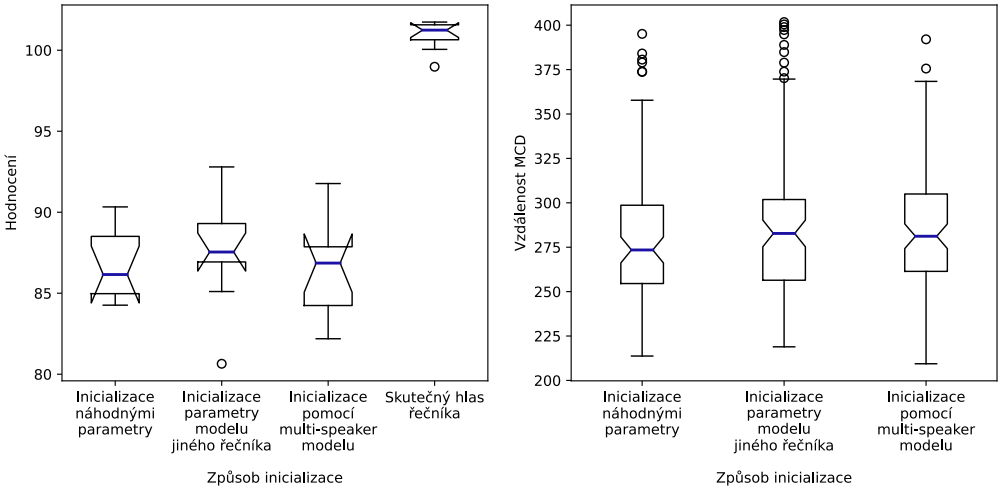
Obrázek 6.7: Porovnání modelů syntetizujících hlas profesionálního řečníka při použití 3 h řeči a různém způsobu inicializace parametrů modelu. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD



Obrázek 6.8: Porovnání modelů syntetizujících hlas profesionálního řečníka při použití 6 h řeči a různém způsobu inicializace parametrů modelu. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD

6.2 Hlas amatérského řečníka

V podkapitole 6.1 byly popsány výsledky experimentů zkoumajících kvalitu syntetické řeči generované pro hlas profesionálního řečníka v závislosti na způsobu inicializace modelu a na množství použitých trénovacích dat. Stejné experimenty byly provedeny také s hlasem amatérského řečníka. Výsledkům těchto experimentů je věnována tato podkapitola.



Obrázek 6.9: Porovnání modelů syntetizujících hlas profesionálního řečníka při použití 12 h řeči a různém způsobu inicializace parametrů modelu. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD

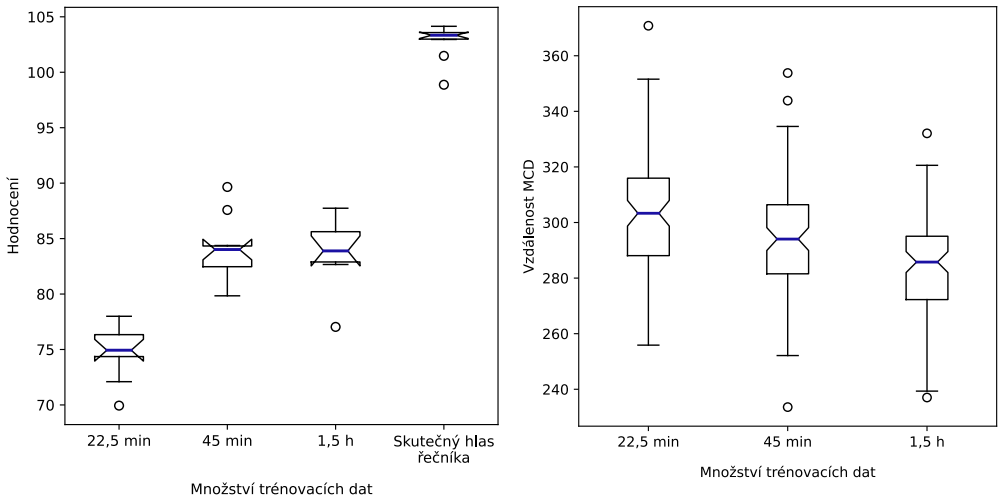
6.2.1 Vliv množství trénovacích dat na kvalitu syntetické řeči

Cílem prvního experimentu bylo zjistit, jak závisí kvalita řeči generované syntetizérem na množství použitých trénovacích nahrávek, inicializujeme-li před trénováním parametry modelu náhodně. Výsledky experimentu jsou vykresleny ve formě krabicových diagramů na obrázku 6.10.

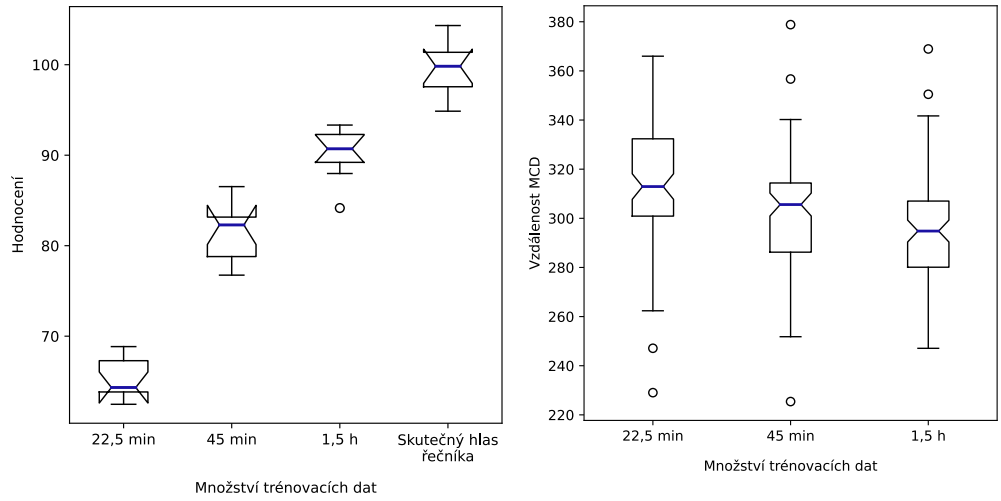
Z hodnocení na základě vzdálenosti MCD lze usuzovat, že větší množství trénovacích dat implikuje lepší kvalitu syntetické řeči. Poslechový test potvrdil, že při použití pouhých 22,5 minut trénovacích nahrávek není kvalita výsledné syntézy příliš vysoká, mezi modely natrénovanými s využitím 45 minut řeči a 1,5 hodiny řeči však nebyl prokázán statisticky významný rozdíl.

Velmi podobná situace nastala při hodnocení modelů, jejichž trénování bylo inicializováno parametry modelu natrénovaného pro hlas jiného řečníka. V tomto případě již poslechový test zcela jasně prokázal, že větší trénovací korpus vede na lepší kvalitu řeči generované natrénovaným syntetizérem, viz obrázek 6.11.

Při inicializaci parametry multi-speaker modelu dosáhl kupodivu model natrénovaný s využitím 22,5 minut řeči v rámci poslechového testu lepších výsledků nežli model využívající 1,5 hodiny řeči, viz obrázek 6.12. Jinak však na základě tohoto experimentu nebyl prokázán žádný statisticky významný trend.



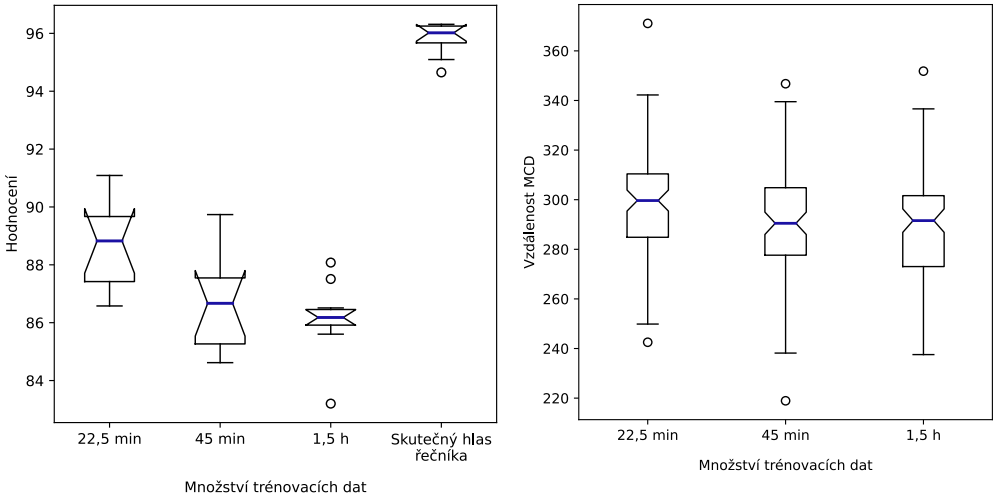
Obrázek 6.10: Porovnání modelů syntetizujících hlas laického řečníka při náhodné inicializaci parametrů a různém množství trénovacích dat. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD



Obrázek 6.11: Porovnání modelů syntetizujících hlas laického řečníka při inicializaci parametrů pomocí modelu natrénovaného pro jiného řečníka a různém množství trénovacích dat. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD

6.2.2 Vliv způsobu inicializace parametrů modelu na kvalitu syntetické řeči

Zabývejme se dále otázkou, jak v případě hlasu laického řečníka ovlivní kvalitu syntetické řeči způsob inicializace modelu při trénování, máme-li k dispozici konkrétní množství trénovacích dat.



Obrázek 6.12: Porovnání modelů syntetizujících hlas laického řečníka při inicializaci parametrů pomocí multi-speaker modelu a různém množství trénovacích dat. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD

Z modelů využívajících pouze 22,5 minuty trénovacích dat dosáhl při poslechovém testu nejlepšího výsledku model inicializovaný pomocí multi-speaker modelu, nejhůře naopak dopadl model inicializovaný parametry modelu natrénovaného pro hlas jiného řečníka, viz obrázek 6.13. Na skutečnost, že při fine-tuningu tohoto modelu nedosahuje generovaná řeč příliš vysoké kvality, ostatně poukazuje také hodnocení na základě objektivní míry MCD.

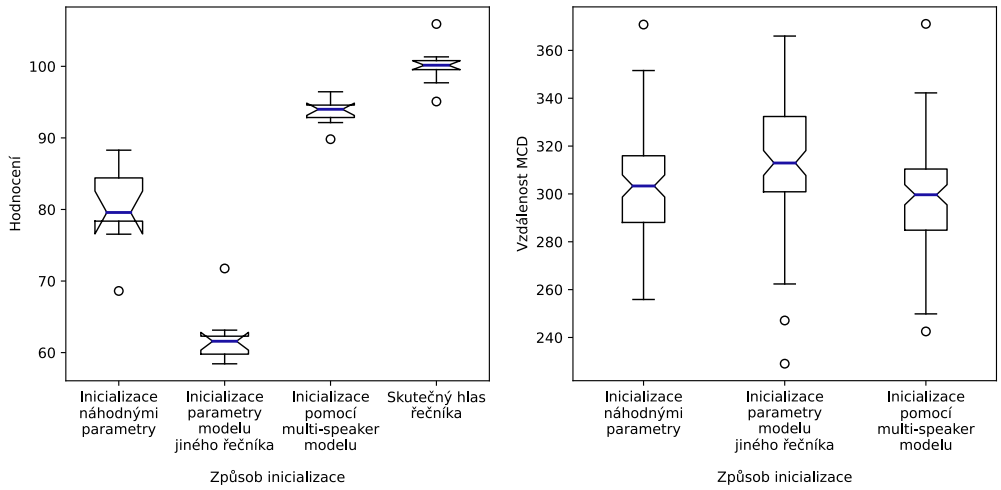
Při srovnávání modelů natrénovaných na základě 45 minut řeči již nebyly rozdíly mezi jednotlivými modely tak výrazné, přesto však bylo z výsledků poslechových testů i z hodnocení pomocí vzdálenosti MCD zcela zřejmé, že nejhorší kvality stále dosahuje model dotrénovaný z parametrů modelu jiného řečníka, viz obrázek 6.14.

Ani v případě, kdy byl k dispozici řečový korpus o rozsahu 1,5 hodiny, neobdržel syntetizér využívající předtrénovaný model hlasu jiného řečníka příliš vysoká hodnocení, viz obrázek 6.15.

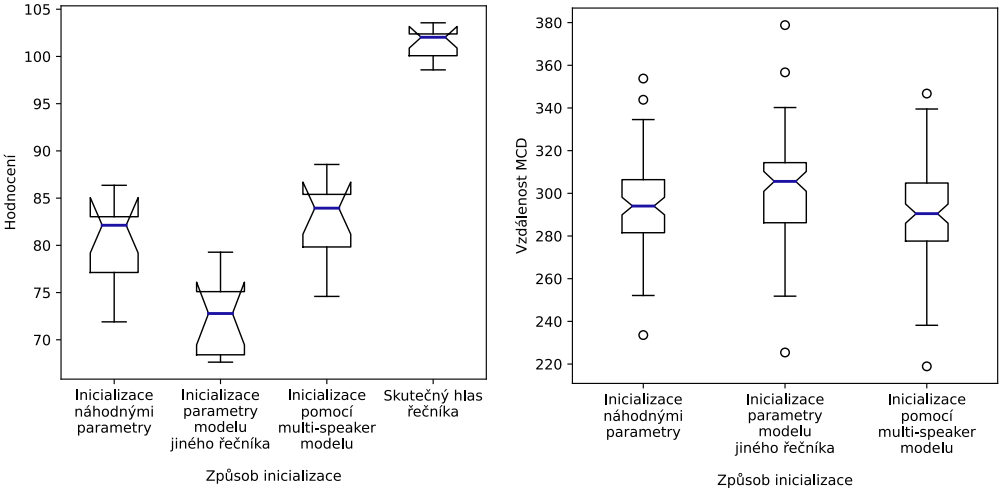
6.3 Shrnutí výsledků experimentů

Shrňme nyní výsledky experimentů popsané v předchozích podkapitolách a zformulujme na jejich základě obecné rady, jimiž by se měl návrhář systému TTS řídit, aby dosáhl co možná nejlepší kvality generované syntetické řeči.

Máme-li možnost zvolit množství trénovacích dat použitých pro trénování, pak lze na základě provedených experimentů jednoznačně doporučit použití korpusu



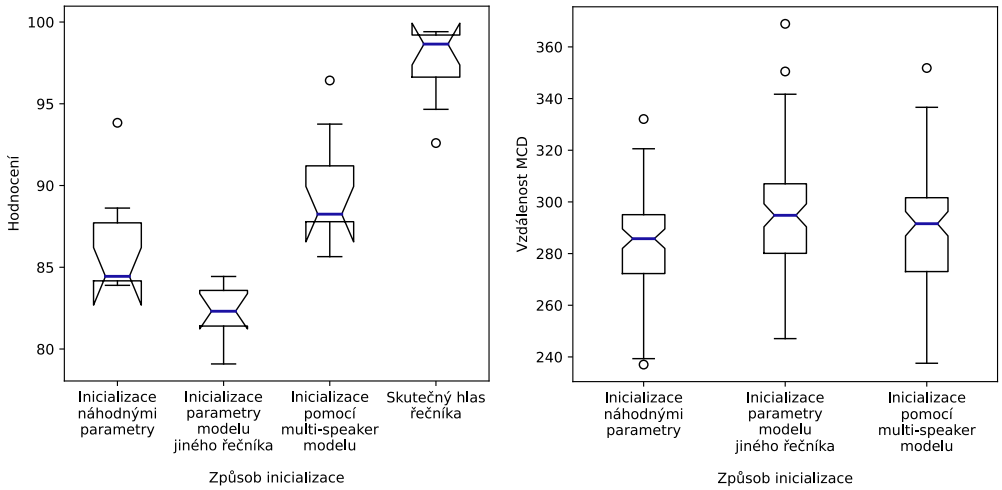
Obrázek 6.13: Porovnání modelů syntetizujících hlas laického řečníka při použití 22,5 min řeči a různém způsobu inicializace parametrů modelu. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD



Obrázek 6.14: Porovnání modelů syntetizujících hlas laického řečníka při použití 45 min řeči a různém způsobu inicializace parametrů modelu. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD

obsahujícího minimálně 45 minut řeči. Není přitom podstatné, zda rekonstruujeme hlas profesionálního či nezkušeného řečníka, ani to, zda hodláme model inicializovat náhodnými parametry či parametry modelu natrénovaného z nahrávek hlasu jiného řečníka. Použití více než 45 minut trénovacích dat již ve většině experimentů neprokázalo zlepšení kvality syntetické řeči.

Chceme-li inicializovat trénovaný model parametry multi-speaker modelu, pak



Obrázek 6.15: Porovnání modelů syntetizujících hlas laického řečníka při použití 1,5 h řeči a různém způsobu inicializace parametrů modelu. Levý diagram je založen na výsledcích poslechového testu a pravý diagram na vzdálenosti MCD

na základě provedených experimentů nelze konstatovat, použití jakého množství trénovacích dat je nejvhodnější. Zdá se, že multi-speaker model sám o sobě natolik dobře vystihuje lidskou řeč, že již velikost použitého řečového korpusu kvalitu syntetické řeči výrazně neovlivní.

Zabývejme se dále situací, při níž máme pevně stanovené množství trénovacích dat a můžeme rozhodnout pouze o tom, jakým způsobem inicializujeme parametry modelu při trénování. Nemáme-li k dispozici více než 3 hodiny trénovacích dat, pak lze na základě provedených experimentů doporučit použití multi-speaker modelu. Náhodná volba parametrů vedla v případě většiny experimentů k podobně kvalitní syntetické řeči, rozhodně však nelze doporučit inicializaci parametrů pomocí modelu natrénovaného pro hlas jiného řečníka.

Skutečnost, že fine-tuning vedl k horším výsledkům nežli náhodná volba parametrů, je poněkud překvapivá, přesto však k podobným jevům v oblasti strojového učení občas dochází. Oblast, pro kterou je natrénován model použitý pro fine-tuning (v případě syntézy řeči hlas daného řečníka) může být zdánlivě podobná cílové oblasti, tato podobnost však může být zavádějící a použitím principu transfer-learning kvalitu výsledného systému nezlepšíme, ale naopak zhoršíme. Tento fenomén se nazývá *negativní transfer* a zmiňuje jej např. článek [29].

Máme-li však k dispozici větší množství trénovacích dat (alespoň 6 hodin řečových nahrávek), pak již lze fine-tuning modelu jiného řečníka doporučit.

6.4 Omezení platnosti vyvozených závěrů a náměty pro další výzkum

Nutno poznamenat, že závěry popisované v podkapitole 6.3 nemusejí být platné ve všech situacích. V první řadě je třeba mít na paměti, že experimenty byly prováděny pouze s end-to-end modelem VITS, vliv velikosti řečového korpusu a přístupu transfer-learning na kvalitu řeči generované jinými neurálními modely může být zcela jiný.

Dále poznamenejme, že vzhledem k rozsahu této práce a k řečovým korpusům, které byly k dispozici, byly experimenty prováděny pouze s jedním profesionálním a s jedním amatérským hlasem. Aby však bylo možné zcela přesvědčivě prokázat platnost vyvozených závěrů, bylo by vhodné provést tytéž experimenty s větším množstvím hlasů, a to s hlasy ženskými i mužskými.

Stejně tak by bylo vhodné při fine-tuningu neurálních syntetizérů otestovat větší počet předtrénovaných modelů, neboť lze předpokládat, že kvalita výsledné syntetické řeči závisí na tom, do jaké míry je hlas cílového řečníka podobný hlasu, pro nějž byl natrénován předtrénovaný model.

Nakonec uveďme také fakt, že byl multi-speaker model používaný při experimentech natrénován pouze na základě hlasů 6 řečníků. Pokud bychom měli k dispozici model, pro jehož natrénování byly použity nahrávky např. stovek řečníků, pak bychom jej mohli považovat za skutečně obecný model lidské řeči a bylo by zajímavé pozorovat, jak fine-tuning tohoto modelu ovlivní kvalitu výsledné syntetické řeči. Stejně tak by bylo možné zkoumat, jaký vliv má na kvalitu syntetické řeči fine-tuning tzv. *multi-language* modelu, tj. modelu, pro jehož natrénování byly použity nahrávky řeči v různých jazycích.

6.5 Poznatky získané během experimentů

Uveďme nakonec dvě zajímavé skutečnosti, které vyplynuly z prováděných experimentů.

V první řadě se jedná o problematickou výslovnost fonému /r/, na kterou bylo upozorněno již v podkapitole 5.3. V případě všech 27 natrénovaných modelů docházelo poměrně často k tzv. ráčkování, ačkoliv řečníci, jejichž hlasy byl syntetizovány, touto vadou řeči netrpěli. Jedná se patrně o obecný problém spojený s neurálním modelem VITS a je třeba upozornit na to, že při syntéze českých textů jde o značný nedostatek.

Při trénování jednoho z modelů došlo k dalšímu zajímavému úkazu – natrénovaný syntetizér generoval řeč s nepřirozeně rychlým tempem, a to především při syntetizování delších promluv. Po provedení několika dalších tisíc trénovacích

kroků však problém zmizel. Velmi pravděpodobně se jedná o jakýsi problém spojený s trénováním bloku *stochastic duration predictor*, který v rámci modelu VITS predikuje délky jednotlivých fonémů (viz podkapitola 3.4.2).

Hledání příčiny popsaných jevů přesahuje rámec této práce, řešení zmíněných problémů se však nabízí jako předmět dalšího výzkumu.

Závěr 7

Cílem této práce bylo objasnit, jaký vliv má množství trénovacích dat na kvalitu syntetické řeči, a také, jak se na kvalitě syntézy odrazí trénování s využitím předtrénovaných modelů.

V první části práce byly popsány některé významné metody, které lze využít ke generování umělé řeči, přičemž zvláštní pozornost byla věnována moderním metodám syntézy řeči založeným na neurálních modelech. Byly zde rovněž popsány přístupy k hodnocení syntetické řeči, které byly následně použity v praktické části práce.

Praktická část diplomové práce byla věnována experimentům provedeným s end-to-end syntetizérem VITS. V rámci těchto experimentů bylo natrénováno celkem 27 modelů vzájemně se lišících velikostí použitého trénovacího korpusu a také způsobem inicializace parametrů při trénování. Tyto modely byly porovnány pomocí poslechových testů a pomocí objektivní míry MCD a na základě výsledků tohoto porovnání byly následně formulovány zásady, jaké volit množství trénovacích dat, resp. jak inicializovat parametry trénovaného modelu, aby výsledná syntetická řeč dosahovala nejvyšší možné kvality.

Během experimentů se ukázalo, že model VITS trpí problémy s výslovností fonému /r/ a v některých případech též nepřirozeným tempem syntetické řeči. Řešení těchto nedokonalostí bylo předloženo jako námět pro další výzkum.

Bibliografie

1. TIHELKA, Daniel et al. Save Your Voice: Voice Banking and TTS for Anyone. In: *INTERSPEECH 2021*. Brno, Česká republika, 2021, s. 2195–2196.
2. GORDON, J. Ramsay. Mechanical Speech Synthesis in Early Talking Automata. *Acoustics Today*. 2019, roč. 15, č. 2, s. 11–19.
3. BRACKHANE, Fabian. Kempelen vs. Kratzenstein - researchers on speech synthesis in times of change. In: *First International Workshop on the History of Speech Communication Research*. Dresden, Německo, 2015, s. 42–49.
4. YOUNG, Thomas. *A Course of Lectures on Natural Philosophy and the Mechanical Arts*. Sv. 2. J. Johnson, 1807.
5. KEMPELEN, Wolfgang von. *Wolfgangs von Kempelen k.k. wirklichen Hofraths Mechanismus der menschlichen Sprache: nebst der Beschreibung seiner sprechenden Maschine*. Wien : J.V. Degen, 1791.
6. TAN, Xu; QIN, Tao; SOONG, Frank; LIU, Tie-Yan. *A Survey on Neural Speech Synthesis*. 2021. Dostupné z arXiv: 2106.15561 [eess.AS].
7. COKER, Cecil H. A Model of Articulatory Dynamics and Control. *Proceedings of the IEEE*. 1976, roč. 64, s. 452–460.
8. FANT, Gunnar. *Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations*. De Gruyter Mouton, 1971. D A C S R Series. ISBN 9789027916006.
9. TABET, Youcef; BOUGHAZI, Mohamed. Speech synthesis techniques. A survey. In: *International Workshop on Systems, Signal Processing and their Applications, WOSSPA*. Tipaza, Alžírsko, 2011, s. 67–70. Dostupné z DOI: 10.1109/WOSSPA.2011.5931414.
10. PSUTKA, Josef; MÜLLER, Luděk; MATOUŠEK, Jindřich; RADOVÁ, Vlasta. *Mluvíme s počítačem česky*. Praha, Česká republika: Academia, 2006. ISBN 80-200-1309-1.

11. HUNT, Andrew J.; BLACK, Alan W. Unit selection in a concatenative speech synthesis system using a large speech database. In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Atlanta, Spojené státy americké, 1996, sv. 1, s. 373–376. Dostupné z DOI: 10.1109/ICASSP.1996.541110.

12. ZEN, Heiga; TOKUDA, Keiichi; BLACK, Alan W. Statistical parametric speech synthesis. *Speech Communication*. 2009, roč. 51, č. 11, s. 1039–1064. ISSN 0167-6393. Dostupné z DOI: https://doi.org/10.1016/j.specom.2009.04.004.

13. SHEN, Jonathan et al. *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. 2018. Dostupné z arXiv: 1712.05884 [cs.CL].

14. PING, Wei et al. Deep Voice 3: 2000-Speaker Neural Text-to-Speech. In: *International Conference on Learning Representations*. Vancouver, Kanada, 2018. Dostupné také z: https://openreview.net/forum?id=HJtEm4p6Z.

15. LI, Naihan et al. *Neural Speech Synthesis with Transformer Network*. 2019. Dostupné z arXiv: 1809.08895 [cs.CL].

16. REN, Yi et al. FastSpeech: Fast, Robust and Controllable Text to Speech. In: WALLACH, H. et al. (ed.). *Advances in Neural Information Processing Systems*. Vancouver, Kanada: Curran Associates, Inc., 2019, sv. 32.

17. JEONG, Myeonghun; KIM, Hyeongju; CHEON, Sung Jun; CHOI, Byoung Jin; KIM, Nam Soo. Diff-TTS: A Denoising Diffusion Model for Text-to-Speech. In: *INTERSPEECH 2021*. Brno, Česká republika, 2021, s. 3605–3609.

18. MIAO, Chenfeng et al. Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Španělsko, 2020, s. 7209–7213. Dostupné z DOI: 10.1109/ICASSP40776.2020.9054484.

19. OORD, Aaron van den et al. *WaveNet: A Generative Model for Raw Audio*. 2016. Dostupné z arXiv: 1609.03499 [cs.SD].

20. PRENGER, Ryan; VALLE, Rafael; CATANZARO, Bryan. *WaveGlow: A Flow-based Generative Network for Speech Synthesis*. 2018. Dostupné z arXiv: 1811.00002 [cs.SD].

21. KUMAR, Kundan et al. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. In: *Neural Information Processing Systems*. Vancouver, Kanada, 2019.

22. YANG, Geng et al. Multi-Band Melgan: Faster Waveform Generation For High-Quality Text-To-Speech. In: *2021 IEEE Spoken Language Technology Workshop (SLT)*. Shenzhen, Čína, 2021, s. 492–498. Dostupné z DOI: 10.1109/SLT48900.2021.9383551.

23. KONG, Jungil; KIM, Jaehyeon; BAE, Jaekyoung. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Vancouver, Kanada: Curran Associates Inc., 2020. NIPS'20. ISBN 9781713829546.

24. PING, Wei; PENG, Kainan; CHEN, Jitong. ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech. In: *7th International Conference on Learning Representations, ICLR 2019*. New Orleans, Spojené státy americké: OpenReview.net, 2019.

25. KIM, Jaehyeon; KONG, Jungil; SON, Juhee. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In: *ICML 2021*. Wien, Rakousko, 2021.

26. KINGMA, Diederik P.; WELLING, Max. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*. 2019, roč. 12, č. 4, s. 307–392. Dostupné z DOI: 10.1561/2200000056.

27. KOMINEK, John; SCHULTZ, Tanja; BLACK, Alan W. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In: *Proc. Speech Technology for Under-Resourced Languages (SLTU-2008)*. Hanoi, Vietnam, 2008, s. 63–68.

28. CERŇAK, Miloš; RUSKO, Milan. An Evaluation of Synthetic Speech Using the PESQ Measure. In: *Forum Acusticum*. Budapest, Maďarsko, 2005, s. 2725–2728.

29. ZHUANG, Fuzhen et al. *A Comprehensive Survey on Transfer Learning*. 2020. Dostupné z arXiv: 1911.02685 [cs.LG].

30. GRŮBER, Martin; CHÝLEK, Adam; MATOUŠEK, Jindřich. Framework for Conducting Tasks Requiring Human Assessment. In: *INTERSPEECH 2019*. Graz, Rakousko, 2019, s. 4626–4627.

Seznam obrázků

Seznam tabulek