FAKULTA MECHATRONIKY,
INFORMATIKY A MEZIOBOROVÝCH
STUDIÍ TUL

Diplomová práce

Separace řečových signálů pomocí metod strojového učení

Studijní program: N0613A140028 Informační technologie
Autor práce: **Bc. Martin Matoušek**
Vedoucí práce: Ing. Jiří Málek, Ph.D.
Ústav informačních technologií a elektroniky

Liberec 2023

FAKULTA MECHATRONIKY,
INFORMATIKY A MEZIOBOROVÝCH
STUDIÍ TUL

Zadání diplomové práce

Separace řečových signálů pomocí metod strojového učení

Jméno a příjmení:	**Bc. Martin Matoušek**
Osobní číslo:	M21000162
Studijní program:	N0613A140028 Informační technologie
Zadávající katedra:	Ústav informačních technologií a elektroniky
Akademický rok:	2022/2023

Zásady pro vypracování:

1. Úloha separace řečových signálů spočívá v odhadu promluv jednotlivých řečníků z jejich směsi, kde je případně přítomen i šum prostředí. Velká skupina algoritmů, které řeší tento problém, je založena na strojovém učení. Separační model je u těchto metod odvozen z rozsáhlé databáze směsí řečových signálů na principu trénování s učitelem [1]. Na databázích simulovaných směsí pořizených v přibližně fixních akustických podmínkách fungují tyto trénované separační metody velmi dobře. Robustnost modelů při použití na mírně odlišných datech ovšem zůstává důležitou otevřenou otázkou [2]. Robustnost lze sice zvýšit přidáním dalších trénovacích dat, ale není realistické zařadit do trénovací množiny všechny možné kombinace akustických podmínek. Cílem diplomové práce je:
2. Seznámit se s problematikou separace realistických jednokanálových směsí řečových signálů metodami strojového učení
3. Zprovoznit trénování/testování některé moderní metody, např. sítě ConvTasNet [3].
4. Analyzovat robustnost natrénovaného separačního modelu na tzv. off-domain datech, tedy datech, která se silně podobají trénovací množině, ovšem liší se v některém podstatném aspektu. Může se jednat např. o jazyk promluv, tedy síť byla trénována na angličtině ale pokusíme se separovat směs čínských řečníků (viz [2]). Dále se může jednat o přítomnost jiného typu šumu, zvětšenou dobu dozvuku místnosti či časově proměnnou aktivitu separovaných řečníků.
5. Volitelně: zvolit jednu z metod adaptace neuronových sítí a porovnat, jak je adaptace (pomocí menšího množství off-domain dat) efektivní ve srovnání s novým trénováním modelu na rozšířeném trénovacím datasetu.

Rozsah grafických prací:	dle potřeby dokumentace
Rozsah pracovní zprávy:	40-50 stran
Forma zpracování práce:	tištěná/elektronická
Jazyk práce:	Čeština

Seznam odborné literatury:

[1] HERSHEY, John R., Zhuo CHEN, Jonathan LE ROUX a Shinji WATANABE. Deep clustering: Discriminative embeddings for segmentation and separation. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [online]. IEEE, 2016, 2016, s. 31-35 [cit. 2022-09-23]. ISBN 978-1-4799-9988-0. Dostupné z: doi:10.1109/ICASSP.2016.7471631

[2] HAN, Jiangyu, Yanhua LONG, Lukas BURGET a Jan CERNOCKY. DPCCN: Densely-Connected Pyramid Complex Convolutional Network for Robust Speech Separation and Extraction. In: ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [online]. IEEE, 2022, 2022-5-23, s. 7292-7296 [cit. 2022-09-22]. ISBN 978-1-6654-0540-9. Dostupné z: doi:10.1109/ICASSP43922.2022.9747340

[3] LUO, Yi a Nima MESGARANI. Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing [online]. 2019, 27(8), 1256-1266 [cit. 2022-09-23]. ISSN 2329-9290. Dostupné z: doi:10.1109/TASLP.2019.2915167

Vedoucí práce:	Ing. Jiří Málek, Ph.D. Ústav informačních technologií a elektroniky
Datum zadání práce:	24. října 2022
Předpokládaný termín odevzdání:	22. května 2023

L.S.

prof. Ing. Zdeněk Plíva, Ph.D.
děkan

prof. Ing. Ondřej Novák, CSc.
vedoucí ústavu

V Liberci dne 24. října 2022

Prohlášení

Prohlašuji, že svou diplomovou práci jsem vypracoval samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mé diplomové práce a konzultantem.

Jsem si vědom toho, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci nezasahuje do mých autorských práv užitím mé diplomové práce pro vnitřní potřebu Technické univerzity v Liberci.

Užiji-li diplomovou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti Technickou univerzitu v Liberci; v tomto případě má Technická univerzita v Liberci právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Současně čestně prohlašuji, že text elektronické podoby práce vložený do IS/STAG se shoduje s textem tištěné podoby práce.

Beru na vědomí, že má diplomová práce bude zveřejněna Technickou univerzitou v Liberci v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů.

Jsem si vědom následků, které podle zákona o vysokých školách mohou vyplývat z porušení tohoto prohlášení.

11. května 2023 Bc. Martin Matoušek

Poděkování

Děkuji svému vedoucímu diplomové práce Ing. Jiřímu Málkovi, Ph.D. za vlídný a zodpovědný přístup, nápomocné konzultace a také konstruktivní, odborné rady a kritiku. Děkuji také Ing. Lukáši Matějů, Ph.D. za tipy pro práci s MetaCentrem. Dále bych chtěl poděkovat své rodině a přátelům za podporu během studií a v neposlední řadě své přítelkyni, která mi byla během studií i psaní diplomové práce velkou oporou.

Anotace

Tato diplomová práce se zabývá problematikou separace řeči, zkoumá chování moderních separačních sítí na off-domain datech a zabývá se rozšířením trénovací množiny za účelem zprovoznění separace řeči na těchto datech. Nejprve porovnává účinnost různých metod, které řeší úlohu separace řeči na datech s dvěma řečníky. Po porovnání byla pro experimenty vybrána konvoluční síť Conv-TasNet jako poměrně účinná metoda, která má zároveň rychlé trénování a poměrně malou velikost modelu.

Hlavním tématem této práce je zkoumání toho, jak se metoda separace řeči s učitelem chová na off-domain datech. Tento problém může nastat například změnou jazyka mluvčích, změnou dozvuku prostředí nebo počtu řečníků. Z těchto rozsáhlých alternativních možností byla jako hlavní náplň práce zvolena změna jazyka, která byla podrobně zkoumána kvalitativně i kvantitativně. Do menší míry a nad rámec zadání byly zkoumány i experimenty s proměnlivým počtem řečníků.

V rámci změny jazyka řečníků je tato změna dat problém a model trénovaný na angličtině při použití na taiwanském korpusu, který v tomto případě představuje off-domain data, nefunguje. V rámci experimentů pro zprovoznění modelu na různých jazycích, byly modely trénované na rozšířených korpusech kvalitní i na datech obsahujících taiwanštinu. Důležité je ale zmínit, že při přítomnosti různých jazyků ve směsi, je nutné do trénovací sady přidat kromě korpusů v angličtině a taiwanštině i korpus, který je kombinuje. Tento koncept rozšíření datové sady pro zprovoznění modelů na různých jazycích se ukázal jako efektivní.

Částečně bylo zpracováno i téma různého počtu mluvčích. I v rámci testování na datech s různým počtem řečníků se ukázalo, že model trénovaný na dvou řečnících na data s jedním řečníkem není účinný. Rozšíření datové sady o taková data umožnilo modelu, i přes určitá omezení, pracovat s daty s jedním řečníkem. Také ale vyšly najevo nedostatky sítě Conv-TasNet při práci s jiným počtem řečníků.

Klíčová slova

separace řeči, off-domain data, různé jazyky, angličtina, taiwanština, různý počet řečníků, rozšíření datové sady, Conv-TasNet

Annotation

This thesis addresses the problem of speech separation, investigates the behavior of modern separation networks on off-domain data and explores training set extension to allow speech separation on this data. First the effectiveness of different methods that address speech separation on two speaker data is compared. After the comparison, Conv-TasNet was selected for the experiments as a relatively efficient method that has both fast training and relatively small model size.

The main focus of this paper is to determine how supervised speech separation method behaves on off-domain data. From these extensive alternatives, language change was chosen as the main focus of the thesis and was investigated in detail both qualitatively and quantitatively. To a lesser extent and beyond the scope of this thesis, experiments with a variable number of speakers were also investigated.

Changing the language of the speakers poses a problem and the model trained on English does not perform well when used on the Taiwanese corpus, which represents off-domain data in this case. In experiments to create model functioning on different languages, the models trained on the extended corpora were effective even on data containing Taiwanese speakers. However, it is important to mention that when both languages are present in the mixture, it is necessary to add corpus combining both languages in addition to English and Taiwanese corpora to the training set. This concept of extending the dataset of the models for different languages has proven to be effective.

The topic of the different number of speakers was partially researched. While testing on data with different number of speakers, it turned out that the model trained on two speakers is not effective on data with one speaker. Extending the dataset with such data allowed the model to work with single-speaker data, despite some limitations. However, the shortcomings of Conv-TasNet in working with a different number of speakers also became apparent.

Keywords

speech separation, off-domain data, different languages, English, Taiwanese, different speaker count, dataset extension, Conv-TasNet

Obsah

Seznam použitých zkratek

DFT – Diskrétní Fourierova transformace (Discrete Fourier transform)

STFT – Krátkodobá Fourierova transformace (Short-time Fourier transform)

ASA – Auditory scene analysis

CASA – Computational auditory scene analysis

PCA – Analýza hlavních komponent (Principal component analysis)

SDR – Poměr signálu ke zkreslení (Signal-to-Distortion ratio)

SI-SDR – SDR nezávislý na měřítku (Scale-Invariant SDR)

PIT – Permutačně invariantní trénování (Permutation Invariant Training)

TCN – Časová konvoluční síť (Temporal Convolutional Network)

RNN – Rekurentní neurální síť (Recurrent Neural Network)

DPRNN – Duální RNN (Dual-path RNN)

WSJ0 – Wall Street Journal corpus

WHAM! – WSJ0 Hipster Ambient Mixtures

ee – model s datovou sadou eng + eng

et – model s datovou sadou eng + tai

tt – model s datovou sadou tai + tai

Seznam obrázků

Seznam tabulek

1 Úvod

Problematika zpracování řeči je v této době velmi důležitou a zároveň velmi rychle se rozvíjející oblastí.

Rychlý rozvoj je ovlivněn primárně díky rozsáhlému využití v oblasti telefonie a přenosu zvuku přes internet, například při telekonferencích. Další aplikací jsou chytří asistenti či hlasové ovládání, které je využíváno v oblasti IoT, ale i v automobilovém průmyslu. Úlohy z oblasti zpracování řeči se využívají také v inteligentních pomůckách pro nedoslýchavé. Důležitou roli hraje i ve vojenství, kde se využívá například pro zlepšení komunikace v hlučných prostředích, jako například v letadlech či vrtulnících.

Jednou z důležitých úloh zpracování řeči je odstranění nechtěných signálů ze směsí řeči a šumu. Řešení této úlohy je možné rozdělit do tří kategorií:

1. V případě, že nechtěné signály jsou pouze šum, jedná se o úlohu zlepšování řeči. Tato úloha je ve své původní formě nejstarší a také nejjednodušší z těchto kategorií. A to proto, že řeč má jiné vlastnosti, než šum prostředí a na základě těchto vlastností je lze odlišit.
2. Případ, ve kterém je cílem získat všechny řečové zdroje se označuje jako separace řeči. V tomto případě jsou ve směsi kromě nechtěných šumových signálů také nechtěné promluvy konkurenčních řečníků. Úloha je definována jako slepý odhad všech řečových signálů ze směsi signálů, tedy bez znalosti informací o řečnících.
3. Extrakce cílového řečníka označuje úlohu, při které je cílem získat řečový signál jednoho cílového řečníka. Nechtěným signálem je přitom šum i nechtěná řeč dalších přítomných řečníků. Extrakce cílového řečníka může být označena za nejnáročnější úlohu z těchto kategorií. To je způsobeno tím, že oproti separaci je nutné navíc identifikovat správného řečníka mezi ostatními mluvčími.

V posledních letech se v tomto oboru dostaly do popředí metody založené na neurálních sítích, které umožnily posun v dosavadních možnostech zpracování řeči. Cílem této diplomové práce je přiblížit problematiku separace jednokanálových směsí řečových signálů metodami strojového učení. Následně zprovoznit trénování a testování některé z moderních neurálních sítí pro separaci řeči a prozkoumat její funkčnost a případná omezení. V případě této práce jde o síť Conv-TasNet z roku 2019 [1], která je součástí hlavního proudu aktuálních řešení problematiky separace řeči.

Další úlohou této práce je analyzovat, jak reaguje metoda separace řeči s učitelem na takzvaná off-domain data. To jsou testovací data, která se silně podobají trénovací sadě, ale liší se v některém z podstatných aspektů. Existuje celá řada možných variant off-domain dat, například změna jazyka mluvčích, dozvuku místnosti nebo počtu aktivních řečníků. Z těchto rozsáhlých alternativ je pro hlavní část práce zvolena změna jazyka promluv. Jazykem viděným v použité trénovací sadě je angličtina a jako off-domain jazyk slouží taiwanština. Nad rámec zadání je pak v menší míře zkoumána ještě změna počtu současně aktivních řečníků. V práci jsou následně porovnány výsledky testování na datech původní domény s výsledky testování na off-domain datech a natrénovány modely na rozšířených korpusech s cílem zprovoznit separaci řeči na off-domain datech.

2 Zlepšování řeči

V úloze zlepšování řeči (speech enhancement) [2] je cílem odstranit šum ze směsi jednoho řečového signálu a nechtěného šumu. To je podobné i v případě extrakce cílového řečníka, kde ovšem uvažujeme, že mezi nechtěnými signály jsou také jiné řečové signály. V obou případech systém začíná se zkreslenou řečí na vstupu a končí s čistou řečí na výstupu.

Hlavní cílem při návrhu algoritmu pro zlepšení řeči je utlumit šum bez zavedení znatelného zkreslení do řeči signálu. Šum je v úloze zlepšení řeči nechtěnou složkou signálu, která je obecně velmi odlišná od signálu řečníka. Těchto odlišností algoritmus využívá k co nejefektivnějšímu odstranění šumu.

V závislosti na počtu mikrofonů se zlepšování řeči dělí na monofonní s jedním mikrofonem a zlepšování řeči založené na polích (array based) s více mikrofony. Počet mikrofonů může výrazně ovlivnit výkon algoritmu zlepšování řeči. S větším počtem senzorů je možné získat více informace. Tuto dodatečnou informaci pak lze využít pro vylepšení výsledků systému zlepšování řeči. V případě mikrofonních polí to je prostorová informace. Ta umožňuje určit, kde se nachází zdroj signálu na základě zpoždění dozvuků na jednotlivých mikrofonech. Například pokud je alespoň jeden z mikrofonů v blízkosti zdroje hluku, či šumu, je možné jeho signál efektivně odečítat. Původní a běžnější metodou je ale zlepšování řeči ze signálu získaného jedním mikrofonem. Metody zlepšování řeči staví na několika základních principech. Jedním je například odhad vlastností řeči a/nebo šumu přímo ze zkresleného signálu. Na tomto principu je založeno například spektrální odečítání [3]. Druhým principem je nalezení zobrazení, které ze zarušeného signálu odhadne signál nezkreslený. Tohoto využívají metody založené na statistickém modelu [4], nebo i moderní algoritmy založené na datech a strojovém učení [5–7] (například odhlučňovací autoenkodéry).

2.1 Spektrální odečítání

Spektrální odečítání [3] je metoda pro získání řečového signálu v časové oblasti pozorovaného v šumu. Získává se odečtením odhadu průměrného magnitudového spektra šumu od magnitudového spektra signálu s hlukem. Spektrum šumu je odhadováno a aktualizováno z period, kde se nenachází řečový signál, ale pouze šum.

V případě spektrálního odečítání se očekává, že šum je stacionární nebo alespoň více stacionární než řeč. To znamená, že se v průběhu času nemění spektrum šumu, či případně pouze minimálně. Pro získání signálů v časové oblasti je nutné odhadnutou okamžitou hodnotu

magnitudového spektra kombinovat s fází zašuměného signálu a následně transformovat diskrétní Fourierovou transformací do časové oblasti.

Spektrální odečítání je relativně nenáročné na výpočetní složitost. Vzhledem k náhodným variacím šumu, ale může být výsledkem odečítání záporná hodnota magnitudy či výkonového spektra. Magnituda a výkonové spektrum jsou nezáporné veličiny a je nutné záporné hodnoty převést na nezáporné. Tato úprava zkresluje rozložení získaného signálu řeči. Zkreslení se navyšuje s klesajícím poměrem signálu řeči k šumu.

2.2 Statistické modely

Další skupina metod formuluje zlepšování jako statistický odhad neznámých parametrů. Existuje mnoho technik pro získání těchto odhadových funkcí. Jeden případ je lineární Wienerovo filtrování a další metody jsou nelineární [2]. Oba případy jsou popsány níže v této kapitole. Různé metody se liší hlavně v předpokladech o parametrech (např. determinismus či náhodnost) a využitém optimalizačním kritériu.

Statistické modely využívají k získání zlepšeného signálu řeči optimalizaci ztrátových funkcí. Například v případě Wienerova filtrování se pomocí střední kvadratické odchylky hledá optimální hodnota koeficientů diskrétní Fourierovy transformace (DFT) čistého signálu. Tento přístup produkuje lineární odhadové funkce komplexního spektra signálu a je optimální v případě, že koeficienty diskrétní Fourierovy transformace šumu a řeči jsou nezávislé náhodné Gaussovské proměnné.

Další metody se zaměřují na (potenciálně) nelineární odhad magnitudového spektra řeči. Ty produkují nelineární odhadové funkce magnitudy za využití různých statistických modelů a ztrátových funkcí. Tyto nelineární odhadové funkce zpracovávají funkce hustoty pravděpodobnosti koeficientů DFT šumu a řeči a v některých případech využívají negaussovské rozdělení. Odhadové funkce jsou často kombinovány s modifikacemi, které berou v potaz pravděpodobnost přítomnosti řeči.

2.3 Metody hlubokého učení

Konvenční algoritmy zlepšování řeči spoléhají na statistické předpoklady o signálech řeči a šumu. Pomocí těchto předpokladů poté odvozuje pravidla potlačení šumu. Protože model často pouze aproximuje realitu, mohou být tyto metody na reálných datech suboptimální.

Hluboké neuronové sítě zaznamenaly úspěchy v související metodě rozpoznání řeči [8, 9], kterou se rozumí automatický převod mluvené řeči do textové podoby. V návaznosti na tyto úspěchy se projevil zájem aplikovat metody hlubokého učení na úlohu zlepšování řeči [5]. V tomto případě se hluboká neurální síť využívá v regresní úloze k natrénování komplexní transformace ze zašuměné řeči do čisté řeči. Využití přístupu založeném na hlubokém učení má výhodu v tom, že neodhaduje statistické vlastnosti signálu a může fungovat i pro rychle se měnící nestacionární šum.

K trénování nelineární transformace ze zašuměné do čisté řeči využívají metody hlubokého učení synchronní nahrávky čisté a zašuměné řeči. Zašuměná řeč představuje vstupní data neuronové sítě. Čistý signál řeči je tedy referencí pro trénování sítě a zároveň požadovaným výstupem sítě. Ze vstupních dat je vypočítána hodnota krátkodobé Fourierovy transformace, jejíž výsledkem jsou magnitudy a fáze rámců. Výstupem sítě je čisté magnitudové spektrum v časovém rámci. V řadě úloh rozpoznávání řeči se využívá nekauzální [10] (využívající minulé i budoucí rámce spektogramu STFT) a často symetrický (se stejným počtem minulých a budoucích rámců spektogramu STFT) kontext. V některých úlohách, které upřednostňují zpracování v reálném čase před přesností, je možné použít i kauzální kontext [12]. Při využití kauzálního kontextu ale systém nevyužívá budoucí rámce, jejichž využití velmi napomáhá přesnosti, jak při strojovém zlepšování řeči, tak i v samotném lidském vnímání řeči [11]. Cílem je získat vektor příznaků posledního rámce spektogramu kontextového okna. Výhodou využití kauzálního kontextu je možnost zpracování v reálném čase, naopak nevýhodou je chybějící kontext budoucích rámců, který může být užitečný.

3 Separace řeči

Cílem separace řeči [12] je slepý odhad všech řečových signálů ze směsi signálů. Separace řeči je základní úkol v problematice zpracování signálů a má mnoho praktických využití. Ta zahrnují naslouchátka, mobilní telekomunikaci a robustní automatické rozpoznávání řeči a řečníka. Problematika separace řeči je nazývána také „cocktail party problém“ – tento název byl vytvořen Edwardem Colinem Cherry v roce 1953 [13].

Schopnost oddělit řeč od hluku v pozadí je zásadní, neboť signál cílové řeči je v mnoha případech poškozen aditivními zvuky z jiných zdrojů, případně i dozvuky nebo povrchovými odrazy řeči samotné. Častou překážkou v oblasti zpracování řeči je i překrývání řečníků (Speaker overlapping) [14]. Tento problém se běžně vyskytuje ve vícečlenné konverzaci. V situacích, které zahrnují volnou výměnu názorů, diskuzích či debatách se překrytí řečníků objevuje ve více než 20 % času řeči. Přestože lidé dokáží tyto problémy vyřešit velmi jednoduše a automaticky, je velmi složité zkonstruovat automatizovaný systém, který by v tomto zdánlivě jednoduchém úkonu odpovídal lidskému zvukovému ústrojí.

Předpokládaná situace pro úlohu separace řeči [12] je více (*J*) řečníků v místnosti s nebo bez dalšího zdroje hluku. V úloze separace řeči je možné použít jeden či více (*K*) mikrofonů. Signály zaznamenané mikrofony mohou být modelovány jako:

$$y^{(m)}(t) = \sum_{j=1}^{J} a_j^{(m)}(t) * s_j(t) + v^{(m)}(t)$$

(3.1)

kde *t* značí index vzorku, $y^{(m)}(t)$ značí pozorovanou směs signálů na mikrofonu m, $s_j(t)$ značí řečový signál řečníka j, $a_j^{(m)}(t)$ je impulsní odezva místnosti (room impulse response) mezi řečníkem j a mikrofonem m a $v^{(m)}(t)$ značí signál hluku včetně impulsní odezvy místnosti ze zdrojů hluku. Pokud je úloha zaznamenávána jedním mikrofonem, je vynechán index mikrofonu m. Operace * v tomto případě značí konvoluci.

Překrytí řečníků v experimentech [15] výrazně snižuje také úspěšnost systémů pro rozpoznání řeči. Například ve výzvě CHiME-5, která se zabývá problematikou rozpoznání vzdálené řeči, kdy se senzory nachází ve větší vzdálenosti od řečníka (například pro využití pro řečově ovládané asistenty) se i v nejlepším ze soutěžních systému navýšila chybovost z 40 % na datové sadě s nízkým překrytím řečníků na 60% chybovost při vysokém překrytí řečníků [16].

Překrývající se řeč je jednou z největších výzev i pro aktuální moderní diarizační systémy. Diarizace řečníků [17] je úloha určení toho kdo, kdy v dané nahrávce mluvil. To se děje zaznamenáváním důležitých částí nahrávky, jako například přechod mezi samotným šumem a řečí, případně změna řečníka. Poté se diarizační systémy snaží rozdělit zvuková data a shlukovat je podle identity řečníka.

3.1 Pre-neurální přístupy

Problematika řečové separace je pro svou důležitost rozsáhle studována již po několik desetiletí. Mezi první pokusy strojově oddělit řeč patří článek „Computational Auditory Scene Analysis“ (CASA) z roku 1994 [18], který vychází z Bregmanovy publikace „Auditory Scene Analysis“ (ASA) popisující model vnímání směsi zvuků člověkem [19]. Mezi další možné postupy se řadí slepá separace řeči (například Faktorizace nezáporných matic) [20] či faktoriálový skrytý Markovův model [21].

3.1.1 Výpočetní analýza sluchové scény

Výpočetní analýza sluchové scény (Computational auditory scene analysis) [18] je pre-neurální metoda založená na percepčních principech analýzy zvukové scény (Auditory scene analysis). Tento systém se skládá ze čtyř částí.

V první části je modelována sluchová periferie pomocí pásmových filtrů a simulací neuromechanických přenosů vnitřními vláskovými buňkami lidského ucha. V druhé fázi systému je vyjádřena periodicita, frekvence, nástupy a posuny vyjádřeny v sluchových reprezentacích. Těmi jsou reprezentace sluchové mapy založené na organizaci vyšších sluchových drah. Informace ze sluchových map poté slouží ke konstrukci symbolického popisu sluchové scény ve třetí fázi.

V poslední fázi systému je použita vyhledávací strategie, která seskupuje zvukové prvky podle podobnosti jejich frekvencí, časů nástupu a posunu. Po prohledání může být signál syntetizován a vyhodnocen neformálními poslechovými testy.

I přesto, že metody výpočetní analýzy sluchové scény jsou důležitou součástí historie výzkumu akustiky, tak nedosahují takových výsledků jako pozdější metody založené na datech.

3.1.2 Slepá separace řeči

Slepá separace řeči je metoda obnovení neznámých signálů nebo zdrojů signálu ze směsi signálů. Základním předpokladem slepé separace je, že zdrojové signály nejsou známé a nejsou

dostupné žádné dodatečné informace o směsi signálů. Tato neznalost je kompenzována statistickým předpokladem, například že jednotlivé zdroje jsou na sobě nezávislé [22].

Slepé metody jsou obecnější, mají tedy potenciál širšího využití a jsou robustnější vůči rozmanitosti dat. Obecný model je ale jen aproximace reality. Pokud jsou tedy dostupné dostatečně přesné znalosti o situaci, je vždy výhodnější použít model s předchozí informací o směsi signálů.

3.1.3 Faktorizace nezáporných matic

Faktorizace nezáporných matic (Non-negative matrix factorization) [20] je jednou z metod slepé separace. Dlouhou dobu byla velmi populárním způsobem pro separaci řeči a často je používána jako výchozí bod pro pozdější přístupy. Faktorizace nezáporných matic byla původně vytvořena jako metoda pro redukci dimenze dat a je součástí sady postupů využívaných pro rozklad datových matic do dvou matic.

Základním modelem je takzvaná bilineární faktorizace nezáporného vstupu **V** do dvou nezáporných matic **W** a **H**, nebo $\boldsymbol{V} \approx \boldsymbol{WH}$, kde obě faktorové matice mohou být nižšího řádu než matice **V**. Tento model má podobný koncept jako například PCA [23]. Všechny tyto maticové faktorizace se totiž dají vyjádřit stejnou rovnicí.

V případě faktorizace nezáporných matic je unikátní právě podmínka nezápornosti faktorových matic. Vzhledem k těmto požadavkům mohou být natrénované komponenty kombinovány pouze aditivně, a ne se navzájem vylučovat. Tento fakt zajišťuje, že vektory, z kterých se skládají matice **W** a **H,** mohou být interpretovány jako konstruktivní stavební bloky vstupu.

Když je faktorizace nezáporných matic zavedena pro data, která byla vytvořena smícháním nezáporných zdrojů, tak vypočtené komponenty velmi dobře odpovídají původním zdrojům a při procesu dekompozice dokáže oddělit prvky ze všech zdrojů. Vzhledem k tomu, že faktorizace nezáporných matic dokáže operovat bez původní znalosti jednotlivých zdrojů dat, je velmi dobře použitelná v učení bez učitele a problematice slepé separace.

4 Neurální přístupy separace řeči

V poslední době se neurální sítě využívají k širokému spektru úkolů v oblasti strojového učení a v problematice zpracování řeči. Mezi tyto úkoly patří také identifikace a verifikace řečníka. Experimenty ukazují, že v mnoha případech výrazně překonávají původní přístupy. Stejnou tendenci ukazují neurální sítě i při využití v úkolech separace řeči, kde nasazení neurálních sítí postupně vede k výraznému zlepšení výkonu [12].

4.1 Oblasti vstupu a výstupu

Existují různé druhy reprezentace řečových signálů. Ty jsou využity jako vstup i výstup systému pro separaci řeči. Mezi nejběžnější se řadí reprezentace v časové oblasti a frekvenční oblasti [12].

4.1.1 Frekvenční oblast

Frekvenční oblast řeči [12] je logickou volbou pro využití při analýze řečových signálů. To je způsobeno tím, že řeč je tvořena periodickým procesem hlasivkami a zároveň je vnímána sluchovým ústrojím jako frekvence. Dřívější pokusy v separaci řeči i extrakci cílového řečníka používaly krátkodobou Fourierovu transformaci pro reprezentaci směsi řečových signálů. Vstupem neurální sítě je v tom případě magnituda směsi signálů $|Y(n,f)|$, případně logaritmická magnituda $log|Y(n,f)|$. Taková neurální síť v mnoha případech vrací výstupní data ve stejné oblasti jako vstupní data, případně systém vrací masku $\mathrm{M}(n,f)$ (takzvaná T-F maska) [24]. *Y* zde označuje krátkodobou STFT y z rovnice (3.1), *n* jsou časové rámce a *f* je frekvence signálu.

T-F maska je založena na krátkodobé frekvenční reprezentaci signálu. Maska je matice obsahující hodnoty, často od 0 do 1, které určují poměr zachování magnitudy původního signálu. Maska se aplikuje na spektrogram signálu násobením po jednotlivých prvcích matice. U binárních metod se odvození masky chová jako klasifikační problém, tudíž může být diskriminačně trénováno pro přesnost, bez obětování rychlosti systému. Maska je poté využita v procesu syntézy odhadu signálu řeči k filtrování původního signálu na základě výsledků trénování. Syntéza odhadu signálu řeči probíhá provedením inverzní krátkodobé Fourierovy transformace na násobku spektogramu směsi s fází zarušeného signálu a maskou.

V těchto případech neurální síť nepředpovídá fázi výstupu, ale využívá fázi vstupní směsi signálů pro rekonstrukci separované řeči. To není optimálním řešení a ukazuje se, že využití

komplexní reprezentace může v mnoha případech vést k lepším výsledkům než využití pouze magnitudy [1].

4.1.2 Časová oblast

V systémech pro separaci řeči, pracujících s reprezentací řeči v časové oblasti, je první vrstvou kodér, který obsahuje jednu konvoluční vrstvu. Tato konvoluční vrstva přijímá signál v časové oblasti [12] a kóduje jej do reprezentace $S^{(enc)} \in R^{N \times K}$, kde K označuje počet filtrů konvoluce a N je počet výsledných rámců, určený posunem okna v konvoluci.

Taková vrstva může implementovat Fourierovu transformaci a reprezentace v takovém případě odpovídá krátkodobé Fourierově transformaci. V tomto případě jsou ale parametry naučeny při trénování neurální sítě, a tak mohou být optimalizovány pro danou úlohu.

Pro rekonstrukci separovaných signálů řeči se používá dekodér. Ten získává výstupní signál z odhadů zakódované reprezentace a často implementuje transponovaný konvoluční vrstvu s parametry naučenými též při trénování sítě.

Změna reprezentace z frekvenční oblasti na časovou oblast může způsobit celkové vylepšení výkonu systému. To je způsobeno několika vlivy: Síť může využívat a předpovídat informace o fázi, která je v předchozích přístupech (4.1.1) vynechána. Nadále ztrátová funkce na časové oblasti využívaná v časových přístupech je lépe kompatibilní s běžnými metrikami pro hodnocení kvality separace a může tedy vést k lepším výsledkům. A také má na úspěšnost vliv velikost okna v kodéru, která je výrazně menší než pro krátkodobou Fourierovu transformaci.

4.2 Ztrátové funkce

Hodnota ztrátové funkce označuje rozdíl mezi očekávaným výstupem systému a reálným signálem vygenerovaným neurální sítí. Pro úlohy ve frekvenční oblasti se využívají primárně binární křížová entropie (Cross-Entropy loss), ztrátová funkce nula jedna (Zero-One loss) nebo střední kvadratická chyba (Mean squared error). Pro úlohy v časové oblasti je to hlavně poměr signálu ke zkreslení (Signal to Distortion ratio), případně jednodušší poměr signálu k šumu (Signal to Noise ratio).

4.2.1 Zero-One loss

Prvním typem ztrátových funkcí ve frekvenční oblasti jsou klasifikační kritéria pro masky. Velmi často používanou ztrátovou funkcí používanou při úlohách klasifikace je ztrátová funkce

nula jedna (Zero-One loss) [25]. Tato ztrátová funkce se využívá v modelech, kde je cílem rozdělit příklady do dvou, nebo i více tříd (binární a vícetřídová klasifikace). Přiřazuje ztrátovou hodnotu 0 pro případy správné klasifikace a 1 pro nesprávné.

$$L_{01} = \frac{1}{n}\sum_{i=1}^{n} \delta(x_i, \hat{x}_i), kde\ \delta(x_i, \hat{x}_i) = \begin{cases} 0, pokud\ x_i = \hat{x}_i \\ 1, pokud\ x_i \neq \hat{x}_i \end{cases}$$

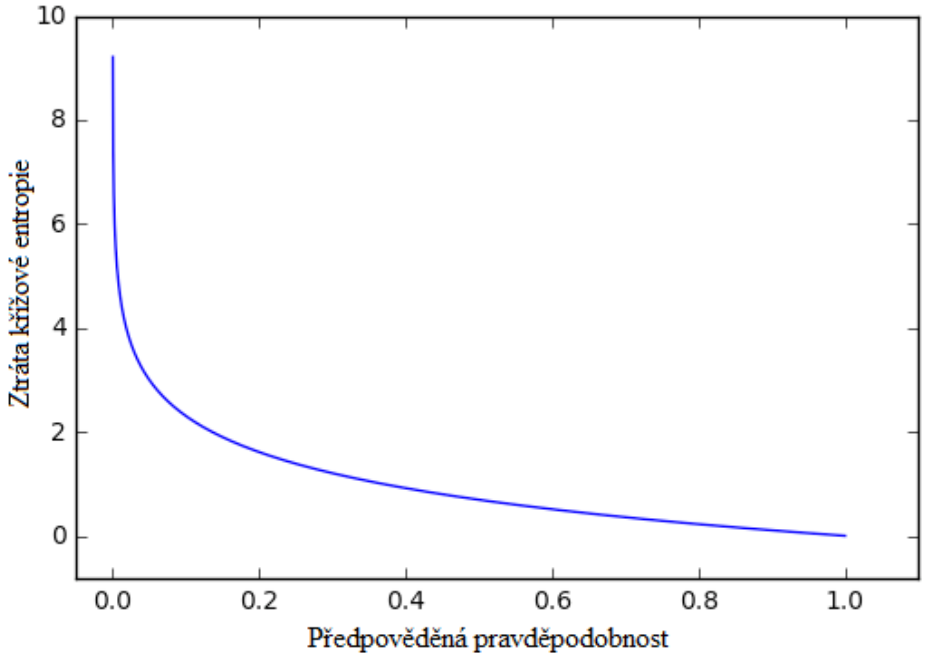
(4.1)

kde x_i značí bin s indexem i očekávané masky a $\hat{x}_i$ je bin s indexem i výstupní odhadnuté masky.

V případech, kdy je klasifikace citlivá na různé druhy chyby, se přiřazuje ztrátová hodnota s různými váhami. To je možné využít pro rozlišení chyb prvního (falešně pozitivní) a druhého typu (falešně negativní) v případech kde je nutné jim přiřadit různé hodnoty. Například hodnota ztráty při nediagnostikování nemocného pacienta má větší váhu než diagnostika nemoci u zdravého člověka. To systému umožňuje upřednostnit omezení či v nějakých případech i vyloučení falešných negativ i na úkor celkové úspěšnosti klasifikační úlohy.

4.2.2 Binární křížová entropie

Ztráta křížové entropie [26] měří úspěšnost klasifikačního modelu, který vrací hodnoty pravděpodobnosti od nuly do jedné. Ztráta křížové entropie narůstá s velikostí odchylky odhadnuté pravděpodobnosti od skutečné třídy. Hodnota třídy je 1, což znamená, že s klesající hodnotou odhadnuté pravděpodobnosti narůstá hodnota ztráty. Teoretický dokonalý model by měl ztrátu rovnou nule. V případě odhadu ideální binární masky se využívá binární křížová entropie. V tom případě poté třída 1 odpovídá řeči a třída 0 odpovídá šumu.



Graf 4.1: Ztráta křížové entropie [31]

4.2.3 Střední kvadratická chyba

Druhou kategorií ztrátových funkcí ve frekvenční oblasti jsou ztrátové funkce počítané přímo z odhadnutého signálu. Tento výpočet probíhá ve frekvenční oblasti a v mnoho případech jde o magnitudu krátkodobé Fourierovy transformace.

Střední kvadratická chyba [25] je průměr kvadratických chyb predikcí na všech prvcích testovací sady. Chyba predikce je v tomto případě rozdíl mezi pravými hodnotami a odhadnutými hodnotami pro daná data.

$$L = \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}$$

(4.2)

kde x_i je opravdová cílová hodnota testu, $\hat{x}_i$ je odhadnutá hodnota a n značí celkový počet testů.

4.2.4 Ztrátové funkce v časové oblasti

Poměr signálu ke zkreslení (SDR) [27] je objektivním měřítkem výkonu pro hodnocení výkonu algoritmů pro zpracování řeči.

V základní formě je SDR definováno pro referenční signál x a odhadnutý signál $\hat{x}$ jako:

$$SDR = 10\log_{10}\left(\frac{||x||^2}{||\hat{x}-x||^2}\right)$$

(4.3)

kde $||x||$ označuje normu (či velikost) vektoru x. Tato forma je závislá na měřítku signálu. Chyba odhadu $\hat{x}-x$ by měla být minimalizována. Hodnota ztrátové funkce SDR je tedy rovna $L_{SDR} = -SDR$.

Varianta SDR invariantní vůči měřítku zpracovaného signálu se označuje SI-SDR (Scale Invariant Signal to Distortion Ratio).

SI-SDR je definováno jako:

$$SI\text{-}SDR = 10\log_{10}\left(\frac{||ax||^2}{||ax-\hat{x}||^2}\right)$$

(4.4)

kde $a = argmin||ax-\hat{x}||^2$ je měřítko signálu. Měřítko vstupního signálu x zajišťuje, že SI-SDR je invariantní vůči měřítku $\hat{x}$, což může být při jeho implementaci výhodné, protože metoda zpracování řeči nezajistí, aby zpracovaný signál měl správné měřítko. Maximalizace ztrátové funkce SI-SDR je ekvivalentní maximalizaci korelace mezi x a $\hat{x}$. Hodnota ztrátové funkce SI-SDR je tedy rovna $L_{SI-SDR} = -SI\text{-}SDR$.

4.3 Rané neurální sítě

Při řešení úloh s více než jedním řečníkem jsou výstupem jednotlivé signály řeči pro každého z řečníků. Pro získání signálu řeči dvou řečníků je možné použít síť s dvěma výstupy. Při trénování této sítě ale můžou nastat dvě různé situace, kde na každý z výstupů může být přiřazen jeden či druhý signál. To samozřejmě ovlivňuje hodnotu ztrátové funkce, pokud je řečník přiřazen na výstup, na kterém je porovnáván se signálem druhého řečníka. Obecně je počet těchto permutací *J!*, kde *J* je počet řečníků a *!* značí faktoriál. Toto se označuje jako permutační problém [28].

V raných přístupech využívajících neurální sítě [12] byla problematika separace řeči často velmi omezena a zjednodušena, právě kvůli permutačnímu problému. Řada experimentů se soustředila pouze na případy, ve kterých se v směsi signálů nachází právě jeden ženský a právě jeden mužský hlas [29–31]. Tato skutečnost poměrně zjednodušuje problematiku separace řeči, protože předchází permutačnímu problému. To je způsobeno odlišnostmi mužského a ženského

hlasu. Výstup byl v tomto případě dělen na jednu část pro ženský hlas a druhou část pro mužský hlas, které lze ve většině případů odlišit jednodušeji než dva hlasy stejného pohlaví.

Dalším omezení této úlohy bylo v některých případech trénování na předem určených dvojicích řečníků [32–34]. Problémem v tomto případě je předpoklad, že trénovací sada bude obsahovat dostatečné množství dat pro každého přítomného řečníka. Dále je velmi obtížné rozšířit tento model pro neznámého řečníka. To je pro využití v realistických situacích neuspokojivé a je tedy využitelné pouze pro omezené experimenty.

I přes tyto limitace je ale možné poznatky a principy z těchto metod využít pro vývoj robustnějších systémů, které mohou posouvat možnosti v této problematice a rozšiřovat použití pro obecnější případy.

4.4 Hluboké shlukování

Hluboké shlukování bylo prvním způsobem, který poměrně úspěšně řešil permutační problém. V úloze hlubokého shlukování [35] probíhá nejprve výpočet příznaků (speaker embeddings) pro každý segment krátkodobé Fourierovy transformace vstupní zarušené řeči. Speaker embeddings označují reprezentaci řečníka pomocí vektoru příznaků konstantní velikosti, nezávisle na délce signálu. Tyto vektory jsou trénovány jako unikátní reprezentace stylu řeči pro každého řečníka, ale neměly by být závislé na obsahu sdělení. Výsledné vektory mají při porovnání stejného řečníka malou vzdálenost a pro různé řečníky jsou naopak vzdálené. Shlukování těchto příznaků realizuje přiřazení každého segmentu k jednomu ze zdrojů. To probíhá právě za využití speaker embeddings. Segmenty jsou podle jejich vzdáleností přiřazeny nejbližšímu shluku. Výsledkem trénování jsou vektory, přiřazené ke všem segmentům signálu, takové že shlukování příznaků vede k extrakci dominantních segmentů každého řečníka. Hluboké shlukování probíhá ve frekvenční oblasti.

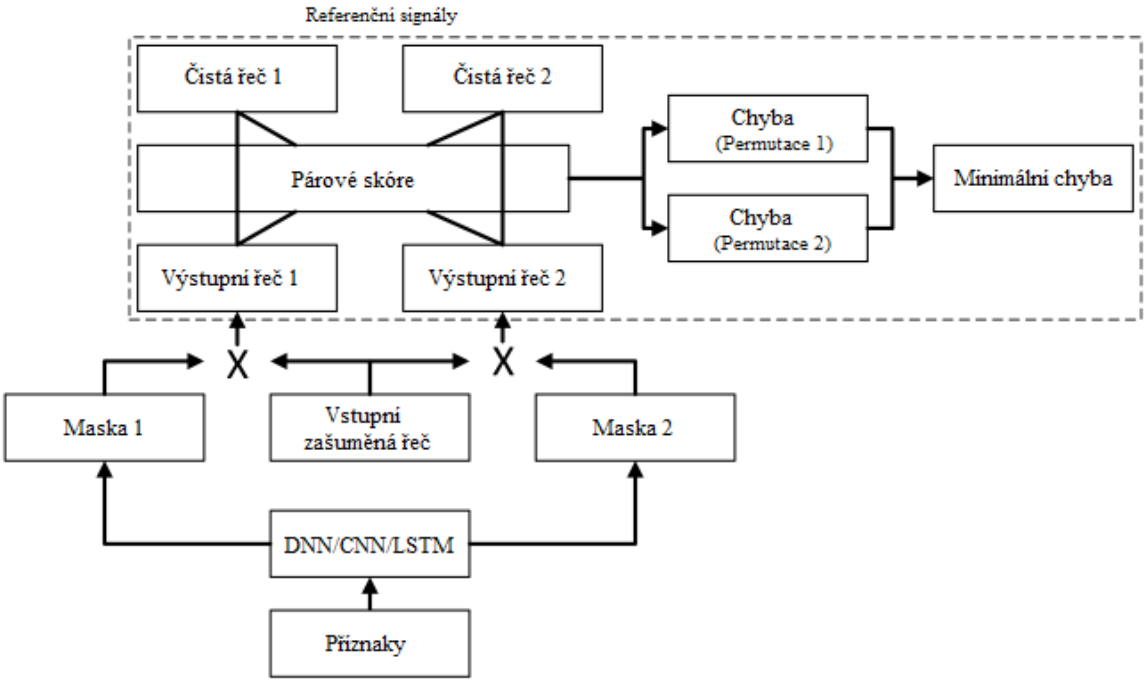
Původní systém hlubokého shlukování byl zamýšlen pouze k získání binárních masek všech zdrojů, přičemž by ponechával získání chybějících příznaků na jiné části systému. Předchozí přístupy v této oblasti používaly ručně navržené příznaky ke shlukování částí zvukového záznamu spektra, patřících ke stejnému zdroji.

4.5 Permutačně invariantní trénování

Permutační problém lze řešit porovnáním ztrátových funkcí jednotlivých výsledků (permutací) při trénování, nalezení správného řešení s nejmenším rozdílem a následné provedení zpětné

propagace se správným výsledkem. Takové trénování se poté nazývá permutačně invariantním trénováním (PIT) [36].

Pro získání správných permutací výstupů je nejprve nutné určit *J!* různých permutací výstupů (*J* je počet řečníků) a následně vypočítat ztrátové funkce pro každou z nich. Jako správný výsledek je následně zvolena ta permutace výstupů s nejnižším součtem hodnot ztrátové funkce u jednotlivých výstupů. Model je následně optimalizován pro snížení právě této hodnoty. Tento postup je ilustrován níže (Obrázek 4.1).



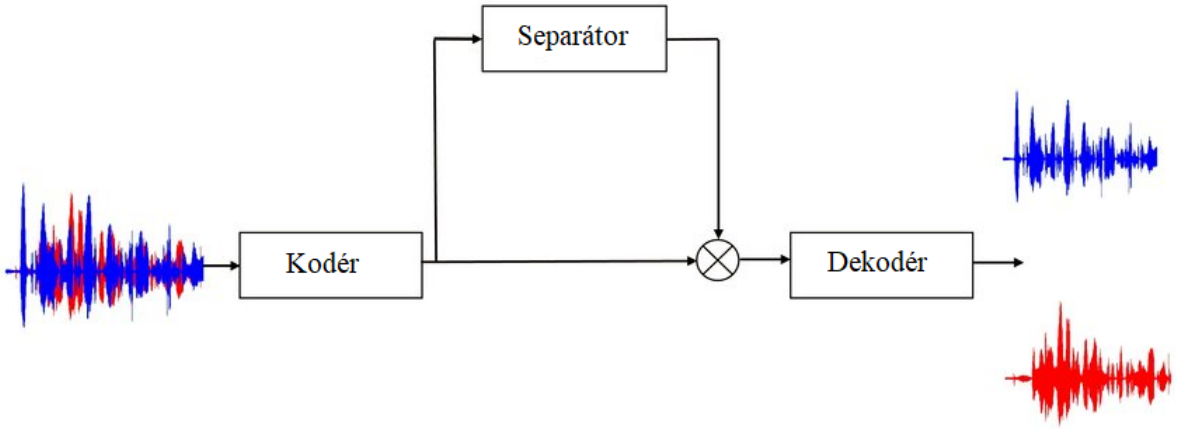
Obrázek 4.1: Permutačně invariantní trénování [36]

4.6 Conv-TasNet

Mnoho předchozích metod formulovalo problém separace za využití reprezentace smíšeného signálu v krátkodobé frekvenční oblasti. To má určité nevýhody, jako například oddělení fáze a magnitudy, protože odhaduje pouze hodnoty magnitudy a využívá zarušené fáze, která ovlivňuje srozumitelnost výsledné řeči. Další nevýhodou je suboptimálnost této reprezentace pro separaci řeči a dlouhá odezva při výpočtu spektogramů.

Konvoluční síť pro oddělení zvuku v časové oblasti (Conv-TasNet) [1] byla navržena k odstranění těchto nedostatků. Conv-TasNet je framework hlubokého učení pro end-to-end separaci řeči v časové oblasti.

Architektura Conv-TasNet se skládá z tří hlavních částí: kodér, separátor a dekodér (viz Obrázek 4.2). Kodér je konvoluční vrstvou, která transformuje signál časové oblasti do vyšší reprezentace. Stejným způsobem na konci dekodér navrací tuto reprezentaci zpět do signálu v časové oblasti pomocí transponované konvoluční vrstvy. Nejdůležitější částí systému je separátor, jenž odhaduje masku, kterou poté aplikuje na kódovanou reprezentaci signálu, aby oddělil jednotlivé řečníky. Separátor se skládá z několika opakovaných průchodů, přičemž při každém dalším průchodu se zvyšuje dilatace sekvence konvolučních bloků časové konvoluční sítě (TCN). To umožňuje síti modelovat dlouhodobé závislosti řečového signálu při zachování malé velikosti modelu.



Obrázek 4.2: Architektura Conv-TasNet [37]

Tato architektura byla vybrána z důvodu předchozích úspěchů časových konvolučních sítí [38]. Zvyšování dilatace v konvolučních filtrech pomáhá zvětšovat velikost kontextu modelem.

Conv-TasNet výrazně překonává předchozí metody časově-frekvenčního maskování v směsích dvou a tří řečníků. Conv-TasNet navíc překonává několik ideálních časově-frekvenčních masek v separaci řeči dvou řečníků hodnocených mírou objektivního zkreslení a subjektivním hodnocením kvality lidským posluchačem. Conv-TasNet má také výrazně menší velikost modelu a kratší minimální latence než předchozí metody T-F maskování, což z něj dělá vhodné řešení pro offline aplikace i aplikace pro separaci řeči v reálném čase.

4.7 Duální rekurentní neuronové sítě

Pokrok v oblasti separace řeči za využití hlubokého učení navýšil zájem o přístupy v časové oblasti. V porovnání se standardními časově-frekvenčními metodami jsou metody časové oblasti navrženy tak, aby najednou modelovaly magnitudu a fázi spektra a umožňovaly přímou optimalizaci vzhledem k časovým diferenciovatelným kritériím.

Tyto metody, včetně například Conv-TasNet, jsou kompromisem mezi délkou kontextu vstupních dat a latencí výstupu neuronové sítě. Delší kontext umožňuje lépe modelovat časové závislosti v řečovém signálu a vede tak k lepším výsledkům, ovšem za cenu zvýšení latence výstupu. V případě, že nezáleží na délce latence, je tedy výhodnější použít delší kontext. To ovšem také obvykle zvětšuje velikost modelu (počet trénovatelných vah). Jedním ze způsobů, který umožňuje využít delší kontext bez významného zvětšení modelu je DPRNN (duální rekurentní neuronová síť).

DPRNN řadí vrstvy RNN tak, aby modelovala dlouhé sekvenční vstupy jednoduchou cestou. Tento způsob spočívá v rozdělení vstupních sekvencí do kratších částí a prokládá dvě RNN, jednu interní pro lokální a jednu externí pro globální modelování.

V DPRNN se nejprve zpracovávají krátké úseky signálu nezávisle na sobě, poté zpracovává souvislosti mezi dílčími bloky. Tímto způsobem je zajištěno, že bloky DPRNN pracují na kratších sekvencích, ale trénují se i dlouhodobé závislosti mezi bloky.

5 Extrakce cílové řečníka

V problematice separace řeči je úkolem systému extrahovat řečový signál všech řečníků, kteří na nahrávce mluví. Tento úkol je v řadě případů řešen neurálními sítěmi. Extrakce cílové řeči [39] staví také na těchto modelech. Oproti separaci řečníka se extrakce řeči zabývá extrakcí jednoho cílového řečníka od hluku v pozadí nebo ze směsi řeči.

Stejně jako u separace, existují k řešení extrakce dva hlavní přístupy: extrakce založená na naměřených datech a na modelech odvozených z obecných statistických předpokladů o signálech.

Extrakce založená na datech [40] se provádí pomocí metod strojového učení s učitelem (supervised learning). Pro natrénování modelu se v tomto případě využívají velké trénovací sady dat. Přístupy založené na datech využívají techniky strojového a hlubokého učení pro odhad extrakčních modelů [41]. Je možné dosáhnout poměrně vysoké kvality extrakce, pokud jsou data dostatečně relevantní k dané problematice.

Metody založené na modelu či učení bez učitele mají jednu velmi důležitou výhodu, a tou je fakt, že nepotřebují žádná trénovací data. Metody založené na modelu využívají pouze obecné statistické předpoklady o vstupních datech a vyžadují pouze minimální informace o dané situaci [41]. Tyto přístupy jsou teoreticky použitelné v širokém rozsahu úloh bez dodatečného přizpůsobení. Tato širší využitelnost je ale dosažena za současného obětování vyšší přesnosti, protože využité statistické modely pouze odhadují reálné podmínky.

Řada studií se snaží dosáhnout extrakce trénováním neurální sítě pouze na datech cílového řečníka [32, 33]. Tím se vytváří model, který je specifický právě na jednoho specifického řečníka. Takové modely mohou být trénovány v závislosti na dvojici řečníků, kde neurální síť má data o cílovém řečníkovi i dalším (rušivém) řečníkovi [32]. Dalším způsobem je model závislý pouze na cílovém řečníkovi. Ten může být generalizován pro jakékoli rušivé řečníky [33]. V obou případech je ale nutné mít dostatečné množství dat cílového řečníka a neumožňují extrakci neznámého řečníka.

Metodou, která umožňuje pracovat s předem neznámým řečníkem je SpeakerBeam [42]. Tato metoda nejprve trénuje model nezávislý na cílovém řečníkovi. Je ho tedy možné trénovat i s několika řečníky. Pro extrakci specifického řečníka následně využívá dodatečné informace o cílovém řečníkovi. Tyto informace jsou získány z referenční nahrávky daného řečníka.

Neurální síť poté využívá tyto informace, aby se zaměřila na cílového řečníka a považuje všechny ostatní signály za rušení.

5.1 Souvislost se separací řeči

Doktorka Kateřina Žmolíková, autorka neurální sítě pro extrakci cílového řečníka SpeakerBeam [42], popisuje ve své disertační práci [12] výhody extrakce cílového řečníka vůči úloze separace řeči a výhody separace řeči takto:

5.1.1 Výhody extrakce cílového řečníka

1. Není nutné počítat řečníky. Klasická neurální síť pro úkoly separace řeči má tolik výstupů, kolik má vstupní signál řečníků. To způsobuje, že architektura systému je závislá na počtu řečníků ve směsi. Extrakce cílové řeči se těmto problémům kompletně vyhýbá, jelikož daná neurální síť má vždy pouze jeden výstup, kterým je zvukový signál cílového řečníka. To způsobuje úplnou nezávislost systému na počtu řečníků ve směsi zvuků.
2. Nemusí řešit permutace výstupů. Při využití neurálních sítí pro separaci řeči má systém jeden výstup pro každého řečníka. To vede k permutačnímu problému. Během tréninku nelze určit v jakém pořadí se vyskytují jednotliví řečníci. Všechny možné permutace musí být správně vyhodnoceny jako korektní výsledek. Extrakce cílového řečníka se tomuto problému zcela vyhýbá, protože využívá dodatečné informace o cílovém řečníkovi.
3. Konzistence výstupu pro delší nahrávky. Modely separace řeči někdy naráží na chyby prohození řečníků během zpracování sekvence. Toto chování je penalizováno při trénování, přesto je ale obtížné udržet konzistentní pořadí řečníků ve výstupu. Tento problém může být častým důvodem chyb systémů separace řeči. Při extrakci cílového řečníka je výstup konzistentní díky využití informací o řečníkovi.

5.1.2 Výhody separace řeči

1. Není potřeba referenční záznam řečníka. Pokud není dostupný předchozí zvukový záznam řečníka a není možné jej získat, je nemožné využít modely pro extrakci řeči. V takovém případě není využití extrakce řečníka definováno, protože závisí na předchozí znalosti informací o jeho řeči. Je tedy nutné využít model pro separaci řeči, který je schopný zvukový signál řečníka získat bez předchozí znalosti daného řečníka.

2. Možnost upravit výběrový modul samostatně. V některých případech může být výhodné explicitně upravit pouze proces výběru řečníka. Například lze využít předtrénovaný model řečníka. Dále je v některých situacích možné určit, zda je řečník ve směsi či ne. Pro takové případy může být výhodné mít možnost explicitně nastavit práh, který podle podobnosti řeči rozhoduje, zda se cílový řečník ve směsi nachází. Při extrakci cílového řečníka je tento práh naučen neurální sítí a není jednoduché ho upravit.
3. Méně výpočtů pro extrakci více řečníků. Pokud je cílem úkolu extrakce všech, či obecně více řečníků ze zvukové směsi, tak je využití modelu řečové separace výrazně výpočetně efektivnější než extrakce cílového řečníka. Pro extrakci cílového řečníka je v takovém případě nutné procházet data jednou pro každého řečníka. V případě separace řeči stačí jeden průchod pro získání signálů pro všechny řečníky.

5.2 Související úlohy

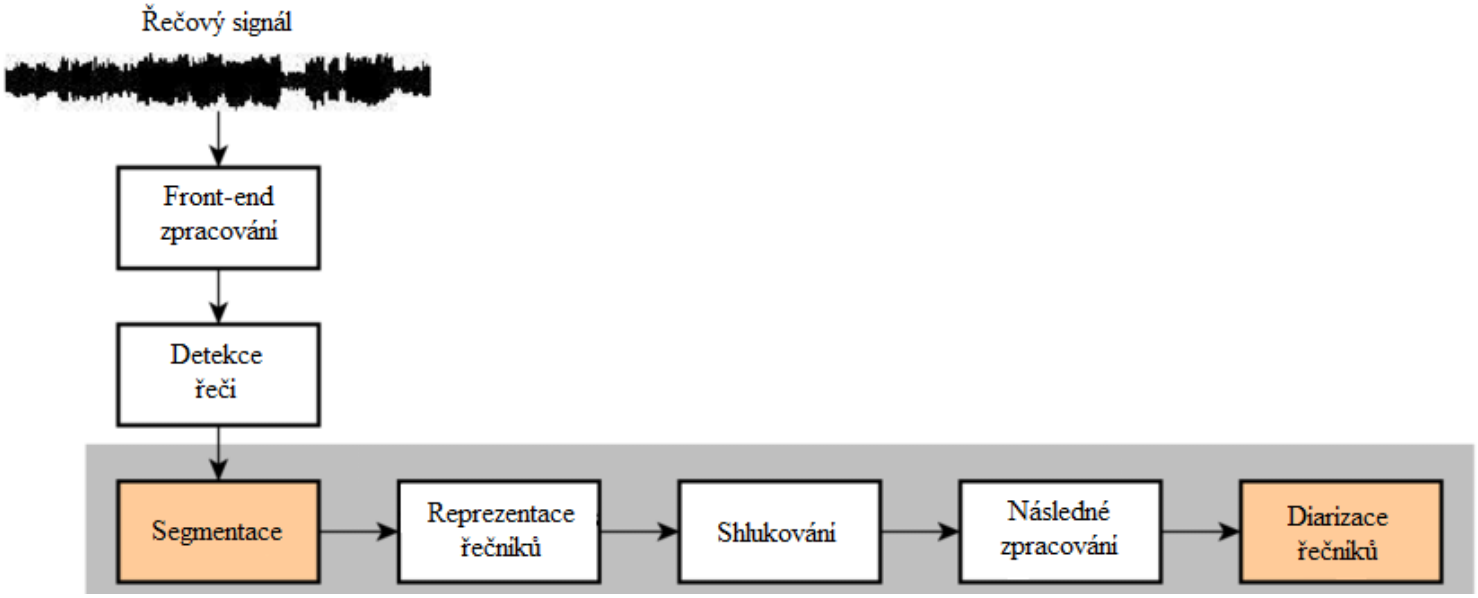
5.2.1 Rozlišení řečníků

Rozlišení řečníků (speaker diarization) [17] je úloha snažící se o rozdělení nahrávky do intervalů odpovídajících jednotlivým řečníkům. Ve zkratce lze popsat jako úlohu určení toho, kdo kdy mluvil. Tento systém by tedy měl být schopný rozpoznat kolik řečníků se nachází v nahrávce a v kterých částech každý z nich mluví. Na první pohled je tato úloha velmi blízká extrakci cílového řečníka. Teoreticky je možné opakovaně použít extrakci a získat tak nahrávky jednotlivých řečníků. To je v praxi ale velmi neefektivní, metoda extrakce cílového řečníka je náročnější než rozlišení řečníků.

Během procesu diarizace je zvuková stopa rozdělena a shlukována do skupin řečových signálů se stejným označením řečníka. Výsledkem je automatická detekce přechodů mezi řečí a „tichem“ a přechodů mezi jednotlivými řečníky. Výhodou je také, že v mnoho případech není nutné mít žádnou předchozí znalost o řečnících, o jejich řeči a o počtu řečníků ve zvukových datech.

Mnoho systémů rozlišení řečníků se skládá z vícero nezávislých modulů (viz Obrázek 5.1). Pro zmírnění rušení v zvukovém signálu se využívají různé techniky front-end zpracování, například zlepšování řeči, dereverbace, separace řeči a podobně. Poté přichází na řadu detekce řeči, pomocí které jsou odděleny části signálu s řečí od ticha (částí, kde se nemluví). Samotné signály řeči jsou poté transformovány do reprezentace akustickými příznaky, které jsou následně ve fázi shlukování rozřazeny do skupin a označeny podle třídy řečníka. V posledních fázích

algoritmu jsou v následném zpracování (post-processing) dále zpřesňovány výsledky shlukování. Každý z těchto modulů může být upravován a optimalizován samostatně, což umožňuje větší kontrolu nad jednotlivými částmi systému.



Obrázek 5.1: Architektura systému diarizace řečníků

5.2.2 Adaptace řečníka

V problematice rozpoznání řeči je adaptace řečníka [43] technika, při které se systém rozpoznání řeči adaptuje na akustické příznaky specifického řečníka za použití krátkých ukázek řeči daného řečníka. Přesněji se jedná o postup pro úpravu parametrů rozpoznávacího modelu systému pro rozpoznávání řeči. Tato úprava probíhá na základě informací ze vzorků řeči určitého řečníka.

Základní princip je tedy velmi podobný úkolu extrakce cílového řečníka. Upravuje dopředný průchod neurální sítí pomocí referenčních nahrávek řečníka. Mnoho metod extrakce cílového řečníka vychází, případně se inspiruje z problematiky adaptace řečníka. Nicméně je zde několik klíčových rozdílů, které je nutné zmínit. Při adaptaci řečníka je cílem lehce pozměnit výstup neurální sítě změnou informací o řečníkovi, přičemž model může bez obtíží fungovat i bez adaptace pro specifického řečníka. Naopak při extrakci cílového řečníka jsou informace o řečníkovi základním prvkem a jejich modifikace by měla v určitých případech významně změnit výstup systému.

6 Faktory ovlivňující výkon separace řeči

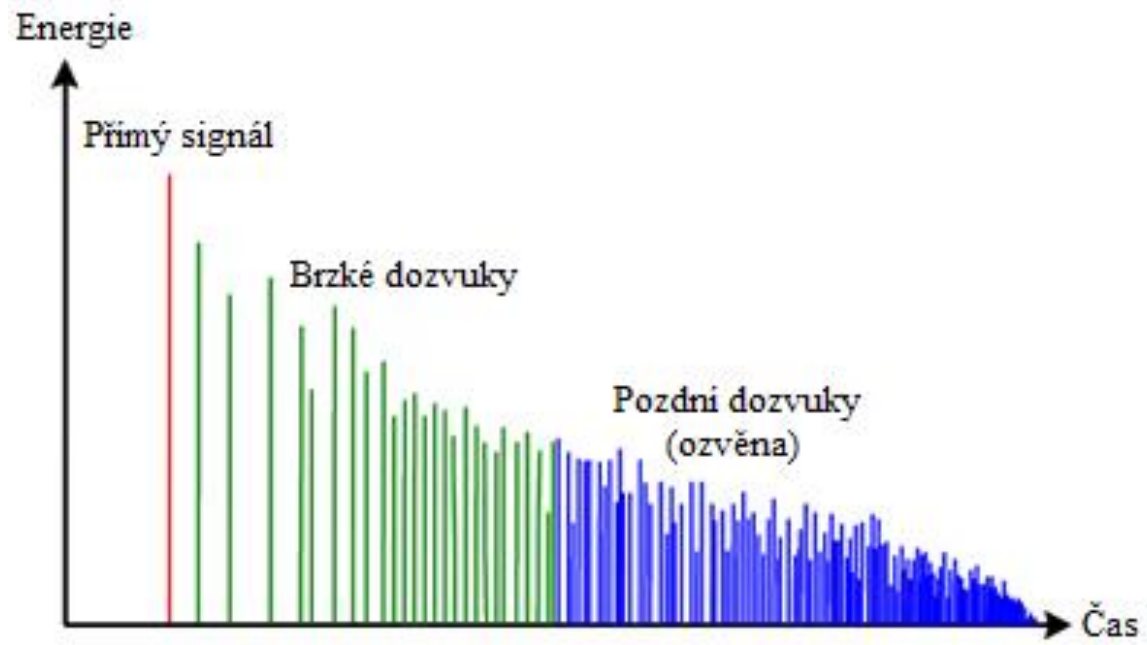
6.1 Šum

Šum, nebo hluk v pozadí [2] je jakýkoli zvukový signál, který je odlišný než monitorovaný signál. V případě úlohy separace řeči jsou to kromě zvuků okolí i nechtěné řečové signály. Zvuk je definován jako vibrace v médiu – často ve vzduchu. Šumem může být jakákoli nechtěná složka signálu – například zvuky okolí, elektrické rušení, případně srozumitelná i nesrozumitelná mluva v pozadí. Je vyjádřen pomocí frekvence, intenzity, periodicity a trvání. Frekvence zvuku je vyjádřena jako počet cyklů za sekundu – hertz [Hz], přičemž 1 hertz znázorňuje 1 cyklus za sekundu [s^{-1}]. Lidské sluchové ústrojí obecně dokáže vnímat frekvence mezi 20 až 20 000 Hz, ale je nejcitlivější na zvuky mezi 500 a 3 000 Hz.
Hlasitost zvuku je měřena v pascalech [Pa] nebo decibelech [dB]. Pascal udává, jak velká síla v newtonech působí na jednotkovou plochu – $N \cdot m^{-2}$ nebo také $kg \cdot m^{-1} \cdot s^{-2}$. Decibel je jednotka užívaná pro měření hladiny intenzity zvuku. Decibel je fyzikálně bezrozměrná logaritmická jednotka, kde navýšení o 3 dB značí dvojnásobný výkon, a naopak snížení o 3 dB značí poloviční výkon. Vnímání hlasitosti zvuku záleží částečně i na jeho frekvenci. Například pro vyrovnání vnímané hlasitosti zvuku o frekvenci 1000 Hz a hlasitosti 40 dB je zapotřebí přes 60 dB o frekvenci 10 000 Hz.

Při separace může šum způsobovat různé komplikace. Šum může například překrývat části řeči a tím v rámci separačního algoritmu ztížit rozlišení jednotlivých řečových signálů. Dále může šum v závislosti na hlasitosti a typu obecně zhoršit kvalitu vstupního signál. V některých případech může šum mít podobnou charakteristiku jako řeč, a to v případech kdy je šumem nechtěný řečový signál. V takových případech může být náročné určit která část signálu je řeč a co je pouze nechtěný šum.

6.2 Dozvuk

V reálném prostředí je zvukový signál v řadě případů poškozen dozvukem z prostorových odrazů od stěn a objektů v místnosti. Dozvuk v signálu je při jeho zpracování náročnou výzvou, zejména v kombinaci s hlukem v pozadí. Dozvuk (reverberace) [44] označuje proces více-směrové propagace zvuku ze zdroje do mikrofonu. Signál získaný mikrofonem se poté skládá z přímého zvuku, odrazů, které dorazí brzy po přímém zvuku (brzké dozvuky) a odrazů, které doráží po brzkých dozvucích (pozdní dozvuky).



Obrázek 6.1: Okamžitá energie impulsní odezvy místnosti [45]

Kombinace přímého zvuku a brzkých dozvuků se označuje za brzkou komponentu řeči. Brzké dozvuky primárně dokreslují spektrální zabarvení signálu a brzké dokonce napomáhají zlepšovat porozumění řeči. Pozdní dozvuky ale pozměňují signál ve formě postupně více utlumených ozvěn. To vede k tomu, že mikrofon přijímá několik kopií původní řeči s různým zpožděním a utlumením. Tento jev vede k zhoršení porozumění řeči a snižuje tak efektivitu algoritmů pro separaci řeči. V případě posluchače se zvuk zdá vzdálenější a s ozvěnou.

6.3 Charakteristiky hlasu

Zatímco výkon průměrného systému pro separaci směsi dvou řečníků v poslední době významně vzrostl, lze ve specifických referenčních směsích stále poznat velkou variaci výsledků. Studie v této problematice [46] odhaluje poměrně významnou mezeru mezi směsmi stejného pohlaví a směsmi rozdílného pohlaví. Pro dvojici stejného pohlaví tedy systémy dosahují podprůměrných výsledků. Vzhledem k tomu, že právě toto rozložení řečníků je v reálných situacích velmi časté, nelze systémy řečové separace optimalizovat pouze pro průměrnou výkonnost systému. Je třeba brát v potaz tyto negativní krajní případy a porozumět, které parametry rozdílných řečníků jsou v takových situacích dominantní pro predikci systému.

Těmito parametry jsou [46]:

Délka hlasového ústrojí – délka hlasivkové trubice, která začíná u hlasivek a končí u úst. Tato hodnota je pro konkrétního řečníka konstantní a nelze ji změnit. Lze ji odhadnout z řečového signálu.

- Základní frekvence řeči – časově proměnlivá hodnota, kterou lze měřit během znělých rámců. Mění se podle aktuální intonace, Lombardova efektu, stavu řečníka a lze ji vlastní vůlí měnit. Lombardův efekt je jev, při kterém řečník nevědomě mění způsob svého mluvení, aby kompenzoval okolní hluky a bylo řeči lépe rozumět.

Výše zmíněná studie ukazuje, že rozdíly základních frekvencí řeči dvou řečníků ve směsi mají důležitý vliv na úspěšnost řečové separace. Naopak rozdíl v délce hlasového ústrojí úspěšnost neovlivňuje.

6.4 Neshoda oblastí

Řada prací v oblasti separace řeči je založena na metodách založených na datech. V těchto případech se nejprve využívá trénovací datová sada pro trénování parametrů systému, který se poté využívá v testovací fázi. Pokud je tedy distribuce trénovacích a testovacích dat nesouměrná, velmi klesá úspěšnost systému.

Z tohoto důvodu lze systémy pro separaci řeči pouze těžko zobecnit pro různé nahrávky z jiných reálných zdrojů [47]. V této době se mnoho systémů pro separaci řeči a extrakci cílového řečníka trénuje na uměle vytvořených směsích zvukových dat. Důvodem je náročnost nahrání tak velkého množství dat s překrývajícími se řečníky s paralelními referencemi jednoho řečníka. Tyto uměle vytvořené podmínky nemusí splňovat různé specifické charakteristiky řeči, které se vyskytují v reálných situacích, například Lombardův efekt, přirozené překrývání řečníků nebo realistická odezva místnosti a hluk v pozadí.

7 Realizace

7.1 Použité knihovny

7.1.1 PyTorch

Pytorch [48] je knihovna strojového učení, která kombinuje snadnou použitelnost a vysokou rychlost. Sdružuje nativní styl programování Pythonu, který umožňuje psát jednoduše čitelný kód, usnadňuje debugging a je konzistentní s dalšími populárními vědeckými knihovnami. Zároveň ale zůstává efektivní a podporuje hardware akcelerátory, jako například GPU.

Pytorch umožňuje okamžité provedení dynamických výpočtů nad tensory s automatickou diferenciací a akcelerací pomocí GPU, a zároveň zachovává výkon, který je srovnatelný s nejrychlejšími knihovnami pro hluboké učení neuronových sítí.

Každý aspekt knihovny PyTorch je běžným Python programem, který má uživatel plně pod kontrolou.

7.1.2 PyTorch lightning

PyTorch Lightning [49] je framework rozšiřující knihovnu PyTorch, který odděluje vědecké komponenty a technické detaily z kódu napsaném pro knihovnu PyTorch a efektivně tím zjednodušuje jeho implementaci. Zároveň také rozšiřuje možnosti využití – poskytuje například výkonnostní a bottleneck profiler, model checkpointing, logování, zobrazení metrik, vizualizaci, předčasné zastavení trénování a podobně. PyTorch Lightning umožňuje implementacím učících algoritmů být nezávislé na hardwaru a také dostupnější pro výzkumníky s horšími výpočetními prostředky. Zprostředkuje totiž běh stejného kódu na libovolném hardwaru. Kód je oproti knihovně PyTorch také čitelnější, jelikož detailní technický kód je abstrahován v souborech knihovny.

Pytorch Lightning také umožňuje využití přesně stejného rozdělení datových sad, jemného ladění parametrů, kritérií pro brzké zastavení trénování a stejné zpracování dat pro zajištění konzistentních výsledků napříč různými systémy.

7.1.3 Asteroid

Asteroid [50] je sada nástrojů pro separaci zvukových zdrojů založená na knihovně PyTorch. Inspirací pro vytvoření této knihovny byly nejúspěšnější neurální systémy pro separaci zdrojů

zvuku. To znamená, že Asteroid obsahuje všechny neurální stavební bloky, které jsou potřebné pro vybudování systému pro separaci zdrojů.

Asteroid také obsahuje takzvané recepty na běžných datových sadách pro separaci zvukových zdrojů. Těmi jsou předem připravené návody a ukázky implementací separačních algoritmů ve formátu Jupyter Notebook. Tyto recepty je možné vyzkoušet ve službě Google Colab, či například v lokální Python distribuci.

Asteroid je založen na hlubokém učení separace zvukových zdrojů a augmentaci řeči. Velmi důležitou předností této sady nástrojů je uživatelská přívětivost pro vývojáře a výzkumníky. Kód je jednoduše rozšiřitelný, což umožňuje snadné experimentování. Také je reproduktovatelný na různých zařízeních.

Asteroid preferuje co nejmenší abstrakci kódu, což znamená, že využívá co nejvíce nativního kódu Pytorch Lightning. Umožňuje importování kódu třetí strany s minimálními změnami. Asteroid poskytuje kompletní sadu nástrojů pro vytvoření systému – od přípravy dat až po evaluaci výsledků.

7.2 Metacentrum

Virtuální organizace MetaCentrum [51] (MetaVO) je tzv. "catch-all" virtuální organizace sdružující všechny uživatele registrované v MetaCentru. Je otevřená všem akademickým pracovníkům, zaměstnancům a studentům vědeckovýzkumných institucí v České republice. Po registraci do MetaCentra mají uživatelé možnost bezplatně využívat výpočetní a úložné kapacity.

Metacentrum nabízí zdroje pro takzvané gridové počítání. Grid označuje spojenou síť počítačů na různých místech. Uživatelské výpočetní úlohy čekají ve frontách na přiřazení zdrojů, o které uživatel zažádal. Přiřazování zdrojů probíhá v rámci plánovacího systému, který na základě specifických požadavků, priorit uživatele, náročnosti úlohy a dalších podmínek určuje pořadí úloh a přidělení specifického stroje. Struktura se dělí na takzvané čelní uzly, ve kterých uživatelé připravují a posílají své úlohy a na výpočetní uzly, ve kterých probíhá samotné počítání.

Pro komunikaci s čelními uzly se využívá protokol SSH a pro přenos souborů protokol SCP (v případě Windows nástroj PuTTY, případně WinSCP). Úlohy v Metacentru se dělí na interaktivní a dávkové. Interaktivní úlohy slouží k manuální práci s terminálem. To je vhodné například pro nastavení prostředí, úvodní stažení závislostí nebo odladění skriptu. Naopak dávkové

úlohy slouží ke spouštění předem připravených odladěných skriptů. Skript po spuštění proběhne kompletně samostatně.

Při provádění náročnějších úloh je vhodné využít možnost žádat o specifické zdroje. Mezi základní parametry patří nastavení minimální velikosti operační paměti, velikost a typ disku pro dočasné soubory (v rámci MetaCentra označováno jako *Scratch*), počet a rychlost CPU. Dále pro strojové učení velmi důležitý počet GPU, paměť GPU a výpočetní možnosti grafické karty pro CUDA (technologie umožňující počítat úlohy na grafické kartě). Dále je potřeba nastavit maximální délku trvání úlohy, která je pro krátké úlohy omezena na 24 hodin a u dlouhých úloh na 120 hodin. Toto rozlišení je nutné kvůli oddělení front pro kratší a delší úlohy – u těch se očekává delší čekání na výpočetní zdroje.

7.3 Datové sady

7.3.1 WSJ0

Tvorba datové sady CSR-I [52], do které spadá i WSJ0 započala v roce 1991 iniciativou agentury ministerstva obrany pro pokročilé výzkumné projekty (DARPA). Jejím cílem bylo vytvořit novou knihovnu dat pro podporu výzkumu systému pro kontinuální rozpoznávání řeči s obsáhlým slovníkem.

Mezi první dva vytvořené korpusy patří právě WSJ0 a jeho následovník WSJ1. Skládají se primárně z četby textů vybraných ze strojově čitelných novinových článků z Wall Street Journal. V dalších rozšířeních korpusu CSR-I se nachází hlavně severoamerické obchodní noviny a také další zpravodajská média.

Texty vybrané ke čtení byly zvoleny z 5 000 nebo 20 000 slovní podmnožiny textového korpusu WSJ. Nadále obsahuje i texty hypotetických novinových článků diktované novináři.

Pro nahrání řeči byly použity 2 mikrofony – blízký mikrofon Sennheiser HMD414 a také sekundární mikrofon, který byl při různých nahráváních rozdílný. Tato datová sada je tedy dostupná v třech konfiguracích: řeč z mikrofonu Sennheiser, řeč z druhého mikrofonu a kombinace řeči z obou. Všechny tyto sady obsahují přepis řeči, testy, dokumentaci apod.

7.3.2 WHAM!

Datová sada WHAM! [53] (WSJ0 Hipster Ambient Mixtures) slučuje všechny nahrávky z korpusu wsj0-2mix s unikátním hlukem v pozadí z korpusu WHAM noises.

Korpus wsj0-2mix obsahuje kombinace dvou řečníků z datové sady WSJ0, a to v minimální verzi, kde jsou delší signály zkráceny a maximální verzi, kde je ke kratším signálům připojeno ticho. Obě verze jsou dostupné s vzorkovací frekvencí 16 kHz a 8 kHz. Existuje také verze s třemi řečníky.

Audio pro hluk v pozadí (WHAM noises) bylo nasbíráno v různých městských lokacích v oblasti San Francisco Bay Area v roce 2018. Prostředí se skládá především z kaváren, restaurací, barů a parků. K nahrání hluků byl použit binaurální mikrofon Apogee Sennheiser na stativu ve výšce 1 až 1,5 metru od země. Datová sada šumu byla zpracována tak, aby byly odstraněny všechny části signálu obsahující srozumitelnou řeč.

Tato datová sada po namixování wsj0-2mix a WHAM noises obsahuje 20 000 nahrávek (30 hodin) v trénovací sadě, 5 000 nahrávek (10 hodin) ve validační sadě a 3000 nahrávek (5 hodin) v testovací sadě. Trénovací a validační sady obsahují stejné řečníky, ale řečníci v testovací sadě jsou odlišní.

Rozšířením této datové sady je korpus WHAMR!, ve kterém je kromě hluků na pozadí k řečovým signálům přidáván také uměle generovaný dozvuk řeči.

8 Modely

V konfiguraci clean je pro trénování použita datová sada dvou řečníků bez šumu (wsj0-2mix). Trénování modelu skončilo po 64 hodinách a 199 epochách, při kterých dosáhlo hodnoty SI-SDR na validační sadě 16,91 dB.

S takto natrénovaným modelem bylo na testovací sadě provedeno několik experimentů. Prvním experimentem bylo testování modelu na shodných testovacích datech, tedy směs dvou řečníků bez šumu a následné porovnání výsledků s dřívějšími ale i s modernějšími metodami. Následující tabulka obsahuje porovnání hodnot ztrátových funkcí SI-SDRi na testovací sadě korpusu wsj0-2mix. Výsledky těchto metod jsou čerpány z článků, které se zabývají tématem separace řeči a též své metody testovali na testovací sadě korpusu wsj0-2mix.

Model	SI-SDRi	Velikost modelu
Hluboké shlukování [54]	10,8 dB	
TasNet [55]	10,8 dB	
Chimera++ [56]	11,5 dB	
TasNet v2 [57]	13,2 dB	
Conv-TasNet	16,3 dB	5,1 M
Deep CASA [58]	17,7 dB	12,8 M
DPRNN [59]	18,8 dB	2,6 M
DPTNET [60]	20,2 dB	2,7 M
Wavesplit [61]	21,0 dB	29 M
Wavesplit v2 [61]	22,2 dB	29 M
Separate and Diffuse [62]	23,9 dB	

Tabulka 8.1: Porovnání modelu Conv-TasNet trénovaného na korpusu wsj0-2mix s dalšími metodami

Při separaci řeči na zašuměných datech může šum způsobovat překrytí signálu, či zhoršení kvality řečového signálu. To může vést k navýšení složitosti separačního algoritmu a horší kvalitě separovaných řečových signálů. Zároveň ale vyšší náročnost vstupních dat umožňuje natrénovat robustnější model, který může mít v určitých situacích lepší výsledky než model trénovaný bez šumu.

Pro variantu noisy byla při trénování využita datová sada dvou řečníků s přidaným šumem z korpusu WHAM noises (WHAM!). Trénování v tomto případě skončilo již po 46 hodinách a 140 epochách a dosáhlo hodnoty SI-SDR na validační sadě 8,81 dB.

I v tomto případě bylo prvním experimentem testování modelu na shodných testovacích datech, tedy směs dvou řečníků s přidaným šumem z korpusu WHAM noises. Výsledky byly následně porovnány s dalšími metodami. I zde jsou hodnoty ztrátové funkce SI-SDR ostatních metod čerpány z článků, ve kterých jsou metody testovány na testovací sadě korpusu WHAM!.

Model	SI-SDRi	Velikost modelu
Chimera++ [56]	9,9 dB	
BLSTM [63]	12,0 dB	23,6 M
Conv-TasNet	13,2 dB	5,1 M
Filterbank [64]	12,9 dB	
TDANet [65]	14,8 dB	2,3 M
Gated DualPathRNN [66]	15,2 dB	2,6 M
TDANet Large [65]	15,2 dB	2,3 M
Wavesplit [61]	15,4 dB	29 M
Wavesplit v2 [61]	16,0 dB	29 M

Tabulka 8.2: Porovnání modelu Conv-TasNet trénovaného na korpusu WHAM! s dalšími metodami

I přes to, že některé z metod dosahují pro korpusy wsj0-2mix i WHAM! lepších výsledků, je důležité poznamenat, že Conv-TasNet je poměrně jednoduchý a má menší model než většina z výkonnějších modelů. Zároveň je možné ho natrénovat poměrně rychle. Je tedy velmi dobrou volbou pro účely této práce, kdy je rychlé trénování modelů výhodou pro provádění více experimentů. Výhodou je v této oblasti také uživatelská přívětivost při případné úpravě modelu, korpusu i samotných přípravách experimentů.

9 Experimenty na rozdílných jazycích

Modely jsou rozděleny na typ *clean*, ve kterém jsou ve směsi pouze signály řeči a typ *noisy*, kde je ve směsi navíc šum. V rámci všech experimentů byla využita verze s vzorkovací frekvencí 8 kHz v minimální verzi (viz 7.3.2). Modely byly trénovány (viz 4.6) na gridovém výpočetním systému v rámci MetaCentra (viz 7.2). Byla využita architektura Conv-TasNet. Pro trénování všech modelů byla použita ztrátová funkce SI-SDR. Jako optimalizační funkce byl použit algoritmus Adam [67] a jako aktivační funkce ReLU [68]. Tabulka níže obsahuje použité hyperparametry modelů.

Počet filtrů	512
Počet kanálů v bottlenecku	128
Počet kanálů v konvolučních blocích	512
Velikost jádra	16
Počet konvolučních bloků v každém opakování	8
Počet opakování	3

Tabulka 9.1: Hyperparametry modelů

Trénování bylo pro všechny modely spuštěno na 200 epoch s nastavením pro dřívější zastavení po 30 epochách s téměř nevýznamným snížením hodnoty ztrátové funkce (0,01 dB) na validačních sadě.

Pro experimenty na rozdílném jazyku řeči zůstávají základní testované modely a také konfigurace modelů stejné. Je to tedy model clean trénovaný na datové sadě wsj0-2mix a model noisy trénovaný na datové sadě WHAM!. Výsledky testování se porovnávají pomocí hodnoty SI-SDRi neboli zlepšení ztrátové funkce SI-SDR. Tato metrika značí rozdíl hodnoty ztrátové funkce SI-SDR odhadovaného vygenerovaného signálu a hodnoty ztrátové funkce SI-SDR vstupní zarušené směsi signálů. Tato hodnota vypovídá o zlepšení kvality, a ne pouze o výsledné hodnotě ztrátové funkce, a proto může lépe vypovídat o výkonnosti natrénovaného modelu.

Jako druhý jazyk byla vybrána taiwanština. A to z důvodu, že je velmi odlišná od angličtiny, na které jsou původní modely trénované a také kvůli dostupnosti korpusu TAT [69] (Taiwanese Across Taiwan). Pro účely této práce byl použit korpus TAT Vol2. Vzhledem k tomu že nebyly nalezeny mixovací algoritmy korespondující s korpusem TAT Vol2, tak musely být korpusy

dvou taiwanských řečníků (tt) a kombinace anglického a taiwanského řečníka (et) vytvořeny v rámci vypracování této práce, a to vlastní kombinací řečníků podle logiky z korpusů wsj0-2mix a WHAM!. Ty byly vytvořeny ve verzi clean (bez šumu) i noisy (s přidaným šumem). Byly použity nahrávky taiwanštiny z korpusu TAT, které jsou v kombinované směsi angličtiny a taiwanštiny doplněny o řečníky z korpusu wsj0. Ve verzi noisy jsou přidány nahrávky šumu z korpusu WHAM noises.

9.1 Model clean na rozdílných jazycích

Pro testování funkčnosti na různých datech byly použity oba vytvořené korpusy ve verzi clean. Prvním je kombinace anglické řeči a taiwanské řeči (eng + tai) a druhý korpus tvoří směs dvou taiwanských řečníků (tai + tai). Pro srovnání je zde uveden i původní korpus WSJ0 (eng + eng).

Podle hodnot celkové ztrátové funkce SI-SDRi (viz Tabulka 9.2) je zde jasně vidět, že rozdílný jazyk oproti trénovacímu korpusu znamená pro model významný problém. Při testování na kombinaci anglického a taiwanského řečníka dosahuje 7,7 dB a na směsi dvou taiwanských řečníků pouze 4,1 dB. I v rámci subjektivního hodnocení autora jsou v mnoha případech výstupní data poškozená. To se často projevuje kombinací obou řečníků na výstupech, poškozením řeči, chybějícími částmi řeči a obecně velmi špatnou srozumitelností řeči.

Testovací data	SI-SDRi
eng + eng	16,3 dB
tai + tai	4,1 dB
eng + tai	7,7 dB

Tabulka 9.2: Testování modelu Conv-TasNet trénovaného na korpusu WSJ0 na rozdílných jazycích

9.1.1 Model clean trénovaný na taiwanské řeči

V rámci stanovení výchozích hodnot pro jednotlivé testy byly natrénován model pouze na korpusu směsí dvou taiwansky mluvících řečníků. Takto natrénovaný model a jeho výsledky dále slouží jako měřítko účinnosti modelů s rozšířeným korpusem. Není totiž možné jednoznačně určit efektivitu modelu pouze na základě porovnání s modelem trénovaným na anglickém korpusu. To je zřejmé i z výsledků testování tohoto modelu na datech stejného korpusu (viz Tabulka 9.3), kde je možné sledovat, že tento model dosahuje výrazně nižších hodnot SI-SDRi než anglický model na anglických datech.

Tento model dosahuje při testování na korpusu tai + tai o 3,1 dB lepších výsledků než původní model trénovaný na angličtině. Rozdíl je znatelný i v případě subjektivního hodnocení autora, ale i přesto není srozumitelnost ideální a výstupní nahrávky jsou stále poměrně často poškozené nebo obsahují řeč obou řečníků. To poukazuje na skutečnost, že tento korpus je pro separaci výrazně složitější než korpus anglický. To může být způsobeno různými faktory, například subjektivní kvalita nezkreslených referenčních nahrávek je horší. Specifický důvod je ale složité určit.

Při testech na off-domain datech dosáhl model velmi špatných výsledků. Hodnoty zlepšení ztrátové funkce jsou v tomto případě velmi nízké a pro anglický korpus dokonce záporné. Také v rámci subjektivního hodnocení autora je řeč velmi špatně srozumitelná a ve většině případů výrazně poškozená.

Testovací data	SI-SDRi
eng + eng	-2,4 dB
tai + tai	7,2 dB
eng + tai	1,8 dB

Tabulka 9.3: Testování modelu Conv-TasNet trénovaného na korpusu tai + tai (tt) clean na rozdílných jazycích

9.1.2 Model clean trénovaný na kombinaci taiwanské a anglické řeči

Další model byl natrénován na kombinaci taiwanského a anglického řečníka. I zde byla testována účinnost pro všechny tři korpusy. V rámci testování na korpusu eng + tai dosáhl model hodnoty SI-SDRi 18,4 dB. I výsledky subjektivního hodnocení autora byly pro tento test velmi dobré, výstupy ve většině případů obsahují čistou a dobře srozumitelnou řeč. Tento výsledek předčil i účinnost anglického modelu na anglickém korpusu. To může být pravděpodobně způsobeno výraznějším rozdílem hlasů anglických a taiwanských řečníků. Stejně jako u předchozího modelu je to ale pouze hypotéza a účinnost modelu může být ovlivněna různými faktory.

V rámci testování na off-domain datech jsou výsledky také velmi špatné. Při subjektivním hodnocení autora na korpusu eng + eng jsou výsledky velmi rozmanité, kdy některé z výstupů jsou poměrně kvalitní a nepoškozené, ale některé jsou naopak nesrozumitelné či obsahují řeč obou řečníků zároveň. Tato rozmanitost je znatelná i v rámci jednotlivých hodnot SI-SDRi. Výsledky subjektivního hodnocení autora na datech tai + tai jsou více jednoznačné. Ve většině případů obsahují výstupy řeč obou řečníků a výstupy jsou poškozené a nesrozumitelné. Stejně tak jsou nízké i hodnoty SI-SDRi u jednotlivých nahrávek.

Testovací data	SI-SDRi
eng + eng	1,5 dB
tai + tai	-1,8 dB
eng + tai	18,4 dB

Tabulka 9.4: Testování modelu Conv-TasNet trénovaného na korpusu eng + tai (et) clean na rozdílných jazycích

9.1.3 Rozšíření datové sady

V rámci testování efektivity rozšíření datové sady byly vytvořeny korpusy obsahující části korpusů eng + eng, eng + tai a tai + tai. Tyto korpusy byly pro kombinace dvou korpusů vytvořeny v poměrech 100/100, 80/20 a 50/50. V rámci experimentů byly trénovány modely v těchto kombinacích:

1. Kombinace korpusů eng + eng (ee), tai + tai (tt)

Pro model kombinující anglický a taiwanský korpus je možné sledovat poměrně dobré výsledky v rámci testování na datech eng + eng a tai + tai. Modely dosahují při testování na anglickém korpusu trochu slabších, ale přesto velmi kvalitních výsledků. Model ee50/tt50 má ovšem výrazně nižší účinnost na angličtině. Pro data v taiwanštině se modely, oproti modelu trénovaném pouze na taiwanském korpusu, zlepšily nebo přinejhorším zůstaly stejné jako výsledky původního modelu tai + tai.

V rámci testování na datové sadě eng + tai jsou oproti modelu eng + eng výsledky o něco horší. Modely jsou tedy poměrné účinné na datech, která jsou využita při trénování. Při kombinaci jazyků ovšem selhávají.

Konfigurace testovací sady	Konfigurace trénovací sady					
	eng + eng	tai + tai	eng + tai	ee100/tt100	ee80/tt20	ee50/tt50
eng + eng	**16,3 dB**	-2,4 dB	1,5 dB	13,8 dB	13,7 dB	9,0 dB
tai + tai	4,1 dB	**7,2 dB**	-1,8 dB	8,6 dB	7,4 dB	7,2 dB
eng + tai	7,7 dB	1,8 dB	**18,4 dB**	6,7 dB	5,0 dB	7,3 dB

Tabulka 9.5: Výsledky modelů clean s rozšířenými korpusy na různých jazycích (eng + eng [ee], tai + tai [tt])

2. Kombinace korpusů eng + eng (ee), eng + tai (et)

I v rámci kombinace anglického korpusu se směsí anglického a taiwanského řečníka byly natrénovány tři modely ve výše zmíněných poměrech. V tabulce níže je možné sledovat poměrně dobré výsledky při testování na korpusech eng + eng a eng + tai. Pouze model ee50/et50 má o něco nižší účinnost na anglickém korpusu. Modely ee100/et100 a ee80/et20 poměrně kvalitních výsledků na těchto datech. I přesto, že korpus eng + tai obsahuje taiwanské řečníky, nebyly modely úspěšné v navýšení účinnosti na korpusu tai + tai. Ta byla pro všechny vytvořené modely ještě nižší než u modelu trénovaného na anglickém korpusu.

Konfigurace testovací sady	Konfigurace trénovací sady					
	eng + eng	tai + tai	eng + tai	ee100/et100	ee80/et20	ee50/et50
eng + eng	**16,3 dB**	-2,4 dB	1,5 dB	15,1 dB	14,2 dB	9,6 dB
tai + tai	4,1 dB	**7,2 dB**	-1,8 dB	1,3 dB	3,8 dB	-0,3 dB
eng + tai	7,7 dB	1,8 dB	**18,4 dB**	18,3dB	16,3 dB	17,6 dB

Tabulka 9.6: Výsledky modelů clean s rozšířenými korpusy na různých jazycích (eng + eng [ee], eng + tai [et])

3. Kombinace korpusů eng + eng (ee), eng + tai (et), tai + tai (tt)

V třetím případě byla vytvořena kombinace všech tří korpusů v poměrech 100/100/100, 80/10/10, 50/25/25 a 33/33/33. Modely kombinující všechny tři korpusy mají rozmanitější výsledky než předchozí modely.

Nejlepších výsledků opět dosahuje rozšířený model ee100/et100/tt100, který zachovává velmi dobré výsledky v rámci testování na datových sadách eng + eng a eng + tai. V rámci testování na korpusu tai + tai dosáhl dokonce na SI-SDRi 10 dB, což je o dalších 1,4 dB více než předchozí nejlepší model ee100/tt100 a téměř o 3 dB více než model trénovaný pouze na taiwanštině.

Nejlepším z modelů, které zachovávají stejnou velikost korpusu byl model s kombinací dat 80/10/10. Ten dosahuje o něco nižších, ale stále velmi dobrých hodnot SI-SDRi. I v rámci subjektivního hodnocení autora zůstávají výstupy velmi kvalitní a jen velmi málo poškozené.

Další dva modely už takto robustní nezůstávají. Model ee50/et25/tt25 sice stále dosahuje dobrých výsledků v rámci testování na korpusu tai + tai, ale pro testy na angličtině a směsi anglického a taiwanského řečníka už je mnohem méně účinný. Model ee33/et33/tt33 zachovává

velmi dobré výsledky v rámci korpusů tai + tai a eng + tai, ale pro separaci anglických řečníků už je téměř nepoužitelný.

Konfigurace testovací sady	Konfigurace trénovací sady						
	eng + eng	tai + tai	eng + tai	ee100/ et100/tt100	ee80/ et10/tt10	ee50/ et25/tt25	ee33/ et33/tt33
eng + eng	**16,3 dB**	-2,4 dB	1,5 dB	14,4 dB	14,1 dB	8,7 dB	4,8 dB
tai + tai	4,1 dB	**7,2 dB**	-1,8 dB	10,0 dB	7,5 dB	7,1 dB	7,9 dB
eng + tai	7,7 dB	1,8 dB	**18,4 dB**	17,8 dB	15,3 dB	7,9 dB	16,7 dB

Tabulka 9.7: Výsledky modelů clean s rozšířenými korpusy na různých jazycích (eng + eng [ee], eng + tai [et], tai + tai [tt])

9.1.4 Diskuze výsledků

Při testování separace řeči založené na datech se ukázalo, že je závislá na jazyku mluvčích. V rámci experimentu bylo dosaženo velmi dobrých výsledků na datech kombinující angličtinu a taiwanštinu. Obecně jsou z jednotlivých kategorií nejúčinnější modely s korpusy v poměrech 100/100 a 100/100/100. Modely kombinující dva korpusy dosahují velmi dobrých výsledků na datech, na kterých byly trénovány. Tyto modely ovšem selhávají na neviděném korpusu a pro využití na všech zmíněných situacích je nutné, aby při trénování byly využity data anglická, taiwanská, ale také jejich kombinace.

Jako nejlepší se ukázal model ee100/et100/tt100. Ten při testování na korpusech eng + eng i eng + tai dosahuje téměř stejně kvalitních výsledků jako modely trénované pouze na těchto datech. Pro testování na korpusu tai + tai dosahuje dokonce o téměř 3 dB lepšího výsledku než model trénovaný pouze na korpusu tai + tai.

Nejúčinnější z modelů, které zachovávají stejnou velikost datové sady jako, jsou modely s poměrem dat 80/20 a 80/10/10. Ty kopírují účinnost modelů s větším korpusem se zpravidla o něco málo nižší hodnotou SI-SDRi s maximální odchylkou o 2,5 dB. Jsou tedy také velmi dobrou volbou při výběru modelu funkčního pro angličtinu i taiwanštinu. Jejich výhodou je ale třikrát menší korpus a obecně kratší čas trénování modelu. U modelu ee100/tt100 bylo trénování také poměrně rychlé, protože dosáhl minima ztrátové funkce během méně epoch. V tomto ohledu tedy není možné tuto nevýhodu generalizovat na všechny modely s větším korpusem.

Předmětem dalšího výzkumu by tedy mohl být vliv jednotlivých sad na funkčnost při zachování objemu dat.

V rámci subjektivního hodnocení autora nejlepší modely dosahují velmi dobrých výsledků. Výstupní nahrávky jsou ve většině případů dobře srozumitelné, nejsou poškozené a také neobsahují řeč obou řečníků v jednom z výstupů. I v tomto ohledu bylo tedy trénování modelů s rozšířením korpusu úspěšné. I další subjektivní hodnocení autora u méně účinných modelů odpovídaly výsledné hodnotě SI-SDRi.

9.2 Model noisy na rozdílných jazycích

I pro testování modelu noisy byly využity dříve zmíněné vytvořené datové sady – taiwanský korpus a kombinace anglického a taiwanského řečníka. Ve verzi noisy k nim ale byly přidány nahrávky šumu z korpusu WHAM noises. Je zde vidět velmi podobná klesající tendence, kdy při testování na korpusu eng + tai dosahuje model hodnoty SI-SDRi 6,1 dB a na korpusu tai + tai pouze 4,2 dB. Vzhledem k faktu, že využitý šum je ze stejného korpusu jako v trénovací sadě, poradil si model s jeho odstraněním velmi dobře. I v tomto případě jsou ale výstupní signály velmi špatně srozumitelné a často poškozené.

Testovací data	SI-SDRi
eng + eng	13,3 dB
tai + tai	4,2 dB
eng + tai	6,1 dB

Tabulka 9.8: Testování modelu Conv-TasNet trénovaného na korpusu WHAM! na rozdílných jazycích

9.2.1 Model noisy trénovaný na taiwanské řeči

I pro situaci se šumem byl natrénován model na směsi dvou taiwanských řečníků. Tento model byl testován na všech třech modelových situacích (eng + eng, tai + tai, eng + tai). Ani takto natrénovaný model, ale nedosahuje o mnoho lepších výsledků než původní model, trénovaný na angličtině. V rámci testování na korpusu tai + tai tedy dosahuje o 2 dB lepších výsledků hodnoty SI-SDRi a také v rámci subjektivního hodnocení autora nebylo zlepšení tak výrazné jako u modelu s korpusem tai + tai clean. Přidání šumu v tomto případě tedy poměrně výrazně zvýšilo náročnost separace na tomto korpusu. Ve zbývajících testech model dosahoval velmi špatných výsledků, v rámci hodnoty ztrátové funkce i subjektivního hodnocení autora.

Testovací data	SI-SDRi
eng + eng	-1,6 dB
tai + tai	6,2 dB
eng + tai	1,5 dB

Tabulka 9.9: Testování modelu Conv-TasNet trénovaného na korpusu tai + tai (tt) noisy na rozdílných jazycích

9.2.2 Model noisy trénovaný na kombinaci taiwanské a anglické řeči

Jako další byl natrénován model na směsi taiwanské a anglické řeči. I tento model byl otestován pro všechny tři situace. Tento model v rámci testování na korpusu eng + tai ale vůči předchozím modelům dosáhl výrazného zlepšení a dosáhl hodnoty SI-SDRi 15 dB. I přesto se ale v rámci subjektivního hodnocení autora často vyskytuje v jednom z výstupů poškozená řeč, pozůstatky šumu, případně i fragmenty řeči druhého řečníka.

Stejně jako model trénovaný na korpusu tai + tai se v rámci off-domain testů výrazně zhoršil (viz Tabulka 9.10). V tomto případě jsou ale subjektivní hodnocení autora i jednotlivé hodnoty ztrátové funkce pro data eng + eng velmi variabilní, přičemž v některých případech jsou výsledné výstupy srozumitelné a některé výsledky jsou velmi zarušené a poškozené. Pro situaci tai + tai jsou výsledky jednotnější a ve většině případů jsou výstupy i hodnoty ztrátové funkce velmi špatné.

Testovací data	SI-SDRi
eng + eng	0,0 dB
tai + tai	-2,9 dB
eng + tai	15,0 dB

Tabulka 9.10: Testování modelu Conv-TasNet trénovaného na korpusu eng + tai (et) noisy na rozdílných jazycích

9.2.3 Rozšíření datové sady

I v tomto případě byly natrénovány modely na různých datových sadách, jejichž cílem bylo zprovoznit separaci řeči pro data obsahující taiwanské řečníky. Stejně jako u modelů typu clean zde byly trénovány modely v kombinacích:

1. Kombinace korpusů eng + eng (ee), tai + tai (tt)

U testování kombinace modelů na korpusu kombinujícím eng + eng a tai + tai můžeme sledovat velmi dobré výsledky v rámci testování na těchto korpusech. Jediný z modelů ee50/tt50 je

méně účinný při separaci dvou anglických řečníků. Pro korpus tai + tai ale dokonce dosahuje i lepších výsledků než model trénovaný pouze na těchto datech. Při testování na datech eng + tai lze poznat poměrně výrazné zhoršení vůči modelu eng + tai, a to v rámci hodnoty SI-SDRi a také při subjektivním hodnocení autora.

Konfigurace testovací sady	Konfigurace trénovací sady					
	eng + eng	tai + tai	eng + tai	ee100/tt100	ee80/tt20	ee50/tt50
eng + eng	**13,3 dB**	-1,6 dB	0,0 dB	11,4 dB	12,0 dB	8,6 dB
tai + tai	4,2 dB	**6,2 dB**	-2,9 dB	7,9 dB	6,8 dB	7,0 dB
eng + tai	6,1 dB	1,5 dB	**15,0 dB**	4,9 dB	4,4 dB	5,0 dB

Tabulka 9.11: Výsledky modelů noisy s rozšířenými korpusy na různých jazycích (eng + eng [ee], tai + tai [tt])

2. Kombinace korpusů eng + eng (ee), eng + tai (et)

I v rámci testování kombinace modelů na korpusu kombinujícím anglický korpus a směs anglických a taiwanských řečníku je možné sledovat, že modely dosahují velmi dobrých výsledků v situacích separace na datových sadách eng + eng i eng + tai. U všech modelů je ale možné pozorovat velmi špatné výsledky při testování na korpusu tai + tai. To je způsobeno tím, že jsou tato data při trénování neviděná. Model má s takovými daty problém pracovat, a to i přes skutečnost, že korpus eng + tai taiwanské řečníky obsahuje.

Konfigurace testovací sady	Konfigurace trénovací sady					
	eng + eng	tai + tai	eng + tai	ee100/et100	ee80/et20	ee50/et50
eng + eng	**13,3 dB**	-1,6 dB	0,0 dB	12,7 dB	12,5 dB	8,8 dB
tai + tai	4,2 dB	**6,2 dB**	-2,9 dB	-2,3 dB	-1,4 dB	-1,6 dB
eng + tai	6,1 dB	1,5 dB	**15,0 dB**	15,0 dB	13,8 dB	14,5 dB

Tabulka 9.12: Výsledky modelů noisy s rozšířenými korpusy na různých jazycích (eng + eng [ee], eng + tai [et])

3. Kombinace korpusů eng + eng (ee), eng + tai (et), tai + tai (tt)

Na testování modelů trénovaných na kombinaci všech tří korpusů je znatelné, že jsou nejlepším kompromisem pro využití na všech třech situacích. Pro korpusy eng + eng a eng + tai dosahují téměř stejně dobrých výsledků jako předchozí modely a pro korpus tai + tai dosahuje model

ee100/et100/tt100 dokonce ještě lepších výsledků. Modely trénované na této kombinaci jsou tedy nejrobustnější možností. Model ee100/et100/tt100 a případně i model s omezeným korpusem ee80/et10/tt10 dosahují velmi dobrých výsledků při testování na všech třech vytvořených korpusech.

I tak je ale důležité zmínit, že korpus tai + tai se ukázal jako velmi složitý a ani nejlepší model nedosahuje takové kvality výsledků jako v případě směsi dvou anglických řečníků či při kombinaci anglického a taiwanského řečníka.

Konfigurace testovací sady	Konfigurace trénovací sady						
	Původní	tai + tai	eng + tai	ee100/ et100/tt100	ee80/ et10/tt10	ee50/ et25/tt25	ee33/ et33/tt33
eng + eng	**13,3 dB**	-1,6 dB	0,0 dB	11,9 dB	11,9 dB	9,1 dB	4,5 dB
tai + tai	4,2 dB	**6,2 dB**	-2,9 dB	8,8 dB	6,1 dB	5,3 dB	6,2 dB
eng + tai	6,1 dB	1,5 dB	**15,0 dB**	14,3 dB	12,9 dB	13,5 dB	13,6 dB

Tabulka 9.13: Výsledky modelů noisy s rozšířenými korpusy na různých jazycích (eng + eng [ee], eng + tai [et], tai + tai [tt])

9.2.4 Diskuze výsledků

Zlepšení hodnoty SI-SDRi modelů typu noisy pro data obsahující taiwanské řečníky je téměř totožné jako u modelů clean. Nejúčinnějšími byly také modely s rozšířeným korpusem 100/100, ale pouze na datech použitých při trénování. Nejlepším se ukázal model kombinující všechny tři korpusy v poměru 100/100/100, který dosahuje téměř stejných výsledků na korpusech eng + eng a eng + tai jako modely trénované samostatně na těchto korpusech. Také dosahuje ještě lepší účinnosti při testování na taiwanském korpusu než model trénovaný na korpusu tai + tai.

Jako velmi účinné se také ukázaly modely s 80 % anglického korpusu. Ty se od modelů s většími korpusy liší maximálně o 1,7 dB SI-SDRi. Jejich výsledky jsou tedy ve většině případů téměř srovnatelné. I zde je výhodou těchto model menší velikost datové sady a obecně kratší trénování. Ale i v tomto případě nelze rychlost trénování ani dosažené SDR generalizovat.

Bohužel se ale korpus tai + tai v kombinaci s přidaným šumem ukázal o mnoho náročnější na separaci než v případě dat bez šumu. To znamená, že výsledné výstupní nahrávky nedosahují vysoké kvality ani v případě nejúčinnějších modelů. Při subjektivním hodnocení autora i

v rámci porovnání hodnot ztrátové funkce jsou tedy výsledky oproti původnímu modelu trénovaném na anglickém korpusu o dost lepší, ale stále výsledky nedosahují tak dobré kvality. I přesto ale nejlepší model dosahuje obecně lepších výsledků než model trénovaný pouze na taiwanštině.

10 Experimenty na různém počtu řečníků

Hlavní vybranou variantou off-domain dat z možností v zadání byl rozdílný jazyk mluvčích. Toto téma bylo detailně analyzováno a objektivně kvantitativně vyhodnoceno, původní zadání tedy bylo splněno v předchozích částech této práce. V této části je částečně rozpracována ještě další oblast off-domain dat – proměnný počet současně aktivních řečníků.

Přesto, že se v rámci experimentů podařilo získat poměrně kvalitní výsledky, v rámci testování nebyl nalezen způsob, jak správně počítat hodnoty ztrátové funkce pro data s jedním řečníkem. Výsledky výpočtů v tomto případě neodráželi skutečnou kvalitu výstupních signálů. Následující kapitola tedy popisuje ne zcela kompletní analýzu výsledků založenou pouze na subjektivním kvalitativním hodnocení autora.

10.1 Model clean na různém počtu řečníků

V experimentech je využita datová sada WHAM! [53] (resp. wsj0-2mix, tedy WHAM! bez šumu), vytvořena z kombinací nahrávek dvou řečníků z korpusu WSJ0. Tato datová sada byla pro účely experimentu rozšířena o nahrávky „ticha“, které jsou vytvořeny metodou silent z Python knihovny Pydub, ta vytvoří signál o původní délce vyplněný nulami. Tato data jsou používána na vstupech a referenčních nahrávkách, ve kterých se nenachází řečník.

Pro model clean bylo provedeno testování na datech s různým počtem řečníků. Tyto testy měly ověřit, jak se natrénovaný model bude chovat v případě, že testovací data neodpovídají datům použitým při trénování. V rámci těchto testů byly v případech, že výsledek měl obsahovat pouze jednoho řečníka, zavedeny jako referenční data pro druhý odhadovaný signál nahrávky „ticha“. Při testování pouze na šumu, ticho nahradilo obě referenční nahrávky. Po subjektivním hodnocení autora je u případů s přítomností pouze jednoho řečníka zřejmé, že model se pokouší rozdělit řeč do 2 výstupů podle jemných změn. Proto jsou výsledné odhady řeči poměrně špatné a chybí v nich části řeči či obsahují šum. Zároveň lze pozorovat, že při zpracování směsí obsahující šum, si model neumí poradit s hlukem v pozadí, který se v trénovací a validační sadě vůbec nevyskytuje.

10.1.1 Rozšíření datové sady

V rámci rozšíření modelu ke zprovoznění separace řeči pro různý počet řečníků bylo natrénováno několik testovacích modelů. Tyto modely zůstávají téměř totožné jako tento referenční

model. Na výstupu, kde se řečník nenachází, by bylo v ideálním případě vhodné výstup bez řečníka odhadovat jako ticho, případně co nejvíce ztišený šum.

Stěžejní změnou je přidání dat s jedním řečníkem i tichem. Výsledná datová sada je tedy kombinací směsí 2 řečníků, jednoho řečníka a samotného ticha. Tyto kombinace byly vytvořeny v poměrech 100/100/100, 33/33/33, 50/25/25 a 80/10/10 (uvedeno v procentech původní velikosti korpusu, pro trénování tedy z 20 000 nahrávek) pro tyto 3 případy.

10.1.2 Diskuze výsledků

Trénování s korpusem v poměru 100/100/100 je časově i objemem dat poměrně neefektivní. Tento test byl proveden jako reference k ostatním, kde jsou data již omezená. Trénování tohoto modelu běželo 199 epoch a konečný čas trénování byl více než dvakrát delší než při trénování původního modelu.

Tento model se oproti výchozímu modelu lehce zhoršil v rozpoznávání dat s dvěma řečníky. Pro data s jedním řečníkem model funguje dobře. První výstup s odhadem řečníka obsahuje jasně srozumitelnou řeč. Při využití původního modelu chybí na výstupu části řeči a srozumitelnost je obecně horší. V tomto případě je tedy model pro reálné využití účinnější.

V rámci rešerše [66] a experimentů se ukázalo, že systém Conv-TasNet nepracuje dobře s prázdnými výstupy. V případě zmíněného článku i experimentů provedených v této diplomové práci obsahují výstupy, kde se nenachází řečník, šum či fragmenty řeči (oproti původnímu modelu částečně utlumené, ale stále nedostatečně). Výsledky pro data se šumem jsou, stejně jako v případě původního modelu, velmi špatné, protože v trénovacích datech se šum nevyskytuje.

Obdobně jako u modelů v přechozí kapitole se jako nejrobustnější ukázaly modely s korpusem v poměru 100/100/100 a 80/10/10. Další dva trénované modely už jsou o poznání horší na datech s dvěma řečníky. Všechny modely ale po subjektivním zhodnocení fungují dobře pro data s jedním řečníkem

V účinnosti není po porovnání subjektivního hodnocení autora mezi modelem s korpusem 100/100/100 a 80/10/10 tak výrazný rozdíl. Model s korpusem 100/100/100 je účinnější v situacích se dvěma řečníky a model s korpusem 80/10/10 exceluje v situacích s jedním řečníkem. Oba modely tedy mohou být velmi dobrou volbou pro situace, při kterých mluví jeden nebo dva řečníci zároveň. Výhodou menšího modelu je samozřejmě i třikrát menší velikost korpusu a kratší čas trénování modelu. Proto by byl v mnoha situacích vhodnější volbou.

Po získání výsledků experimentu je možné shrnout několik důležitých bodů o chování modelů a systému Conv-TasNet:

1. Velmi zásadní je reakce systému Conv-TasNet na prázdné výstupy při trénování. Bohužel se ukázalo, že Conv-TasNet práci s takovými daty nezvládá a výstupy bez řečníka obsahují poškozená data namísto prázdných kanálů. Data bez řečníka jsou tedy nejspíše postradatelné. Zároveň je pravděpodobné, že by model mohl dosahovat lepších výsledků pouze s daty dvou řečníků a jednoho řečníka.
2. V rámci tohoto testování lze z výsledků sledovat, že pouhé přidání off-domain dat neznamená jasné zlepšení účinnosti modelu na těchto datech. To ale může být způsobeno i velkým počtem dat bez řečníka, které mohly výsledek trénování narušit.
3. Conv-TasNet je možné, s určitými omezeními, použít pro data s různým počtem řečníků. Je pravděpodobné, že použitelných výsledků by mohlo být dosaženo i v rámci jiných počtů řečníků.

10.2 Model noisy na různém počet řečníků

I pro model noisy proběhlo testování na datech s různým počtem řečníků. Stejně jako v předchozím případě nahradila referenční data pro prázdné signály ticho. Oproti modelu trénovaném na korpusu wsj0-2mix má model trénovaný na korpusu WHAM! zjevnou výhodu na datech obsahujících šum. Proto měl v těchto oblastech obecně lepší výsledky, kde předchozí model selhal na odstranění šumu. Poměrně kvalitní výsledek model zaznamenal i na datech korpusu wsj0-2mix. I v případě tohoto modelu jsou data s jedním řečníkem v rámci subjektivního hodnocení autora často poškozena a chybí v nich části řeči či naopak obsahují šum.

10.2.1 Rozšíření datové sady

I v případě modelu noisy byly provedeny experimenty ke zprovoznění modelu pro různý počet řečníků. Stejně jako v případě předchozího modelu byla při trénování využita ztrátová funkce SDR.

V rámci přidání dat se tento model liší tím, že datovou sadu v tomto případě tvoří kombinace směsí 2 řečníků a šumu, jednoho řečníka a šumu a samotného šumu. Byly použity stejné poměry dat 100/100/100, 33/33/33, 50/25/25 a 80/10/10 pro všechny 3 případy. I zde modely naráží na omezení systému Conv-TasNet, i přesto že modely noisy pracují se zbytkovým šumem o něco lépe. Ale na prázdných výstupech se stále nachází šum či fragmenty řeči.

10.2.2 Diskuze výsledků

Stejně jako v případě modelu clean bylo trénování v poměru 100/100/100 původně provedeno jako reference k ostatním, kde jsou data již omezená. Ale je vhodné zmínit, že trénování tohoto modelu běželo pouhých 85 epoch, takže konečný čas trénování byl pouze o 40 % delší než při trénování původního modelu. Tento model je velmi účinný pro data s jedním řečníkem a zároveň si zachovává vysokou účinnost pro data s dvěma řečníky.

Modely s korpusem s přidaným šumem téměř kopírují výsledky modelů trénovaných bez šumu. Také je ale vhodné zmínit, že modely dosahují velmi dobrých výsledků u dat se šumem, ale i bez šumu. Oba nejlepší modely (s korpusy v poměrech 100/100/100 i 80/10/10) mají tedy dobrý potenciál pro využití v reálných situacích, kde podmínky mohou být závislé na okolnostech nahrávání a v některých případech i proměnlivé.

Pro modely noisy platí obdobné výsledky jako v případě modelů clean, přesto je zde možné sledovat jisté odlišnosti.

1. V tomto případě je oproti modelům clean možné vidět, že model s největším korpusem je skutečně nejlepší volbou pro data s různým počtem řečníků.
2. Modely noisy jsou teoreticky lepší pro reálné využití, protože počítají i s možností, že nahrávky obsahují šum. Nespoléhají tedy na ideální podmínky nahrávání. Jsou robustnější vůči změnám situace, ale zároveň zachovávají poměrně dobrou účinnost i pro data bez šumu.

11 Závěr

Tato práce zkoumá, jak se chovají modely v situacích s off-domain daty a jaké změny v testovací množině je již třeba vnímat jako off-domain data. V rámci této práce byla zkoumána změna jazyka mluvčích a částečně zkoumána i změna počtu aktivních mluvčích. Dále poukazuje na výhody i nevýhody konvoluční sítě Conv-TasNet. Zaměřuje se na variantu volitelného bodu zadání, kterým je adaptace modelu. Vybranou variantou je porovnání rozšíření a částečné změny složení trénovací sady jako možného řešení problematiky off-domain dat. Výsledky modelů využívajících rozšířené datové sady jsou popsány v jednotlivých částech níže.

V této práci jsou nejprve popsány původní metody zlepšování řeči, ze kterého separace řeči vychází. Na tuto část navazuje popis různých metod separace řeči od původních pre-neurálních přístupů až po nejmodernější metody s jejich specifiky i problémy. Separace řeči je také krátce porovnána s problematikou extrakce cílové řeči. Jsou zde rovněž popsány faktory ovlivňující výkon separace řeči. Tím je i neshoda oblastí dat (off-domain data), která je dále řešena i v rámci experimentů.

Pro trénování modelů byla využita konvoluční síť pro separaci zvuku v časové oblasti Conv-TasNet. Během úvodního testování proběhlo také porovnání této sítě s dalšími metodami pro separaci řeči. V rámci těchto testů a rešerše bylo potvrzeno, že Conv-TasNet dokáže poměrně dobře konkurovat i nejmodernějším metodám separace řeči. Zároveň je také výhodou, že má v porovnání s ostatními rychlé trénování a zároveň je výsledný model poměrně malý. To bylo výhodou i při zpracování této práce, protože rychlé trénování umožnilo natrénovat velké množství modelů pro vzájemné porovnání výsledků jednotlivých modelů. Tomu velmi napomohlo i trénování na výkonných výpočetních strojích v rámci gridu MetaCentra. Trénování na běžném osobním počítači by v takovém množství nebylo možné.

V rámci práce byly jako výchozí použity dva korpusy: WSJ0-2mix, který je bez šumu na pozadí a jeho variantu WHAM!, která je rozšířená o šum.

Hlavním cílem této práce je prozkoumání toho, jak se metoda separace řeči chová při mírné změně dat a kdy je tato změna už problémová. Jako hlavní část práce byla z variant off-domain dat nabízených v zadání vybrána varianta zabývající se změnou jazyka mluvčích. Ta byla detailně zkoumána a následně kvantitativně vyhodnocena. Částečně byla nad rámec zadání rozpracována i varianta s proměnným počtem mluvčích.

Při testování původního modelu trénovaného na angličtině bylo zjištěno, že změna jazyka mluvčího je problémová a separace založená na datech je tedy, na rozdíl od slepé separace, jazykově závislá. Při trénování modelu na rozšířené datové sadě, byly k anglickým datům přidány data v taiwanštině. Tento model funguje pro testovací data v angličtině (ee) i v taiwanštině (tt). Přesto jejich kombinace (et) stále představuje neviděná data. Ukázalo se tedy, že v případě kombinace rozdílných jazyků nestačí přidat do trénovací sady data obsahující pouze druhý jazyk, ale je nutné přidat i data kombinující oba jazyky, což multilinguální separaci může dost komplikovat. Úplný trénovací korpus tedy kombinuje všechny tři datové sady. Nejlépe funguje model s korpusem ve variantě ee100/tt100/et100.

Nejúčinnější alternativou s menší datovou sadou odpovídající velikosti původního anglického korpusu byl model s poměry dat 80 % anglických, 10 % taiwanských a 10 % kombinovaných dat. Ten dosahoval pouze o trochu horších výsledků, ale zároveň má třikrát menší velikost datové sady a trénování bylo také téměř třikrát kratší. Oba tyto modely byly vyhodnoceny jako použitelné pro situace, kdy je nutné pracovat s oběma jazyky. Obdobné rozšíření dat by tedy mohlo v podobných případech být velmi účinné pro zprovoznění modelu pro různé jazyky v závislosti na situaci.

V současné době se problém různých jazyků často řeší hrubou silou, tedy přidáváním kombinace všech jazyků. To ale nemusí být jednoduše proveditelné. Pro případné zlepšení porozumění této problematice by mohlo být provedeno více testů na různých jazycích. Je pravděpodobné, že v takových případech bude v rámci účinnosti modelu záležet na podobnosti jazyků, kvalitě nahrávek a dalších specifikách, které by bylo vhodné blíže zkoumat. Je ale pravděpodobné, že řešení této problematiky není pouze technické a je nutné se zamyslet nad rozdíly a specifiky jednotlivých jazyků.

Obecně se v rámci testování úloha se šumem ukázala jako částečně náročnější, a to v rámci hodnoty ztrátové funkce i subjektivního hodnocení autora. Lepší výsledky byl tedy obecně zaznamenány pro data bez šumu.

Další variantou off-domain experimentů bylo testování na datech s proměnným počtem řečníků. Při testování původních modelů se ukázalo, že nefungují v situacích, kdy mluví pouze jeden řečník. V takových situacích byly výsledné výstupy velmi poškozené. V rámci těchto experimentů byly vytvořeny datové sady obsahující nahrávky s jedním řečníkem a také nahrávky ticha, respektive šumu. Na kombinacích těchto korpusů v různých poměrech byly natrénovány modely. Tyto modely umožnily pracovat s daty s jedním řečníkem, na kterých

výchozí modely selhávaly a nejrobustnější modely zároveň zachovaly velmi dobrou účinnost pro data s dvěma řečníky. Ty ovšem byly kvůli problémům s kvantitativními kritérii, vyhodnoceny pouze subjektivně na základě hodnocení autora. V rámci testování a rešerše se ale ukázalo, že Conv-TasNet pracuje velmi špatně s výstupy bez řečníka, které obsahují poškozená data namísto prázdných kanálů.

[1] LUO, Yi a Nima MESGARANI. Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* [online]. 2019, **27**(8), 1256–1266. ISSN 2329-9290, 2329-9304. Dostupné z: doi:10.1109/TASLP.2019.2915167

[2] LOIZOU, Philipos. *Speech Enhancement: Theory and Practice* [online]. 2007. ISBN 978-0-429-09618-1. Dostupné z: doi:10.1201/b14529

[3] VASEGHI, Saeed V. Spectral Subtraction. In: Saeed V. VASEGHI, ed. *Advanced Signal Processing and Digital Noise Reduction* [online]. Wiesbaden: Vieweg+Teubner Verlag, 1996 [vid. 2022-11-18], s. 242–260. ISBN 978-3-322-92773-6. Dostupné z: doi:10.1007/978-3-322-92773-6_9

[4] BANDO, Yoshiaki, Masato MIMURA, Katsutoshi ITOYAMA, Kazuyoshi YOSHII a Tatsuya KAWAHARA. Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* [online]. 2018, s. 716–720 [vid. 2022-12-07]. Dostupné z: doi:10.1109/ICASSP.2018.8461530

[5] MIRSAMADI, Seyedmahdad a Ivan TASHEV. Causal Speech Enhancement Combining Data-Driven Learning and Suppression Rule Estimation. In: *Interspeech 2016*: *Interspeech 2016* [online]. B.m.: ISCA, 2016, s. 2870–2874 [vid. 2022-11-26]. Dostupné z: doi:10.21437/Interspeech.2016-437

[6] HIGUCHI, Takuya, Keisuke KINOSHITA, Marc DELCROIX a Tomohiro NAKATANI. Adversarial training for data-driven speech enhancement without parallel corpus. In: *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*: *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* [online]. 2017, s. 40–47. Dostupné z: doi:10.1109/ASRU.2017.8268914

[7] ERKELENS, Jan, Jesper JENSEN a Richard HEUSDENS. A data-driven approach to optimizing spectral speech enhancement methods for various error criteria. *Speech Communication* [online]. 2007, **49**(7), Speech Enhancement, 530–541. ISSN 0167-6393. Dostupné z: doi:10.1016/j.specom.2006.06.012

[8] LIPPMANN, Richard P. Review of Neural Networks for Speech Recognition. *Neural Computation* [online]. 1989, **1**(1), 1–38. ISSN 0899-7667, 1530-888X. Dostupné z: doi:10.1162/neco.1989.1.1.1

[9] NASSIF, Ali Bou, Ismail SHAHIN, Imtinan ATTILI, Mohammad AZZEH a Khaled SHAALAN. Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access* [online]. 2019, **7**, 19143–19165. ISSN 2169-3536. Dostupné z: doi:10.1109/ACCESS.2019.2896880

[10] SHIFAS, Muhammed PV, Nagaraj ADIGA, Vassilis TSIARAS a Yannis STYLIANOU. A non-causal FFTNet architecture for speech enhancement. In: *Interspeech 2019* [online]. 2019, s. 1826–1830 [vid. 2023-01-24]. Dostupné z: doi:10.21437/Interspeech.2019-2622

[11] GRANCHAROV, V., J. SAMUELSSON a B. KLEIJN. On causal algorithms for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing* [online]. 2006, **14**(3), 764–773. ISSN 1558-7924. Dostupné z: doi:10.1109/TSA.2005.857802

[12] ŽMOLÍKOVÁ, Kateřina. *Neurální extrakce řeči cílového řečníka* [online]. Brno, CZ, 2022. Disertační práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Dostupné z: https://www.fit.vut.cz/study/phd-thesis/1009/

[13] *Some Experiments on the Recognition of Speech, with One and with Two Ears: The Journal of the Acoustical Society of America: Vol 25, No 5* [online]. [vid. 2022-10-08]. Dostupné z: https://asa.scitation.org/doi/10.1121/1.1907229

[14] DIGHE, Pranay, Marc FERRÀS a Hervé BOURLARD. Detecting and labeling speakers on overlapping speech using vector taylor series. In: *Interspeech 2014*: *Interspeech 2014* [online]. B.m.: ISCA, 2014, s. 592–596 [vid. 2022-10-15]. Dostupné z: doi:10.21437/Interspeech.2014-143

[15] BARKER, Jon, Shinji WATANABE, Emmanuel VINCENT a Jan TRMAL. *The fifth „CHiME" Speech Separation and Recognition Challenge: Dataset, task and baselines* [online]. B.m.: arXiv. 28. březen 2018 [vid. 2022-12-16]. Dostupné z: http://arxiv.org/abs/1803.10609. arXiv:1803.10609 [cs, eess]

[16] *Distant microphone speech recognition in everyday environments: from CHiME-5 to CHiME-6 - Jon Barker* [online]. 2019 [vid. 2022-10-11]. Dostupné z: https://www.youtube.com/watch?v=UQci8LgwZOc

[17] PARK, Tae Jin, Naoyuki KANDA, Dimitrios DIMITRIADIS, Kyu J. HAN, Shinji WATANABE a Shrikanth NARAYANAN. *A Review of Speaker Diarization: Recent Advances with Deep Learning* [online]. B.m.: arXiv. 26. listopad 2021 [vid. 2022-10-30]. Dostupné z: http://arxiv.org/abs/2101.09624. arXiv:2101.09624 [cs, eess]

[18] BROWN, Guy J. a Martin COOKE. Computational auditory scene analysis. *Computer Speech & Language* [online]. 1994, **8**(4), 297–336. ISSN 0885-2308. Dostupné z: doi:10.1006/csla.1994.1016

[19] BREGMAN, Albert. Auditory Scene Analysis: The Perceptual Organization of Sound. In: *Journal of The Acoustical Society of America - J ACOUST SOC AMER* [online]. 1990. Dostupné z: doi:10.1121/1.408434

[20] SMARAGDIS, Paris, Cédric FÉVOTTE, Gautham J. MYSORE, Nasser MOHAMMADIHA a Matthew HOFFMAN. Static and Dynamic Source Separation Using

Nonnegative Factorizations: A unified view. *IEEE Signal Processing Magazine* [online]. 2014, **31**(3), 66–75. ISSN 1558-0792. Dostupné z: doi:10.1109/MSP.2013.2297715

[21] VIRTANEN, Tuomas. Speech recognition using factorial hidden Markov models for separation in the feature space. In: [online]. 2006. Dostupné z: doi:10.21437/Interspeech.2006-23

[22] *Independent Component Analysis and Blind Signal Separation* [online]. nedatováno [vid. 2023-01-24]. Dostupné z: https://link.springer.com/book/10.1007/11679363

[23] CHATFIELD, Christopher a Alexander J. COLLINS. Principal component analysis. In: Christopher CHATFIELD a Alexander J. COLLINS, ed. *Introduction to Multivariate Analysis* [online]. Boston, MA: Springer US, 1980 [vid. 2023-01-24], s. 57–81. ISBN 978-1-4899-3184-9. Dostupné z: doi:10.1007/978-1-4899-3184-9_4

[24] *Time-Frequency Masking for Speech Separation and Its Potential for Hearing Aid Design* [online]. [vid. 2023-01-26]. Dostupné z: doi:10.1177/1084713808326455

[25] *Encyclopedia of Machine Learning* [online]. nedatováno [vid. 2022-11-12]. Dostupné z: https://link.springer.com/book/10.1007/978-0-387-30164-8

[26] *Loss Functions — ML Glossary documentation* [online]. [vid. 2022-11-12]. Dostupné z: https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html

[27] LI, Shuai, Hongqing LIU, Yi ZHOU a Zhen LUO. A SI-SDR Loss Function based Monaural Source Separation. In: *2020 15th IEEE International Conference on Signal Processing (ICSP)*: *2020 15th IEEE International Conference on Signal Processing (ICSP)* [online]. 2020, s. 356–360. ISSN 2164-5221. Dostupné z: doi:10.1109/ICSP48669.2020.9321080

[28] YOUSEFI, Midia, Soheil KHORRAM a John H.L. HANSEN. Probabilistic Permutation Invariant Training for Speech Separation. In: *Interspeech 2019*: *Interspeech 2019* [online]. B.m.: ISCA, 2019, s. 4604–4608 [vid. 2023-01-20]. Dostupné z: doi:10.21437/Interspeech.2019-1827

[29] WANG, Yannan, Jun DU, Li-Rong DAI a Chin-Hui LEE. Unsupervised single-channel speech separation via deep neural network for different gender mixtures. In: *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*: *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* [online]. 2016, s. 1–4. Dostupné z: doi:10.1109/APSIPA.2016.7820736

[30] CHIEN, Jen-Tzung a Kuan-Ting KUO. Variational Recurrent Neural Networks for Speech Separation. In: *Interspeech 2017*: *Interspeech 2017* [online]. B.m.: ISCA, 2017, s. 1193–1197 [vid. 2023-01-19]. Dostupné z: doi:10.21437/Interspeech.2017-832

[31] HUANG, Po-Sen, Minje KIM, Mark HASEGAWA-JOHNSON a Paris SMARAGDIS. Deep learning for monaural speech separation. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* [online]. 2014, s. 1562–1566. ISSN 2379-190X. Dostupné z: doi:10.1109/ICASSP.2014.6853860

[32] DU, Jun, Yanhui TU, Li-Rong DAI a Chin-Hui LEE. A Regression Approach to Single-Channel Speech Separation Via High-Resolution Deep Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* [online]. 2016, **24**(8), 1424–1437. ISSN 2329-9304. Dostupné z: doi:10.1109/TASLP.2016.2558822

[33] DU, Jun, Yanhui TU, Yong XU, Lirong DAI a Chin-Hui LEE. Speech separation of a target speaker based on deep neural networks. In: *2014 12th International Conference on Signal Processing (ICSP 2014)*: *2014 12th International Conference on Signal Processing (ICSP)* [online]. Hangzhou: IEEE, 2014, s. 473–477 [vid. 2022-11-26]. ISBN 978-1-4799-2188-1. Dostupné z: doi:10.1109/ICOSP.2014.7015050

[34] ZHANG, Xiao-Lei a DeLiang WANG. A Deep Ensemble Learning Method for Monaural Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* [online]. 2016, **24**(5), 967–977. ISSN 2329-9304. Dostupné z: doi:10.1109/TASLP.2016.2536478

[35] ISIK, Yusuf, Jonathan Le ROUX, Zhuo CHEN, Shinji WATANABE a John R. HERSHEY. *Single-Channel Multi-Speaker Separation using Deep Clustering* [online]. B.m.: arXiv. 7. červenec 2016 [vid. 2022-10-16]. Dostupné z: http://arxiv.org/abs/1607.02173. arXiv:1607.02173 [cs, stat]

[36] YU, Dong, Morten KOLBÆK, Zheng-Hua TAN a Jesper JENSEN. *Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation* [online]. B.m.: arXiv. 3. leden 2017 [vid. 2023-02-16]. Dostupné z: http://arxiv.org/abs/1607.00325. arXiv:1607.00325 [cs, eess]

[37] Fig. 1. The building blocks of Conv-Tasnet. *ResearchGate* [online]. [vid. 2022-10-23]. Dostupné z: https://www.researchgate.net/figure/The-building-blocks-of-Conv-Tasnet_fig1_346511133

[38] LEA, Colin, Rene VIDAL, Austin REITER a Gregory D. HAGER. *Temporal Convolutional Networks: A Unified Approach to Action Segmentation* [online]. B.m.: arXiv. 29. srpen 2016 [vid. 2023-03-03]. Dostupné z: http://arxiv.org/abs/1608.08242. arXiv:1608.08242 [cs]

[39] ŽMOLÍKOVÁ, Kateřina, Marc DELCROIX, Keisuke KINOSHITA, Takuya HIGUCHI, Atsunori OGAWA a Tomohiro NAKATANI. Speaker-Aware Neural Network Based Beamformer for Speaker Extraction in Speech Mixtures. In: *Interspeech 2017*: *Interspeech 2017* [online]. B.m.: ISCA, 2017, s. 2655–2659 [vid. 2022-10-30]. Dostupné z: doi:10.21437/Interspeech.2017-667

[40] WANG, DeLiang a Jitong CHEN. Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* [online]. 2018, **26**(10), 1702–1726. ISSN 2329-9304. Dostupné z: doi:10.1109/TASLP.2018.2842159

[41] MALEK, Jiri, Jakub JANSKY, Zbynek KOLDOVSKY, Tomas KOUNOVSKY, Jaroslav CMEJLA a Jindrich ZDANSKY. Target Speech Extraction: Independent Vector Extraction Guided by Supervised Speaker Identification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* [online]. 2022, **30**, 2295–2309. ISSN 2329-9290, 2329-9304. Dostupné z: doi:10.1109/TASLP.2022.3190739

[42] ZMOLIKOVA, Katerina, Marc DELCROIX, Keisuke KINOSHITA, Tsubasa OCHIAI, Tomohiro NAKATANI, Lukas BURGET a Jan CERNOCKY. SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures. *IEEE Journal of Selected Topics in Signal Processing* [online]. 2019, **13**(4), 800–814. ISSN 1932-4553, 1941-0484. Dostupné z: doi:10.1109/JSTSP.2019.2922820

[43] SHINODA, Koichi. Speaker adaptation techniques for speech recognition using probabilistic models. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)* [online]. 2005, **88**(12), 25–42. ISSN 1520-6440. Dostupné z: doi:10.1002/ecjc.20207

[44] HABETS, Emanuël A. P. Speech Dereverberation Using Statistical Reverberation Models. In: Patrick A. NAYLOR a Nikolay D. GAUBITCH, ed. *Speech Dereverberation* [online]. London: Springer, 2010 [vid. 2022-10-30], Signals and Commmunication Technology, s. 57–93. ISBN 978-1-84996-056-4. Dostupné z: doi:10.1007/978-1-84996-056-4_3

[45] *4th ASA/ASJ Joint Meeting Lay Language Papers - How to compare concert halls by listening to music* [online]. [vid. 2023-01-26]. Dostupné z: https://acoustics.org/pressroom/httpdocs/152nd/behler.html

[46] DITTER, David a Timo GERKMANN. Influence of Speaker-Specific Parameters on Speech Separation Systems. In: *Interspeech 2019*: *Interspeech 2019* [online]. B.m.: ISCA, 2019, s. 4584–4588 [vid. 2022-10-30]. Dostupné z: doi:10.21437/Interspeech.2019-2459

[47] MACIEJEWSKI, Matthew, Gregory SELL, Yusuke FUJITA, Leibny Paola GARCIA-PERERA, Shinji WATANABE a Sanjeev KHUDANPUR. Analysis of Robustness of Deep Single-Channel Speech Separation Using Corpora Constructed From Multiple Domains. In: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* [online]. 2019, s. 165–169. ISSN 1947-1629. Dostupné z: doi:10.1109/WASPAA.2019.8937153

[48] PASZKE, Adam, Sam GROSS, Francisco MASSA, Adam LERER, James BRADBURY, Gregory CHANAN, Trevor KILLEEN, Zeming LIN, Natalia GIMELSHEIN, Luca ANTIGA, Alban DESMAISON, Andreas KÖPF, Edward YANG, Zach DEVITO, Martin RAISON, Alykhan TEJANI, Sasank CHILAMKURTHY, Benoit STEINER, Lu FANG, Junjie BAI a Soumith CHINTALA. *PyTorch: An Imperative Style, High-Performance Deep Learning Library* [online]. B.m.: arXiv. 3. prosinec 2019 [vid. 2022-10-05]. Dostupné z: http://arxiv.org/abs/1912.01703. arXiv:1912.01703 [cs, stat]

[49] FALCON, William a Kyunghyun CHO. *A Framework For Contrastive Self-Supervised Learning And Designing A New Approach* [online]. B.m.: arXiv. 31. srpen 2020 [vid. 2022-10-03]. Dostupné z: http://arxiv.org/abs/2009.00104. number: arXiv:2009.00104arXiv:2009.00104 [cs]

[50] PARIENTE, Manuel, Samuele CORNELL, Joris COSENTINO, Sunit SIVASANKARAN, Efthymios TZINIS, Jens HEITKAEMPER, Michel OLVERA, Fabian-Robert STÖTER, Mathieu HU, Juan M. MARTÍN-DOÑAS, David DITTER, Ariel FRANK, Antoine DELEFORGE a Emmanuel VINCENT. *Asteroid: the PyTorch-based audio source separation toolkit for researchers* [online]. B.m.: arXiv. 8. květen 2020 [vid. 2022-10-05]. Dostupné z: http://arxiv.org/abs/2005.04132. arXiv:2005.04132 [cs, eess]

[51] *O MetaCentru* [online]. [vid. 2023-02-07]. Dostupné z: https://metavo.metacentrum.cz/cs/about/index.html

[52] GAROFOLO, JOHN S., GRAFF, DAVID, PAUL, DOUG a PALLETT, DAVID. *CSR-I (WSJ0) Complete* [online]. B.m.: Linguistic Data Consortium. 30. květen 2007 [vid. 2022-10-04]. Dostupné z: doi:10.35111/EWKM-CG47

[53] WICHERN, Gordon, Joe ANTOGNINI, Michael FLYNN, Licheng Richard ZHU, Emmett MCQUINN, Dwight CROW, Ethan MANILOW a Jonathan Le ROUX. *WHAM!: Extending Speech Separation to Noisy Environments* [online]. B.m.: arXiv. 2. červenec 2019 [vid. 2022-10-04]. Dostupné z: doi:10.48550/arXiv.1907.01160. arXiv:1907.01160 [cs, eess, stat]

[54] HERSHEY, John R., Zhuo CHEN, Jonathan Le ROUX a Shinji WATANABE. *Deep clustering: Discriminative embeddings for segmentation and separation* [online]. B.m.: arXiv. 18. srpen 2015 [vid. 2023-01-20]. Dostupné z: http://arxiv.org/abs/1508.04306. arXiv:1508.04306 [cs, stat]

[55] LUO, Yi a Nima MESGARANI. *TasNet: time-domain audio separation network for real-time, single-channel speech separation* [online]. B.m.: arXiv. 17. duben 2018 [vid. 2023-02-09]. Dostupné z: http://arxiv.org/abs/1711.00541. arXiv:1711.00541 [cs, eess]

[56] WANG, Zhong-Qiu, Jonathan Le ROUX a John R. HERSHEY. Alternative Objective Functions for Deep Clustering. In: *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* [online]. Calgary, AB: IEEE, 2018, s. 686–690 [vid. 2023-02-09]. ISBN 978-1-5386-4658-8. Dostupné z: doi:10.1109/ICASSP.2018.8462507

[57] LUO, Yi a Nima MESGARANI. Real-time Single-channel Dereverberation and Separation with Time-domain Audio Separation Network. In: *Interspeech 2018*: *Interspeech 2018* [online]. B.m.: ISCA, 2018, s. 342–346 [vid. 2023-02-09]. Dostupné z: doi:10.21437/Interspeech.2018-2290

[58] LIU, Yuzhou a DeLiang WANG. *Divide and Conquer: A Deep CASA Approach to Talker-independent Monaural Speaker Separation* [online]. B.m.: arXiv. 24. duben 2019 [vid. 2023-02-09]. Dostupné z: http://arxiv.org/abs/1904.11148. arXiv:1904.11148 [cs, eess]

[59] LUO, Yi, Zhuo CHEN a Takuya YOSHIOKA. *Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation* [online]. B.m.: arXiv. 27. březen 2020 [vid. 2022-10-23]. Dostupné z: http://arxiv.org/abs/1910.06379. arXiv:1910.06379 [cs, eess]

[60] CHEN, Jingjing, Qirong MAO a Dong LIU. *Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation* [online]. B.m.: arXiv. 14. srpen 2020 [vid. 2023-02-09]. Dostupné z: http://arxiv.org/abs/2007.13975. arXiv:2007.13975 [cs, eess]

[61] ZEGHIDOUR, Neil a David GRANGIER. *Wavesplit: End-to-End Speech Separation by Speaker Clustering* [online]. B.m.: arXiv. 2. červenec 2020 [vid. 2023-02-09]. Dostupné z: http://arxiv.org/abs/2002.08933. arXiv:2002.08933 [cs, eess, stat]

[62] LUTATI, Shahar, Eliya NACHMANI a Lior WOLF. *Separate And Diffuse: Using a Pretrained Diffusion Model for Improving Source Separation* [online]. B.m.: arXiv. 25. leden 2023 [vid. 2023-02-09]. Dostupné z: http://arxiv.org/abs/2301.10752. arXiv:2301.10752 [cs, eess]

[63] MACIEJEWSKI, Matthew, Gordon WICHERN, Emmett MCQUINN a Jonathan Le ROUX. WHAMR!: Noisy and Reverberant Single-Channel Speech Separation. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* [online]. Barcelona, Spain: IEEE, 2020, s. 696–700 [vid. 2023-02-09]. ISBN 978-1-5090-6631-5. Dostupné z: doi:10.1109/ICASSP40776.2020.9053327

[64] PARIENTE, Manuel, Samuele CORNELL, Antoine DELEFORGE a Emmanuel VINCENT. *Filterbank design for end-to-end speech separation* [online]. B.m.: arXiv.

28. únor 2020 [vid. 2023-02-09]. Dostupné z: http://arxiv.org/abs/1910.10400. arXiv:1910.10400 [cs, eess]

[65] LI, Kai, Runxuan YANG a Xiaolin HU. *An efficient encoder-decoder architecture with top-down attention for speech separation* [online]. B.m.: arXiv. 24. říjen 2022 [vid. 2023-02-09]. Dostupné z: http://arxiv.org/abs/2209.15200. arXiv:2209.15200 [cs, eess]

[66] NACHMANI, Eliya, Yossi ADI a Lior WOLF. *Voice Separation with an Unknown Number of Multiple Speakers* [online]. B.m.: arXiv. 1. září 2020 [vid. 2023-02-09]. Dostupné z: http://arxiv.org/abs/2003.01531. arXiv:2003.01531 [cs, eess, stat]

[67] KINGMA, Diederik P. a Jimmy BA. *Adam: A Method for Stochastic Optimization* [online]. B.m.: arXiv. 29. leden 2017 [vid. 2023-02-05]. Dostupné z: http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs]

[68] ZEILER, M.D., M. RANZATO, R. MONGA, M. MAO, K. YANG, Q.V. LE, P. NGUYEN, A. SENIOR, V. VANHOUCKE, J. DEAN a G.E. HINTON. On rectified linear units for speech processing. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* [online]. 2013, s. 3517–3521. ISSN 2379-190X. Dostupné z: doi:10.1109/ICASSP.2013.6638312

[69] LIAO, Yuan-Fu, Chia-Yu CHANG, Hak-Khiam TIUN, Huang-Lan SU, Hui-Lu KHOO, Jane S. TSAY, Le-Kun TAN, Peter KANG, Tsun-guan THIANN, Un-Gian IUNN, Jyh-Her YANG a Chih-Neng LIANG. Formosa Speech Recognition Challenge 2020 and Taiwanese Across Taiwan Corpus. In: *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*: *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)* [online]. 2020, s. 65–70. ISSN 2472-7695. Dostupné z: doi:10.1109/O-COCOSDA50338.2020.9295019