ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Fakulta elektrotechnická

Katedra radioelektroniky

Tempo řeči u dětí

Speech Rate of Children

Diplomová práce

Studijní program:	Elektronika a komunikace
Specializace:	Audiovizuální technika a zpracování signálů
Vedoucí práce:	prof. Ing. Roman Čmejla, CSc.

Bc. Jan Vimr

Praha 2020

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Vimr** Jméno: **Jan** Osobní číslo: **457165**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra radioelektroniky**

Studijní program: **Elektronika a komunikace**

Specializace: **Audiovizuální technika a zpracování signálů**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Tempo řeči u dětí

Název diplomové práce anglicky:

Speech Rate of Children

Pokyny pro vypracování:

Na základě akustických analýz vyhodnoťte věkovou závislost tempa řeči v dětských promluvách od zdravých dětí.
a) Proveďte rešerši měření tempa řeči
b) Analyzujte tempo řeči v databázi dětských promluv
c) Implementujte vhodné metody určení tempa řeči a výsledky statisticky vyhodnoťte
d) Analyzujte závislost tempa řeči na typu promluvy a věku a pohlaví dítěte

Seznam doporučené literatury:

[1] V. Aharonson, E. Aharonson, K. Raichlin-Levi, A. Sotzianu, O. Amir, Z. Ovadia-Blechman, 'A real-time phoneme counting algorithm and application for speech rate monitoring' in Journal of Fluency Disorders, Elsevier, vol. 51, pp. 60-68, 2017.
[2] K.J. Logan, C.T. Byrd, E. Mazzocchi, R. Gillam, 'Speaking rate characteristics of elementary-school-aged children who do and do not stutter', Journal of Communication Disorders, 44, pp. 130-147, 2011.
[3] O. Amir, D. Grinfeld, 'Articulation rate in childhood and adolescence: Hebrew speakers', Language and Speech, 54 (2), pp. 225-240, 2011.

Jméno a pracoviště vedoucí(ho) diplomové práce:

prof. Ing. Roman Čmejla, CSc., katedra teorie obvodů FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **12.02.2020** Termín odevzdání diplomové práce: _____

Platnost zadání diplomové práce: **30.09.2021**

prof. Ing. Roman Čmejla, CSc.	doc. Ing. Josef Dobeš, CSc.	prof. Mgr. Petr Páta, Ph.D.
podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

. _____ _____

Datum převzetí zadání Podpis studenta

Čestné prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze dne .....

podpis

Poděkování

Chtěl bych poděkovat prof. Ing. Romanu Čmejlovi, CSc. za vedení této práce, poskytnuté odborné znalosti a vstřícnost při konzultacích. Dále bych rád poděkoval své rodině a přátelům za podporu během studia na ČVUT FEL.

Abstrakt

Tématem této práce je měření tempa řeči v dětských promluvách. V databázi, která obsahovala tři typy promluv od 245 dětí ve věku od 4 do 17 let, bylo manuálně vyhodnoceno tempo řeči. Podařilo se najít silnou korelaci mezi tempem řeči a věkem ve všech třech typech promluv. Dále byl navržen algoritmus pro odhad tempa řeči v promluvě pomocí detekce jednotlivých slabik, založený na krátkodobém výkonu signálu, přítomnosti základní hlasivkové frekvence a počtu průchodů nulou. Výsledky dosažené touto metodou jsou porovnány se dvěma volně dostupnými algoritmy a dále s metodou založenou na využití fonémového rozpoznávače.

Klíčová slova: Tempo řeči, Artikulační tempo, Dětská řeč, Automatická detekce slabik

Abstract

The theme of this thesis is measurement of speech rate in children's speech. Speech rate was manually measured in a database which included three types of utterances from 245 children aged 4 to 17 years. A strong correlation between age and speech rate was found in all type of utterances. An algorithm for speech rate estimation by detecting syllables in utterances was then developed. It is based on short time power of signal, fundamental frequency presence and zero-crossing rate. Results achieved by this this algorithm are compared to two freely available algorithms and a method based on phoneme recognizer.

Keywords: Speech rate, Articulation rate, Children's speech, Automatic syllable detection

Obsah

Seznam zkratek

ANN Artificial neural network - Umělá neuronová síť
ANOVA Analysis of variance - Analýza rozptylu
AR Articulation rate - Artikulační tempo

DCT Diskrétní kosinová transformace
DDK Diadochokineze
DDKR DDK rate - Tempo řeči v DDK promluvě

GMM Gaussian mixture model - Model Gaussovských směsí

IQR Interquartile range - Kvartilové rozpětí

LFME Low frequency modulated energy

ME Mean error
MFCC Mel-frekvenční kepstrální koeficienty

RMSE Root mean square error
RR Recitation rate - Tempo řeči v říkance

SPS Syllable per second - Počet slabik za sekundu
SR Speech rate - Celkové tempo řeči

TCSSBC Temporal correlation and selected sub-band correlation

VEP Vowel end point - Koncový bod samohlásky

ZCR Zero-crossing rate - Počet průchodů nulou

Seznam obrázků

Seznam tabulek

Kapitola 1

Úvod

Tempo řeči je prozodická vlastnost řeči, která se obvykle vyjadřuje počtem řečových jednotek za jednotku času, nejčastěji počtem slabik za sekundu. Často bývá zkoumáno při diagnostice poruch řeči nebo při analýze věkové závislých charakteristik řeči. Odhad tempa řeči lze také využít pro lepší nastavení řečových rozpoznávačů, které mohou mít problémy s velmi rychlými nebo naopak velmi pomalými promluvami.

V dětské řeči dochází vlivem tělesného a duševního vývoje ke změnám některých akustických i prozodických charakteristik. Změny v tempu řeči s rostoucím věkem už byly předmětem mnoha studií, málokdy se však zabývaly dětmi od předškolního věku až po mladistvé. Ve většině z nich se ale podařilo potvrdit, že tempo řeči v dětských promluvách s věkem stoupá. Motivací pro studie tempa řeči u dětí je zejména určení normativních hodnot pro různé věkové kategorie dětí. Ty je možné porovnávat s hodnotami u dětí, které trpí určitou vadu řeči, jako např. koktání [1], [2]. Tempo řeči také může být použito jako jedna z charakteristik při posuzování logopedického věku dítěte [3].

Řada studií také porovnávala tempo řeči v různých typech promluv jako je volná promluva, imitovaná promluva, nebo různé automatické promluvy [4], [5], . Mezi volné promluvy se řadí zejména konverzace a vypravování příběhu podle série předložených obrázků. Imitované promluvy většinou obsahují jednotlivá slova nebo věty opakované po předřečníkovi, což ale zejména u dětí může vést k tomu, že děti kopírují prozódii předřečníka. Mezi automatické promluvy můžeme zařadit ty, kde dítě nemusí dopředu přemýšlet nad tím, co bude říkat. Patří sem jednak diadochokinetické (DDK) promluvy, kde je opakována série slabik, dále repetice určitého slova nebo fráze, a nakonec taky recitace říkanky, kterou dítě umí zpaměti.

Jelikož je manuální určování tempa řeči časově velmi náročné, byla navržena řada metod za účelem automatizace této úlohy. Ty sice nedosahují takové přesnosti, ale nabízí možnost analýzy daleko většího množství promluv než manuální postupy. Některé automatické metody pouze klasifikují promluvy nebo jejich části do diskrétních kategorií jako je pomalá, středně rychlá a rychlá promluva. Jiné se snaží detekovat jednotlivé řečové jednotky jako jsou slabiky nebo fonémy.

V této práci je manuálně vyhodnoceno tempo řeči v databázi promluv, která obsahuje tři různé typy promluv od 245 dětí ve věku od 4 do 17 let. Z nich jsou určena normativní data pro jednotlivé věkové kategorie dětí. Dále jsou zkoumány závislosti tempa řeči na

věku a pohlaví dítěte. V další části je pak navržen algoritmus pro odhad počtu slabičných jader v promluvě a je porovnán s dalšími volně dostupnými algoritmy.

Kapitola 2

Lidská řeč

Obecné informace o lidské řeči prezentované v této kapitole jsou převzaty ze zdrojů [6], [7].

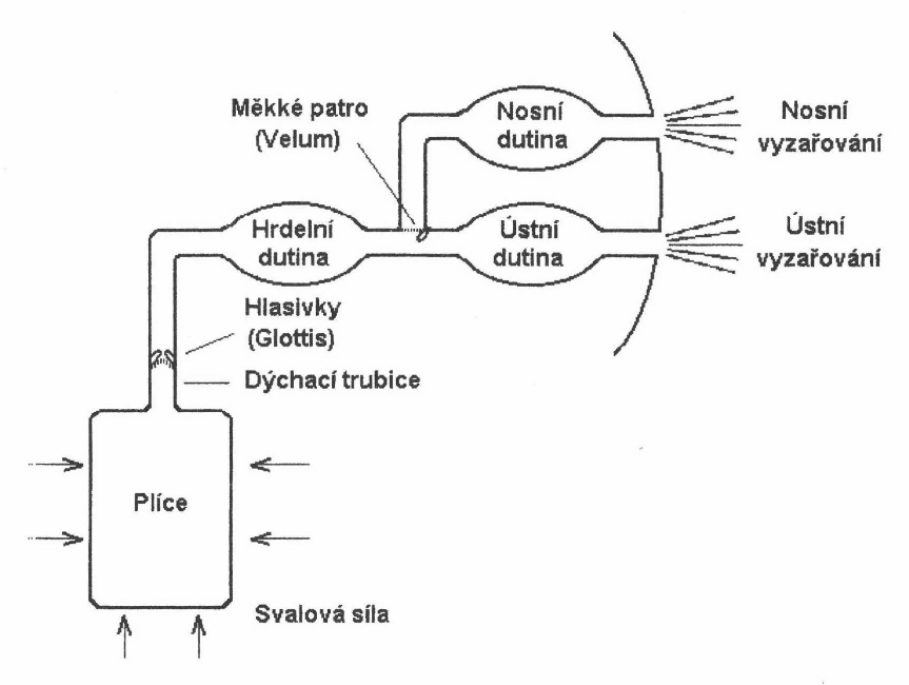
Mluvená řeč je nejstarší a zároveň nejčastěji užívanou formou komunikace mezi lidmi. Přenáší se prostředím ve formě akustických vln. Podstatou akustického signálu je vlnění elastického prostředí v rozsahu slyšitelných kmitočtů. V řečovém signálu je zakódováno několik druhů informace. Vedle samotné akustické složky (amplitudově-frekvenčního časového spektra) bývá z hlediska komunikace nejdůležitější informace lingvistická, protože vyjadřuje význam sdělované myšlenky. Signál dále obsahuje specifické informace o mluvčím, které respektují charakteristiky hlasového traktu řečníka a způsob artikulace (intonaci, rytmus řeči, barvu hlasu atd.), včetně případných anomálií (např. vady řeči), a také informace o emocionálním stavu řečníka.

2.1 Proces vytváření řeči člověkem

Pro vytváření řeči existuje v lidském těle několik skupin orgánů, které se souhrnně nazývají řečové orgány, nebo také artikulační orgány. Produkce řeči zpravidla nebývá primárním úkolem těchto orgánů- jejich základní funkce jsou v lidském těle různé a často spolu navzájem nesouvisejí (např. dýchání, přijímání potravy, cítění). Z hlediska tvorby řeči tyto řečové orgány tvoří hlasový trakt, naznačený na obrázku 2.1, který lze rozdělit na tři základní ústrojí: dechové, hlasové a artikulační.

Dechové ústrojí představuje fundamentální zdroj energie pro řeč. Je tvořeno přívodní dýchací cestou, plícemi a bránicí. Při nádechu dochází k pohybu vzduchu, který tak poskytuje zdroj energie pro řeč. Při výdechu potom v plicích vzniká proud vzduchu, který je základním materiálem pro tvorbu řeči. Proud vzduchu je z plic odváděn průdušnicí, a pak prochází hrtanem a nadhrtanovými dutinami, kde se modifikuje, a jako řečový signál je vyzařován rty do okolního prostoru. Síla výdechového proudu má vliv na sílu hlasu a částečně i na jeho výšku.

Hlasové ústrojí je uloženo v hrtanu, který je s plícemi spojen průdušnicí. Z hlediska tvorby řeči nejdůležitější část hlasového ústrojí tvoří hlasivky, které se nacházejí v hrtanové dutině. Jsou to dvě ostré slizniční řasy, které vedou napříč hrtanem v místě jeho nejužšího průchodu. Jejich základ tvoří hlasový vaz a hlasivkový sval. Prostor mezi hla-



Obrázek 2.1: Schéma hlasového traktu. Převzato z [6].

sivkami tvoří hlasivková štěrbina trojúhelníkového tvaru. Jestliže člověk mlčí, pak hlasivky drží hlasivkovou štěrbinu odkrytou, takže jí může bez odporu procházet vzduch k dýchání. Při vytváření hlasu (fonaci) se stáhnou do tzv. hlasového (fonačního) postavení. Pod tlakem výdechového proudu vzduchu se hlasivky stávají pružnými a začínají kmitat. Střídavě se přitom otevírají a uzavírají, čímž vznikne vzduchová vlna, kde se střídá vždy kvantum hustšího a řidšího vzduchu. Frekvence kmitání hlasivek se označuje f_0, nazývá se základní hlasivková frekvence a odpovídá výšce hlasu. Typicky se pohybuje v rozmezí asi 60 - 400 Hz. Základní hlasivková frekvence tvoří základ pro vznik samohlásek. U znělých souhlásek je napětí hlasivek menší, kmitání je pak méně pravidelné a charakteristika výsledného zvuku již není čistě tónová. Neznělé zvuky jsou naopak tvořeny při klidovém postavení hlasivek, takže základní hlasivkový tón neobsahují.

Artikulační ústrojí se skládá se z nadhrtanových dutin, kam řadíme dutinu hrdelní, ústní a nosní, a z artikulačních orgánů neboli artikulátorů. Ty umožňují vytvářet velké množství různých zvuků, jelikož svým pohybem mění rozměry dutin. Mezi nejvýznamnější artikulátory patří jazyk, rty a měkké patro. Při průchodu základního hlasivkového tónu nadhrtanovými dutinami dochází vlivem rezonance k soustředění akustické energie kolem určitých frekvencí, kterým říkáme formanty. V nosní dutině pak může dojít i k potlačení některých frekvenčních oblastí, tzv. antiformanty. Jednotlivé samohlásky a znělé souhlásky pak získají svůj charakter díky rozdílnému postavení artikulátorů, které vede ke vzniku různých formantových frekvencí. Souhlásky tvořené primárně šumem vznikají při průchodu výdechového proudu vzduchu nadhrtanovými dutinami. Artikulační orgány zde vytvářejí různé překážky, ve formě zúžení, uzavření, nebo naopak uvolnění průchodu. Vznikají tak různé souhlásky v závislosti na charakteru a umístění překážky.

2.2 Samohlásky

Čeština rozlišuje 5 samohlásek (vokálů), které se od sebe liší nastavením hlasového traktu při jejich artikulaci. Jedná se o hlásky /a/, /e/, /i/, o/, /u/ a jejich dlouhé ekvivalenty /á/, /é/, /í/, /ó/, /ú/. Ty se od svých krátkých protějšků liší zejména délkou a jejich artikulace a jejich spektrální vlastnosti jsou v zásadě stejné. Zvláštní postavení v českém jazyce mají dvojhlásky (diftongy). Existují tři, a to /o_u/, /a_u/ a /e_u/, přičemž poslední dvě se objevují pouze ve slovech cizích a přejatých. Při výslovnosti dvojhlásek se artikulační postavení mění z polohy příznačné pro první část dvojhlásky do polohy charakteristické pro druhou část, která však bývá značně potlačena.

2.2.1 Artikulační vlastnosti samohlásek

Na základě pohybu artikulátorů (jazyk, rty, měkké patro, atd.) se mění formantové kmitočty, které značně ovlivňují charakter samohlásky. Nejvýznamnější artikulační charakteristikou českých samohlásek je poloha jazyka. Podle ní rozlišujeme samohlásky přední (palatální), střední (centrální) a zadní (velární) v horizontálním směru a samohlásky vysoké, středové a nízké ve směru vertikálním. Podle postavení rtů rozlišujeme samohlásky zaokrouhlené a nezaokrouhlené. Zaokrouhlení rtů (labializace) je typickým znakem pro zadní samohlásky. Příslušnost českých samohlásek do jednotlivých skupin naznačuje tabulka 2.1. Všechny české samohlásky jsou samohlásky ústní, takže po dobu jejich artikulace uzavírá měkké patro vstup do nosní dutiny.

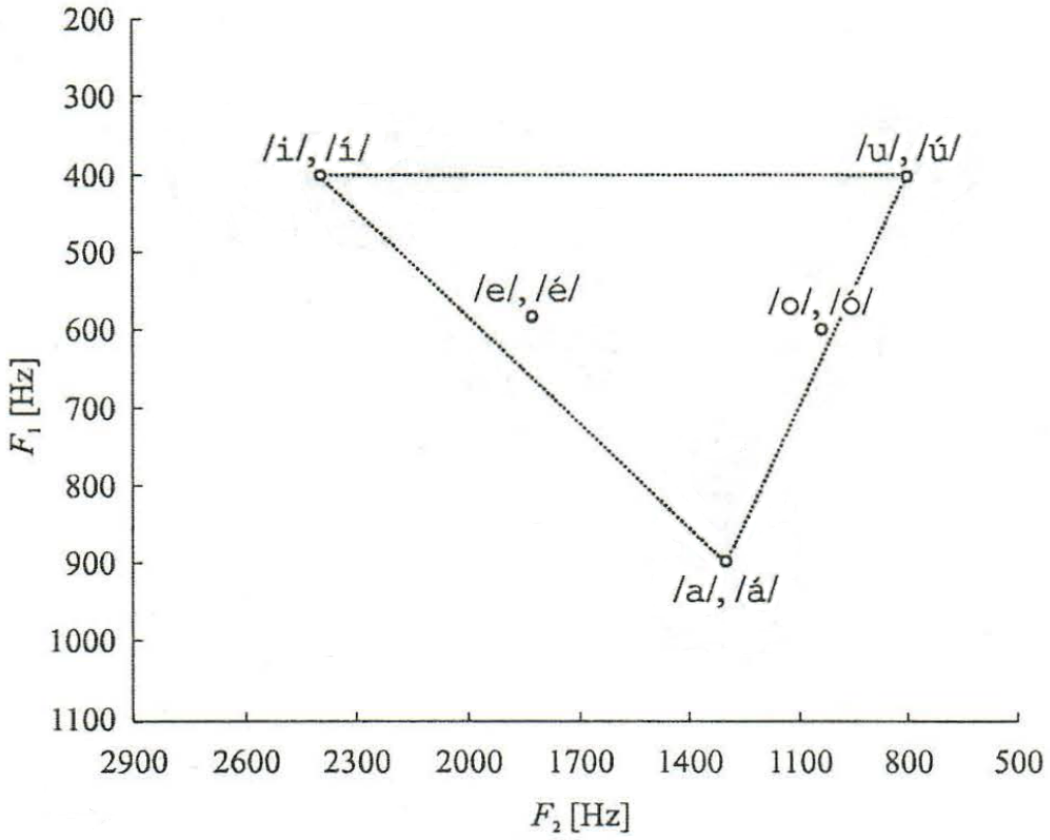
Tabulka 2.1: Dělení českých samohlásek podle artikulace. Převzato z [6].

Rozdělení samohlásek při pohybu jazyka ve směru svislém	Rozdělení samohlásek při pohybu jazyka ve směru vodorovném		
	přední	střední	zadní
vysoké	/í/ /i/		/ú/ /u/
středové	/é/ /e/		/ó/ /o/
nízké		/á/ /a/	
Postavení rtů	nezaokrouhlené		zaokrouhlené

2.2.2 Akustické vlastnosti samohlásek

Z akustického hlediska jsou všechny samohlásky znělé. Jejich akustické signály se vyznačují výraznou kvaziperiodicitou, vysokou amplitudou a delším trváním. Energie samohlásek se soustřeďuje zejména pod 1 kHz a klesá s frekvencí přibližně o 6 dB/oktávu.

Nejvýznamnější akustickou charakteristikou samohlásek jsou formanty, zejména první dva, které rozlišují jednotlivé české samohlásky, viz obrázek 2.2. Frekvence a intenzita formantů závisí na uspořádání, délce, tvaru a průřezu především dutiny ústní, ale i hrdelní. Hlavní anatomické komponenty, které zapříčiňují změny jejich frekvence a intenzity, jsou rty, čelisti a jazyk. S jejich pomocí může řečník měnit velikost nadhrtanového prostoru a tím vytvářet různé samohlásky. Intenzita (amplituda) samohlásek v řeči bývá v porovnání se souhláskami výrazně vyšší. Obecně platí, že intenzita u samohlásek klesá v závislosti na pohybu jazyka směrem nahoru o 5-10 dB, tj. vysoké samohlásky vykazují nižší amplitudu než samohlásky nízké.



Obrázek 2.2: Český vokalický trojúhelník. Převzato z [6].

2.3 Souhlásky

V češtině se rozlišuje celkem 27 souhláskových fonémů: /p/, /b/, /t/, /d/, /ť/, /ď/, /k/, /g/, /f/, /v/, /s/, /z/, /š/, /ž/, /ch/, /h/, /c/, /dz/, /č/, /dž/, /ř/, /m/, /n/, /ň/, /l/, /r/ a /j/. Některé české souhlásky mají více možných alofonických realizací, které se někdy uvádějí zvlášť, např. retozubné /ɱ/, měkkopatrové /ŋ/ nebo neznělé /r̥/. Zatímco samohlásky jsou důležité zejména pro estetické vyznění řeči, přesná výslovnost souhlásek je základní opo-

rou srozumitelnosti řeči. Pro akustické signály souhlásek je typická přítomnost šumu a menší amplituda v porovnání se samohláskami. Shluky jednotlivých souhlásek lze proto poměrně jednoduše rozeznat od samohlásek, ale rozlišit od sebe navzájem signály jednotlivých souhlásek už může být složitější.

2.3.1 Artikulační vlastnosti souhlásek

Při artikulaci souhlásek naráží výdechový proudu vzduchu v nadhrtanových dutinách na překážky nebo zúžení hlasového traktu. Podle místa a způsobu vytvoření překážky vznikají různé souhlásky, tedy zvuky, jejichž podstatou je šum. U českých souhlásek jsou důležité především čtyři artikulační charakteristiky: způsob artikulace, místo artikulace, postavení měkkého patra a činnost hlasivek. Rozdělení českých souhlásek podle jednotlivých charakteristik je naznačeno v tabulce 2.2.

Tabulka 2.2: Dělení českých souhlásek podle artikulace. Převzato z [6].

Místo artikulace \ Způsob artikulace		závěrové			polozávěrové		úžinové					
		ústní		nosní			středové		kmitavé		bokové	glidy
retné	obouretné	p	b	m								
	retozubné			ɱ			f	v				
dásňové	předodásňové	t	d	n	c	d͡z	s	z	r̝	ř	l	
										r		
	zadodásňové				č	d͡ž	š	ž				
patrové	tvrdopatrové	ť	ď	ň								j
	měkkopatrové	k	g	ŋ			x					
hrtanové								h				
Sonory / Šumy		Šu	Šu	So	Šu	Šu	Šu	Šu	Šu	So	So	So
Neznělé / Znělé		N	Z	Z	N	Z	N	Z	N	Z	Z	Z

Podle způsobu artikulace můžeme souhlásky rozdělit na následující skupiny. Závěrové (okluzivy, explozivy) vznikají při úplném přiblížení artikulačních orgánů čímž vzniká tzv. závěr. Ten se projeví jako krátká pauza (okluze), následuje náhlé uvolnění překážky, čímž dojde k úniku nahromaděného vzduchu (exploze). Úžinové nebo též třené souhlásky (konstriktivy, frikativy) vzniknou, pokud artikulační orgány vytvoří úžinu, kde při průchodu výdechového proudu vzniká třecí šum. Do této kategorie spadají i frikativy kmitavé (vibranty), při jejichž vzniku se určitá část artikulátorů rozkmitá a dodá výslednému zvuku částečně tónový charakter, dále frikativy bokové (laterální), u nichž vzduch

uniká bokem místo typické cesty středem, a specifickou frikativou jsou tzv. approximanty (glidy), které jsou typické stejným postavením artikulátorů jako při vzniku samohlásek, ale doba jejich trvání je oproti samohláskám výrazně kratší. U polozávěrových nebo také polotřených souhlásek (semiokluzivy, afrikáty), se postupně vyskytují oba typy překážek, nejprve se uzavře cesta, která se pak postupně otevírá. Závěru odpovídá kratická pauza jako u exploziv, následovaná třecím šumem podobným frikativám.

Druhým hlediskem, podle něhož můžeme rozlišovat souhlásky, je místo artikulace, tedy oblast, v níž dojde k nejvýraznějšímu zúžení či uzavření hlasového traktu. Podle tohoto hlediska dělíme souhlásky do následujících skupin. Retné souhlásky (labiály) vznikají při uzavření rtů nebo přitisknutí horních řezáků na spodní ret, podle čehož se rozlišují souhlásky obouretné a retozubné. Dásňové souhlásky (alveoláry) vznikají při přitisknutí jazyka k dásni a lze je ještě rozdělit na předodásňové a zadodásňové. Patrové souhlásky, mezi které patří tvrdopatrové (palatály) a měkkopatrové (veláry), jsou vytvářeny při přitisknutí jazyka na patro. Poslední skupinou jsou hrtanové souhlásky (laryngály, glotály), které vznikají přímo v hrtanu, bez další modifikace v nadhrtanových dutinách.

Třetím hlediskem je postavení měkkého patra. Pokud se měkké patro uvolní, může se do artikulačního procesu zapojit i nosní dutina. V tom případě je zvuk souhlásky obohacen o poměrně silnou tónovou složku nosní rezonance. Hlásky vzniklé tímto způsobem se nazývají nazální explozivy či jen nazály.

Posledním významným hlediskem, podle něhož můžeme souhlásky rozlišovat, je činnost hlasivek při jejich artikulaci. Je-li foném tvořen za spoluúčasti základního hlasivkového tónu (znělé souhlásky), v nadhrtanových dutinách vznikají rezonancí další tónové složky ve zvuku hlásky podobně jako u samohlásek. Prochází-li vzduch hlasivkami volně (při klidovém postavení hlasivek), zvuky se vytvářejí pouze pomocí překážek v nadhrtanových dutinách (neznělé souhlásky). Některé souhlásky lze zařadit do dvojic, které mají podobnou artikulaci a liší se pouze znělostí, tzv. souhlásky párové.

2.3.2 Akustické vlastnosti souhlásek

Souhlásky mají proti samohláskám výrazně kratší dobu trvání, která navíc není příliš ovlivněna aktuálním tempem řeči. Dále mají typicky menší amplitudu a jejich typickým rysem je přítomnost šumu. Základní akustická charakteristika pro rozlišení jednotlivých souhlásek je poměr šumu a tónové složky. Na základě tohoto parametru lze souhlásky rozdělit na sonorní a šumové.

Sonorní souhlásky (sonory) se vyznačují převládající tónovou složkou. Jsou vždy znělé a díky své převažující harmonické struktuře mohou tvořit jádro slabiky. Jejich akustické signály jsou podobné samohláskám, ale vyznačují se nižší amplitudou a menší energií, která je stejně jako u samohlásek soustředěna na nižších kmitočtech. Mezi sonory patří všechny nazály, při nichž tónová složka vzniká vlivem rezonance nosní dutiny, dále likvidy a glidy.

Šumové souhlásky mohou být znělé i neznělé, ale vždy je patrná silná šumová složka. Řadíme mezi ně všechny párové souhlásky. V praxi se pro /s/, /z/, /š/ a /ž/ vžil termín sykavky a pro /c/, /č/, /dz/ a /dž/ název polosykavky. Šumové frikativy jsou spíše neperiodické a kvůli silnému zúžení hlasového traktu mají mnohem menší amplitudu. Energie je

soustředěna na vyšších kmitočtech. Ve spektru znělých frikativ najdeme slabou formantovou strukturu, u neznělých frikativ se formanty nevyskytují. Šumové explozivy jsou přechodové zvuky, takže jsou artikulačně i akusticky složitější. Okluze se projeví jako ticho a exploze jako krátký výbuch šumu. Afrikáty obsahují postupně okluzi, explozi i výraznou frikci, takže jejich akustické signály připomínají sekvenci explozivy a frikativy. Např. /c/ se artikuluje podobně jako spojení /ts/, apod. Exploze však nebývá tak výrazná a doba trvání bývá kratší.

2.4 Slabiky

Slabiky jsou základní řečové jednotky složené z fonémů. V českém jazyce jsou slabiky nejčastěji tvořeny dvěma nebo třemi fonémy. Ze slabik jsou pak dále tvořena slova. Základem slabiky je slabičné jádro, které je tvořeno samohláskou nebo slabikotvornou souhláskou. V češtině poměrně často tvoří jádro slabiky likvidy /l/ a /r/ a v ojedinělých případech mohou jádro slabiky tvořit také nazály /m/ a /n/. Slabičné jádro zpravidla obklopují tzv. svahy, které mohou být tvořeny jednou nebo více souhláskami nebo nemusí existovat vůbec. Nejčastější slabičné typy v českém jazyce jsou: KV, KVK, KKV, V, KKVK a VK, kde K značí souhlásku (konsonant) a V značí samohlásku (vokál) nebo slabikotvornou souhlásku.

Slabiky nesou většinu informací o prozódii, jako je přízvuk, časování nebo intonace, a proto jsou nejčastěji využívány pro automatický odhad prozodických charakteristik řeči. V této práci jsou slabiky zkoumány zejména v souvislosti s určováním tempa řeči v promluvách, a to jak při manuálním zpracování promluv, tak i při automatickém odhadu tempa řeči.

Kapitola 3

Tempo řeči

Tempo řeči je prozodická vlastnost, která se obvykle vyjadřuje počtem řečových jednotek za jednotku času. Typicky počtem slov za minutu nebo také počtem slabik či fonémů za sekundu. Tempo řeči má významný vliv na délku jednotlivých segmentů řeči. Daleko výrazněji jsou ovlivněny samohlásky než souhlásky, např. délka exploziv se mění jen minimálně [6]. Proto je pro posuzování tempa řeči nejčastěji používanou řečovou jednotkou slabika, jejíž jádro tvoří nejčastěji právě samohláska.

Ve studiích zabývajících se problematikou stanovení řečového tempa se objevují dva různé přístupy jak tempo stanovit. První možností je manuální určení počtu slabik v promluvě, kdy jsou slabiky nebo slova spočítány při poslechu nahrávky, nebo je ručně pořízen přepis nahrávky, ve kterém jsou následně spočítány vybrané řečové jednotky. Druhou možností jsou automatické odhady tempa řeči, které jsou rozebrány v podkapitole 3.2 a některé vybrané volně dostupné metody, které byly použity v této práci, jsou podrobně popsány v podkapitole 4.3. Výhodou manuálního určování počtu řečových jednotek je vysoká přesnost. Hlavní nevýhodou je velká časová náročnost, která téměř vylučuje možnost využití tohoto přístupu na obsáhlejší databáze.

Rozlišují se dvě různé veličiny související s tempem řeči. Jedná se o celkové tempo řeči SR (z anglického „speech rate") a artikulační tempo AR (z anglického „articulation rate"). Pokud je tempo určováno z celé promluvy bez ohledu na pauzy a nespojitosti, pak se jedná o celkové tempo řeči SR. Lze ho určit podle vzorce:

$$\text{SR} = \frac{N_{syl}}{T} \tag{3.1}$$

kde N_{syl} je celkový počet slabik v promluvě a t je celková doba trvání promluvy.

Naproti tomu, pokud je tempo stanoveno výhradně z plynulých úseků promluvy, pak se typicky mluví o artikulačním tempu AR. Plynulé úseky bývají obvykle definovány jako části promluvy, které komunikují určitou myšlenku, nedochází v nich k žádné nespojitosti a neobsahují pauzy delší než 250 ms [8]. V této práci je artikulační tempo určováno z celé promluvy, kde jsou vynechány všechny pauzy delší než 250 ms, podle vzorce:

$$\text{AR} = \frac{N_{syl}}{T - T_p} \tag{3.2}$$

kde N_{syl} značí celkový počet slabik v promluvě, T je celková doba trvání promluvy v

sekundách a T_p je celková doba trvání pauz v promluvě, a to jednak nevyplněných pauz, tedy ticha, a jednak vyplněných pauz, nejčastěji tvořených hezitacemi jako např. „ehm", „em". Celkové tempo řeči je typicky vnímáno jako ukazatel schopnosti komunikovat určitou myšlenku, tedy záleží jak na artikulačních schopnostech jedince, tak na slovní zásobě, schopnosti sestavit větu apod. Naproti artikulační tempo je považováno za ukazatel čistě artikulačních schopností a veškeré ostatní vlivy jsou vyloučeny [9].

3.1 Tempo řeči v dětských promluvách

Vlivem tělesného a duševního vývoje člověka během dětství a dospívání dochází ke změnám některých parametrů, např. k vychýlení formantových kmitočtů nebo k poklesu základní hlasivkové frekvence [10]. Ovlivněny jsou ale i prozodické charakteristiky, mezi které spadá i tempo řeči.

Stanovením tempa řeči v dětských promluvách se již zabývalo mnoho studií. Jejich motivací bylo nejčastěji určení normativních hodnot v jednotlivých věkových kategoriích a zkoumání jejich věkové závislosti. Výsledky pak mohou být využity pro srovnání s dětmi, které trpí poruchou řeči, jako je například koktání. Studie se typicky zabývaly buď to dětmi v předškolním věku, nebo dětmi na prvním stupni základní školy. Pouze několik prací si kladlo za cíl popsat typické hodnoty tempa řeči pro všechny věkové kategorie od předškolních dětí až po mladistvé. Navíc se v různých studiích liší i věkové rozdíly mezi skupinami dětí, nejčastěji je rozdíl jeden nebo dva roky.

Jednotlivé studie se často liší v řečových úlohách, ve kterých je tempo řeči zkoumáno a často se liší také v terminologii. Za volné promluvy se dají považovat úlohy typu vypravování podle předložených obrázků a strukturovaná konverzace s druhým mluvčím, který nejčastěji pokládá otevřené otázky na které je očekávána obsáhlejší odpověď. Dále se pak objevují tzv. imitované promluvy, kdy dítě opakuje promluvu po předřečníkovi. Často se ve studiích objevují i různé rytmické promluvy, kdy má dítě za úkol říci známou říkanku, nebo například dokola opakovat nějaké slovo nebo frázi. Speciálním případem je diadochokinetická (DDK) promluva, kdy má dítě za úkol opakovat skupinu slabik, speciálně zvolených tak, aby byla každá souhláska artikulována v jiné části hlasového traktu [11].

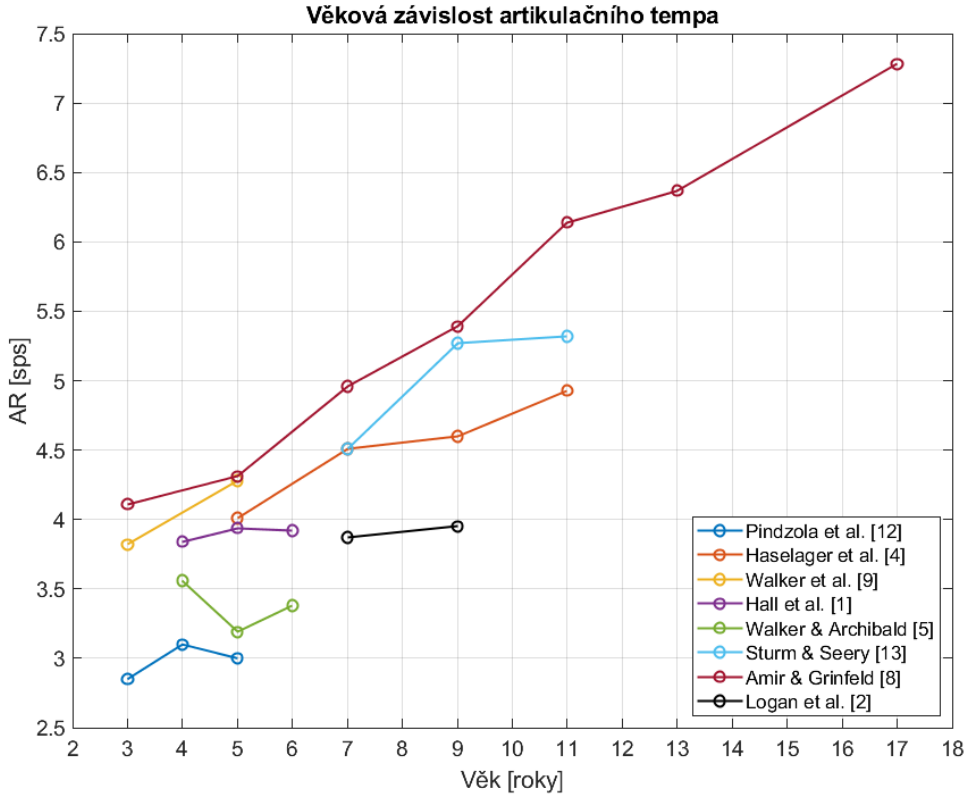
Porovnání výsledků vybraných studií zabývajících se parametrem AR je naznačeno na obrázku 3.1, kde jsou vždy vyobrazeny výsledky jedné vybrané řečové úlohy z dané studie. Byla vybrána vždy úloha, která se co nejblíže podobala úloze, která je analyzována v této práci, tedy vypravování příběhu podle série obrázků. Ve studiích, kde se tato úloha nevyskytovala byla vybrána co nejpodobnější úloha, kterou lze považovat za volnou promluvu, např. konverzace. Tabulka 3.1 obsahuje základní informace o jednotlivých studiích a mimo jiné jsou v ní přehledně vypsány vybrané úlohy zobrazené na obrázku 3.1. Podobně jsou na obrázku 3.2 porovnány výsledky studií zabývajících se parametrem SR a dodatečné informace, včetně vybrané řečové úlohy, jsou vypsány v tabulce 3.2.

Studie zkoumající tempo řeči u předškolních dětí se často liší v závěrech, zda existuje věková závislost tempa řeči už v takto útlém věku. V práci [12] bylo zkoumáno tempo řeči i artikulační tempo ve strukturované konverzaci u anglicky mluvících dětí ve třech skupinách odpovídajících věku 3, 4 a 5 let. V každé skupině bylo 10 dětí (6 chlapců a 4

Tabulka 3.1: Přehled vybraných studií - AR

Studie	Jazyk	Počet dětí	Vybraná úloha
Pindzola et al. [12]	aj (USA)	3 sk. po 10	konv. + vypr.
Haselager et al. [4]	nizozemština	4 sk. po 10	vypravování
Walker et al. [9]	aj (Kanada)	2 sk. po 20	vypravování
Hall et al. [1]	aj (USA)	3 roky, 8 dětí	konverzace
Walker & Archibald [5]	aj (Kanada)	3 roky, 16 dětí	vypravování
Sturm & Seery [13]	aj (USA)	3 sk. po 12	vypravování
Amir & Grinfeld [8]	hebrejština	7 sk. po 20	vypravování
Logan et al. [2]	aj (USA)	2 sk. po 17	vypravování

Význam zkratek: konv. = konverzace, vypr. = vypravování podle obrázku



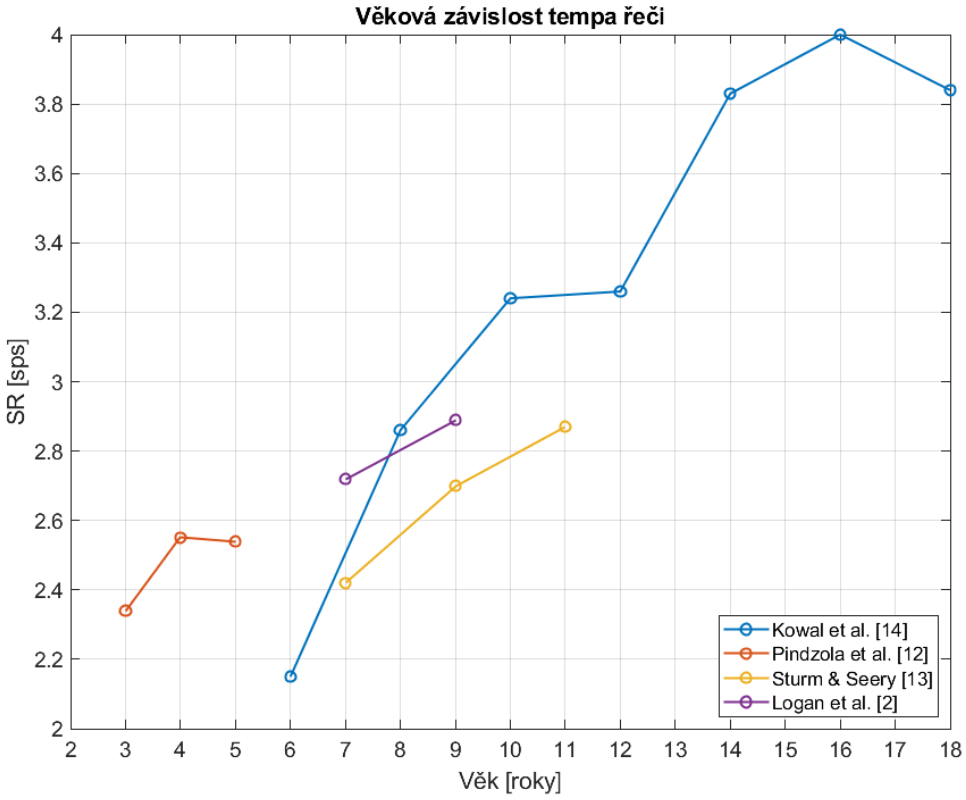
Obrázek 3.1: Výsledky vybraných studií - AR

dívky). Pomocí analýzy rozptylu (ANOVA) se nepodařilo prokázat statisticky významné rozdíly mezi skupinami. Výsledky naznačují zvýšení středních hodnot AR a SR mezi věkem 3 a 4 let, ale mezi věkem 4 a 5 let už nikoliv, naopak dochází k mírnému poklesu. Naproti tomu ve studii [9], kde bylo zkoumáno artikulační tempo tříletých a pětiletých anglicky mluvících dětí ve skupinách po 20 (10 dívek a 10 chlapců), vykazují data ros-

Tabulka 3.2: Přehled vybraných studií - SR

Studie	Jazyk	Počet dětí	Vybraná úloha
Kowal et al. [14]	aj (USA)	7 sk. po 24	vypravování
Pindzola el al. [12]	aj (USA)	3 sk. po 10	konv. + vypr.
Sturm & Seery [13]	aj (USA)	3 sk. po 12	vypravování
Logan et al. [2]	aj (USA)	2 sk. po 17	vypravování

Význam zkratek: konv. = konverzace, vypr. = vypravování podle obrázku



Obrázek 3.2: Výsledky vybraných studií - SR

toucí trend s věkem. Byla zde analyzována jednak spontánní promluva (vypravování) a jednak imitovaná promluva, která byla opakována po předřečníkovi.

Dále bylo publikováno několik podélných studií, kdy byly dětské promluvy nahrávány od jedné skupiny dětí opakovaně v přibližně ročních intervalech. Ve studii [1] byly porovnávány AR zdravých dětí, dětí které koktají a dětí které se z koktání uzdravují, ve věku 3 až 5 let. Všechny děti byly anglicky mluvící. Výsledky studie naznačují, že dochází k nárůstu AR s věkem, ale významný rozdíl byl zaznamenán pouze mezi promluvami z prvního a druhého nahrávání, tedy mezi 3. a 4. rokem života dítěte, ale mezi druhou a třetí návštěvou už nikoliv. Rostoucí trend AR s věkem se naopak nepodařilo potvrdit v práci

[5], kde byly zkoumány 4 typy promluv u skupiny anglicky mluvících dětí postupně ve věku 4, 5 a 6 let. V šesti letech děti sice mluvily v průměru rychleji než ve čtyřech, ale rozdíl nebyl významný. Navíc parametr AR v 5 letech byl ve všech typech promluv nižší než ve 4 a 6 letech.

Další skupina studií se zabývala dětmi v mladším školním věku. V práci [4] je zkoumán parametr AR ve volné a diadochokinetické (DDK) promluvě u věkových skupin 5, 7, 9 a 11 let nizozemsky mluvících dětí. U obou typů promluvy AR narůstá s věkem, výrazněji u DDK promluvy. Ve studii [13] byl zkoumán AR i SR jak ve skupinách odpovídajících přibližně věkům 7, 9 a 11 let v konverzaci a vypravování u anglicky mluvících dětí. Výsledky naznačují nárůst AR i SR mezi skupinami ve věku 7 a 9 let. Výsledky pro skupiny ve věku 9 a 11 let jsou velmi podobné. V práci [2] jsou porovnávány veličiny SR a AR plynule mluvících dětí a dětí které koktají. Všechny děti jsou rodilí mluvčí americké angličtiny. Zároveň s tím byly děti rozděleny na mladší a starší, přibližně 7 resp. 9 let staré. Byly porovnávány 3 různé úlohy, a to konverzace, vypravování a modelová věta, která byla opakována po předřečníkovi. Výsledky naznačují statisticky významné rozdíly mezi věkovými skupinami.

Poslední skupinou studií zabývajících se tempem řeči v dětských promluvách jsou ty, které porovnávají tempo řeči v poměrně širokém věkovém rozpětí. Nejstarší studie z této skupiny je [14], kde jsou zkoumány promluvy od anglicky mluvících dětí a mladistvých ve věkových kategoriích přibližně odpovídajících stáří 6, 8, 10, 12, 14, 16 a 18 let. Pomocí analýzy rozptylu se podařilo potvrdit statisticky významný vliv věku na tempo řeči. Ve studii [8] je zkoumáno artikulační tempo hebrejsky mluvících dětí a mladistvých ve skupinách 3, 5, 7, 9, 11, 13 a 17 let. Děti měly 2 různé úlohy. První byla strukturovaná konverzace a druhou bylo vypravování příběhu podle předložených obrázků. Podařilo se najít statisticky významný vliv věkové skupiny na AR.

Rozdíly mezi výsledky jednotlivých studií mohou být způsobeny mnoha faktory. Jednak byly různé studie prováděny u dětí mluvících různými jazyky, navíc i v rámci jednoho jazyka mohou být ovlivněny různými místními dialekty, např. rozdíl mezi americkou a kanadskou angličtinou. Dále se studie liší v počtu zkoumaných dětí v jednotlivých věkových kategoriích a v délce pořízených nahrávek. V neposlední řadě se liší také přesné zadání a obsah jednotlivých řečových úloh. A nakonec mohou mít vliv i podmínky, za kterých jsou nahrávky pořizovány, např. zda se dítě nachází ve známém prostředí nebo přítomnost známé osoby při nahrávání. Např. v práci [1] si děti během konverzace hrály s plastelínou, v práci [5] si dítě s výzkumným pracovníkem nejprve několik minut povídalo, aby nebylo nervózní, apod.

3.2 Automatický odhad tempa řeči

Existují dva základní přístupy k hodnocení tempa řeči. První přístup je klasifikace promluv do diskrétních kategorií, např. pomalé, střední a rychlé. Tento přístup lze využít například jako předstupeň řečového rozpoznávače, který obsahuje několik různých modelů pro různé tempo řeči. Takto je možné přizpůsobit například krok při segmentaci signálu [15]. Hranice těchto diskrétních kategorií však nejsou nijak pevně definovány.

Druhou možností je pokusit se co nejpřesněji odhadnout skutečné hodnoty vyslovených řečových jednotek v promluvě. Tempo řeči je nejčastěji udáváno jako počet slabik za sekundu. Tím pádem lze úlohu odhadu tempa řeči zjednodušit na detekci počtu slabičných jader v promluvě. Ty lze poměrně snadno detekovat, jelikož jsou nejčastěji tvořeny samohláskami, které se v akustických vlastnostech výrazně odlišují od souhlásek. Ze známého počtu slabik a délky promluvy lze tempo řeči stanovit podle vzorce 3.1. Pro signál samohlásky je typická vyšší energie než pro souhlásky a přítomnost základní hlasivkové frekvence. Proto je základem mnoha algoritmů hledání lokálních maxim ve vyhlazeném průběhu energie signálu, a to buď v celém frekvenčním pásmu signálu, nebo pouze ve vybraných pásmech.

V této sekci jsou dále popsány některé vybrané algoritmy pro odhad tempa řeči v promluvě, které byly v minulosti publikovány. Pouze několik z nich je volně dostupných. Z těch byly v této práci testovány dva algoritmy, a to jmenovitě Praat skript [16] a Theta Oscilátor [17]. Praat skript detekuje slabičná jádra v promluvě na základě průběhu intenzity a přítomnosti základní hlasivkové frekvence. Theta Oscilátor využívá průběh veličiny zvané sonorita. Oba algoritmy jsou podrobněji popsány v podkapitole 4.3.

V práci [18] je získán průběh amplitudy v pásmově filtrovaném signálu. Ve vyhlazeném průběhu amplitudy jsou pak hledána lokální maxima. Za slabičná jádra jsou pak vybrány ty vrcholy, které jsou vyšší než stanovený práh a jsou od sebe vzdáleny minimálně 88 ms. Algoritmus [19] využívá průběh veličiny modifikovaná hlasitost, která je definována jako rozdíl mezi hlasitostí v nižším a vyšším kmitočtovém pásmu. Ve vyhlazeném průběhu modifikované hlasitosti jsou pak hledány vrcholy. Pro zvýšení přesnosti je sledován ještě parametr počet průchodů nulou, který je výrazný zejména v neznělých úsecích, tudíž vylučuje přítomnost samohlásky.

Klasifikace na diskrétní kategorie je použita v práci [20], kde byly natrénovány 3 modely Gaussovských směsí (GMM) pro pomalou, střední resp. rychlou řeč. Na základě modelů jsou určeny pravděpodobnosti příslušnosti k jednotlivým skupinám. Samotné rozdělení do kategorií je provedeno na základě nejvyšší pravděpodobnosti. Pro odhad tempa řeči na spojité škále je možné přidat zobrazovací funkci v podobě neuronové sítě. Jejím vstupem jsou 3 pravděpodobnosti příslušnosti ke kategoriím a výstupem je odhad tempa řeči.

V práci [21] je využita energie ve vybraných subpásmech signálu a přítomnost základní hlasivkové frekvence. Je spočítán průběh signálu TCSSBC (Temporal correlation and selected sub-band correlation), ve kterém jsou hledány lokální maxima. Tento algoritmus byl následně modifikován ve studii [22], kde byl signál rozdělen na úseky tak, aby každý obsahoval právě jeden vrchol v průběhu TCSSBC. Na tyto úseky není aplikováno prahování, jako v předchozím případě, ale je natrénován klasifikátor, který označí daný vrchol buďto jako slabiku nebo nikoliv. Jsou ošetřeny i případy, kdy je jedna slabika tvořena více vrcholy.

Algoritmus [23] využívá příznak nazvaný LFME (Low frequency modulated energy). Vychází z předpokladu, že většina energie samohlásek je na nižších kmitočtech. Jsou vypočteny energetické obálky signálu ve 4 subpásmech. Příznak LFME je získán jako součet energetických obálek tří vyšších pásem, který je vynásoben kvadrátem energetické obálky nejnižšího pásma. V průběhu LFME jsou následně hledány vrcholy, které splňují

zavedená pravidla.

V práci [15] je použita neuronová síť, která klasifikuje jednotlivé segmenty řeči na čtyři kategorie, a to pomalá, průměrná a rychlá řeč, nebo ticho. Vstupem do neuronové sítě jsou mel-frekvenční kepstrální koeficienty (MFCC). S klasifikátorem ve spojení s řečovým rozpoznávačem bylo provedeno několik experimentů, které naznačují, že výsledky rozpoznávače se zlepší, pokud je použita adaptace kroku při segmentaci signálu v závislosti na kategorii tempa řeči. V práci [24] je popsán algoritmus, který se snaží v promluvě detekovat koncové body slabik – VEP (Vowel end point). Na základě těchto bodů je pak řeč segmentována na jednotlivé slabiky.

Kapitola 4

Metodika

V následující kapitole je nejprve popsáno manuální vyhodnocení tempa řeči v databázi dětských promluv, která obsahovala tři typy promluv od každého dítěte, a to volnou promluvu, říkanku a DDK promluvu. Dále jsou zde blíže popsány vybrané metody pro automatický odhad tempa řeči v promluvách, které byly testovány na výše zmíněné databázi promluv. Výsledky jejich testování jsou pak popsány v další kapitole, stejně jako vyhodnocení časových závislostí tempa řeči v dětských promluvách.

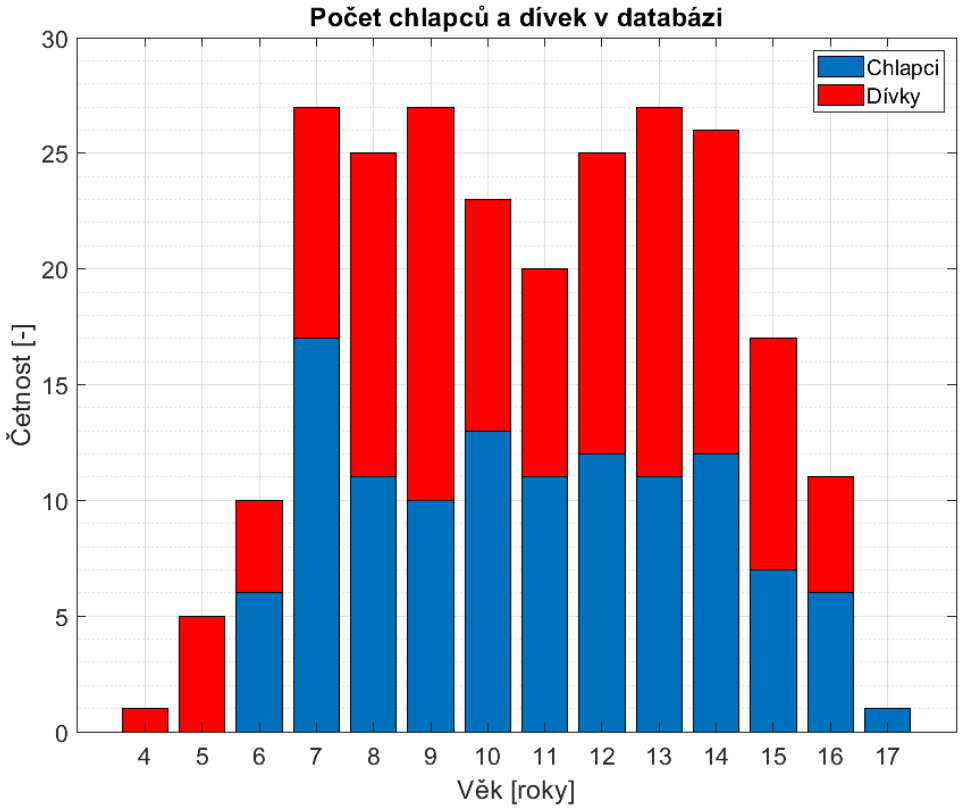
4.1 Databáze promluv

Pro účely této práce byly použity nahrané promluvy od 245 dětí (117 chlapců a 128 dívek) ve věku od 4 do 17 let. Zastoupení chlapců a dívek v jednotlivých věkových kategoriích je ilustrováno na obrázku 4.1. Dataset použitý v této práci obsahuje tři typy promluv od každého dítěte. Prvním typem promluvy je vyprávění příběhu podle série předložených obrázků. Druhý typ promluvy byla říkanka, kterou děti uměly zpravidla odříkat zpaměti. Třetím typem byla diadochokinetická (DDK) promluva, kde děti dostali za úkol co nejrychleji opakovat slabiky /pa/-/ta/-/ka/. Informace o délce promluv a počtu slabik, které obsahují, jsou uvedeny v tabulce 4.1. Jelikož data nemají normální rozdělení, jsou v tabulce uvedeny parametry medián, kvartilové rozpětí (IQR) a rozsah, tedy minimální a maximální hodnota.

Tabulka 4.1: Parametry použité databáze promluv

Typ promluvy	Parametr	Medián (IQR)	Rozsah
Volná promluva	Délka promluvy [s]	19,6 (7,2)	8,8 - 51,6
	Počet slabik [-]	51 (17)	22 - 149
DDK promluva	Délka promluvy [s]	1,7 (0,4)	0,9 - 4,7
	Počet slabik [-]	12 (3)	9 - 12
Říkanka	Délka promluvy [s]	6,8 (3,2)	3,9 - 20,4
	Počet slabik [-]	26 (0)	23 - 32

Promluvy použité v této práci jsou součástí databáze pořízené v roce 2010 za úče-



Obrázek 4.1: Zastoupení chlapců a dívek v databází

lem vyhodnocení věkové závislosti různých charakteristik dětské řeči. Cílem bylo, aby pro všechny zastoupené věkové kategorie dětí byly jednotné promluvy, a aby tím pádem bylo možné srovnávat věkově závislé charakteristiky od dětí v předškolním věku až po mladistvé. Obsah databáze byl konzultován s pracovníky foniatrické kliniky 1. LF UK a VFN. Je zaměřen na akusticko-fonetické jevy, u nichž je předpokládána, nebo již byla potvrzena, věková závislost [3].

Pro nahrávání byl použit kondenzátorový náhlavní mikrofon Bayerdynamic Opus 55.09 MK II SC s velmi plochou frekvenční charakteristikou a byla použita rovněž přiložená protivětrná ochrana proti poryvům některých exploziv. Membrána mikrofonu byla u všech dětí shodně umístěna v rovině obličeje 2 cm vlevo od levého koutku úst. Signál byl digitalizován rekordérem Roland Edirol R-09HR při vzorkovací frekvenci 44,1 kHz a bitové hloubce 16 bitů. Ovladače rekordéru byly nastaveny při každém nahrávání stejně, takže byly všechny děti nahrány se stejným zesílením [3]. Nahrávky byly uloženy ve formátu „.wav", zpravidla v mono stopě, s výjimkou DDK promluvy, která byla uložena ve stereo stopě.

V původní databázi bylo nahráno celkem sedm úloh, z nichž jsou v této práci použity tři, a to volná promluva, říkanka a DDK promluva. Nepoužité nahrávky obsahovaly jednotlivá slova opakovaná po mluvčím nebo podle obrázků, prodloužené fonace samohlásek, prodloužené sykavky a čtený text, který byl nahrán pouze u starších dětí. Některé

věkově závislé parametry dětské řeči zkoumané na této databázi jsou popsány v práci [3]. Kromě jiných parametrů, tam byla vyhodnocována i věková závislost tempa řeči v říkance, ale postup při jeho určování se mírně lišil od postupu použitého v této práci, viz 4.2.2.

4.1.1 Volná promluva

Při nahrávání byla dětem předložena série obrázků, které zobrazovaly jednotlivé činnosti malého chlapce před cestou do školy, jako např. čištění zubů, oblékání, snídání, apod. Děti tak měly možnost použít vlastní slova, ale přitom měly stanovený příběh, kterého se měly držet. V nahrávkách se objevuje velké množství pauz, ať už vyplněných či nevyplněných. Často se také objevují přeřeknutí, zakoktání a opakovaní jednotlivých slov nebo i celých vět. Některé úseky jsou poměrně těžko srozumitelné.

4.1.2 Říkanka

Děti měly za úkol v libovolném tempu zarecitovat říkanku, kterou většinou znaly zpaměti. Nahrávky tak neobsahují téměř žádné nespojitosti, až na několik výjimek, kdy dítě mírně zaváhalo nebo se přeřeklo apod. Jednalo se o říkanku:

„En ten týky,
dva špalíky,
čert vyletěl z elektriky.
Bez klobouku bos,
natloukl si nos."

4.1.3 DDK promluva

Pro posouzení motorických schopností artikulačního ústrojí bývá ve foniatrické praxi často používána diadochokinetická promluva. Ta je založená na co nejrychlejším opakování sekvence slabik, konkrétně kombinací souhláska - samohláska. Aby se v úloze promítla motorická zdatnost celého artikulačního ústrojí, používá se typicky více souhlásek s různým místem artikulace [11]. V databázi, která byla pro tuto práci k dispozici, byla použita sekvence slabik /pa/-/ta/-/ka/.

4.2 Manuální analýza promluv

Pro vyhodnocení tempa řeči v promluvách bylo nutné stanovit počet slabik, které se v jednotlivých nahrávkách objevují. Zdaleka nejnáročnějším typem promluvy pro stanovení počtu slabik byla volná promluva, kde se sice opakovala určitá témata, ale každá nahrávka byla unikátní. Naopak u říkanky byla většina promluv obsahově velmi podobná a lišily se pouze ty, kde dítě zaměnilo některé slovo nebo se přeřeklo. Podobně tomu bylo u DDK promluv, které se lišily zejména počtem opakování trojice slabik /pa/-/ta/-/ka/. To bylo nakonec vyřešeno zkrácením nahrávek na tři nebo čtyři opakování zmíněné trojice slabik.

4.2.1 Analýza volné promluvy

Pro vyhodnocení veličin AR a SR je nutná znalost počtu slabik v dané promluvě. Ten byl stanoven na základě fonetického přepisu každé nahrávky, který byl vytvořen na základě opakovaného poslechu jednotlivých nahrávek a jejich částí. Nahrávky byly přepisovány pomocí fonetické abecedy SAMPA. Byla snaha nahrávky přepsat co nejpřesněji, bez ohledu na jazykovou správnost nebo význam promluvy. V těžko srozumitelných pasážích, byla snaha o co nejbližší napodobení nahrávky. V přepisu byly speciálním znakem vyznačeny hezitační zvuky, jako je např. „ehm". Tyto zvuky nebyly počítány jako slabiky. Pokud se v promluvě vyskytlo zakoktání, tedy několik fonémů bylo zopakováno, bylo to v přepisu naznačeno pomlčkou, např. „pak si o-obo_uval bo-boty".

Z přepisu byl počet slabik stanoven automaticky pomocí skriptu v prostředí MATLAB, kdy je výsledný počet slabik v každé promluvě stanoven jako součet počtu všech samohlásek, dvojhlásek a slabikotvorných souhlásek. Takže např. v úryvku „pak si o-obo_uval bo-boty" by bylo spočítáno celkem 9 slabik, jelikož jsou započítány i slabiky zopakované při zakoktání. Pro výpočet parametru AR je dále nutná znalost délky promluvy bez řečových pauz. Za tímto účelem byly v nahrávkách ručně nalezeny pauzy delší než 250 ms a v programu Audacity byly odstraněny. Nahrávky s vystříhanými pauzami pak byly znovu uloženy ve stejném formátu jako nahrávky původní. Díky tomu bylo možné určit délku promluvy včetně pauz i délku promluvy s odstraněnými pauzami.

4.2.2 Analýza říkanky

U říkanky by se parametry SR a AR příliš nelišily, jelikož ve většině případů neobsahuje žádné pauzy delší než 250 ms. Proto byl určen pouze jeden parametr, nazvaný RR (recitation rete), který byl určen stejně jako parametr SR u volné promluvy. Tedy jako podíl celkového počtu slabik a celkové délky promluvy. Počet slabik byl ve většině případů stejný. Pokud nedošlo k zakoktání nebo nebylo zaměněno některé slovo, obsahovala nahrávka 26 slabik. Pro přesné stanovení počtu slabik v každé promluvě tak stačil většinou jeden poslech nahrávky pro úpravu počtu slabik v případě, že došlo k přeřeknutí, některá část byla zopakovaná nebo naopak chyběla.

V práci [3] byla mimo jiné zkoumána závislost tempa řeči na věku dítěte právě na těchto promluvách. Postup určení tempa řeči se lišil v tom, že z promluvy byly odstraněny pauzy delší než 150 ms a počet slabik byl předpokládán vždy stejný, tedy 26 slabik. V uvedené práci vychází silná korelace mezi tempem řeči v říkance a věkem dítěte.

4.2.3 Analýza DDK promluvy

V diadochokinetické promluvě byla určena veličina, která se v anglické literatuře někdy nazývá „DDK rate", zkráceně tedy DDKR. Jedná se o stanovení tempa řeči v části DDK promluvy, kde nedochází k narušení rytmu, kterým subjekt slabiky opakuje. Původním záměrem bylo z každé nahrávky vybrat 4 opakování trojice slabik /pa/-/ta/-/ka/, vyslovených ve správném pořadí. Byla snaha vybrat úseky, kde je řečové tempo přibližně konstantní. Z toho důvodu často nebyla použita první a poslední trojice slabik v nahrávce.

První trojice byla často vyslovena pomaleji, jakožto zkušební. Poslední trojice často obsahovala změnu v intonaci, jako na konci věty. V některých nahrávkách se nepodařilo získat 4 souvislá opakování trojice slabik /pa/-/ta/-/ka/, a proto byly použity pouze 3 opakování. V nahrávkách, kde se neobjevila ani 3 souvislá opakování této trojice slabik, bylo nutné spojit 2 úseky nahrávky tak, aby tempo nahrávky nebylo narušeno. Úpravy nahrávek byly prováděny v programu Audacity.

4.3 Metody pro automatický odhad počtu slabik v promluvě

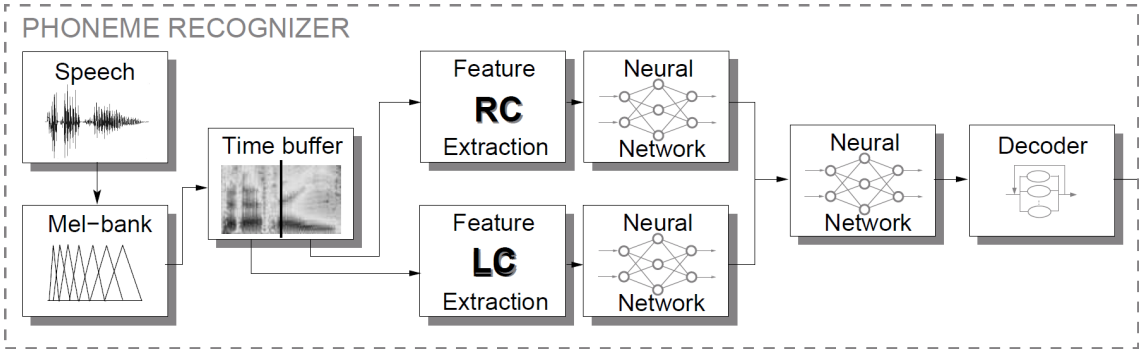
V následující sekci jsou popsány volně dostupné metody pro odhad tempa řeči v promluvě. Metody zahrnují Praat skript [16], který detekuje slabičná jádra v promluvě na základě intenzity a znělosti a Theta Oscilátor [17] který využívá průběh sonority v promluvě. Dále metodu využívající automatický přepis promluvy pomocí fonémového rozpoznávače navrženého na VUT v Brně [25] a vlastní algoritmus využívající krátkodobou energii signálu, přítomnost základní hlasivkové frekvence a počet průchodů nulou.

4.3.1 Rozpoznávač VUT

Tato metoda pro odhad tempa řeči v promluvě využívá automatický přepis promluvy pomocí fonémového rozpoznávače [25], který byl vyvinuto na VUT v Brně. Výsledkem rozpoznávače je seznam rozpoznaných fonémů v promluvě a označení jejich hranic na časové ose. Pro odhad tempa řeči v promluvě pomocí rozpoznávače je nejdůležitější přesné rozlišení samohlásek od souhlásek a dalších zvuků, ale už není tolik podstatné rozpoznání konkrétní samohlásky. Pro zlepšení výsledků rozpoznávače bylo zavedeno několik lingvistických pravidel, které upravují výstup rozpoznávače před analýzou tempa řeči.

Princip fonémového rozpoznávače je naznačen na obrázku 4.2. Signál je filtrován bankou filtrů v melovské frekvenční škále a jsou vybrány úseky o délce 310 ms, kde je spočten vektor 31 hodnot krátkodobé energie signálu pro každé frekvenční pásmo. Vektory jsou v časové oblasti rozděleny na dvě části (levou a pravou) a každá část je váhována příslušnou polovinou Hemmingova okna. Na každou část je následně aplikována diskrétní kosinová transformace (DCT) a z každé části je ponecháno 11 DCT koeficientů. Tyto koeficienty v každé části slouží jako vstup do umělé neuronové sítě (ANN), jejímž výstupem jsou posteriorní pravděpodobnosti fonémů. Tyto výstupy z obou sítí jsou spojeny do jednoho vektoru, a je na ně aplikován logaritmus. Vektor je vstupem další umělé neuronové sítě, jejímž výstupem jsou výsledné pravděpodobnosti fonémů, které jsou dekódovány pomocí Viterbiho algoritmu.

Rozpoznávač byl natrénován pro 4 různé jazyky. Pro češtinu, maďarštinu a ruštinu na databázi SpeechDat-E a pro angličtinu na databázi TIMIT. V rámci této práce byla využita česká varianta rozpoznávače. Vzhledem k tomu, že v databázi SpeechDat-E pro češtinu je pouze velmi nízké zastoupení nahrávek od skupiny 16 let a mladších [26], není očekáváno, že bude výsledek rozpoznávače bezchybný, jelikož dětské promluvy se mohou značně lišit od promluv od dospělých lidí. Aplikace, na kterou je v této práci



Obrázek 4.2: Schéma fonémového rozpoznávače. Upraveno z [25].

rozpoznávač využit, nevyžaduje přesné rozlišení mezi jednotlivými fonémy, ale důležité je zejména rozlišení mezi samohláskami, souhláskami a dalšími zvuky (pauza, hezitace, cizí zvuky).

Na obrázku 4.3 je na krátkém úseku promluvy porovnán výstup rozpoznávače s manuálně určenými hranicemi fonémů. Značka „?" zde značí zvuk cizího původu a značka „-" označuje tichou pauzu. Z obrázku je patrné, že ve většině případů jsou hranice jednotlivých fonémů určeny správně, pouze s drobnými odchylkami od manuálně určených hranic. V několika případech je chybně určena samohláska, např. místo /o/ ve slově „boty" je rozpoznáno /a/, apod. Nicméně na určování tempa řeči toto nemá vliv. Jediná samohláska v této ukázce, kterou se nepodařilo rozpoznat, je /e/ ve slově „se". Aby se eliminovaly tyto drobné nepřesnosti, bylo zavedeno několik lingvistických pravidel, která jsou aplikována na výstup rozpoznávače předtím, než je analyzováno tempo řeči. Tato pravidla by nutně nemusela být vhodná při použití rozpoznávače na promluvy od dospělých lidí, ale u dětských promluv, které byly použity v této práci, vedly ke zpřesnění výsledků. Zavedena byla následující pravidla v uvedeném pořadí.

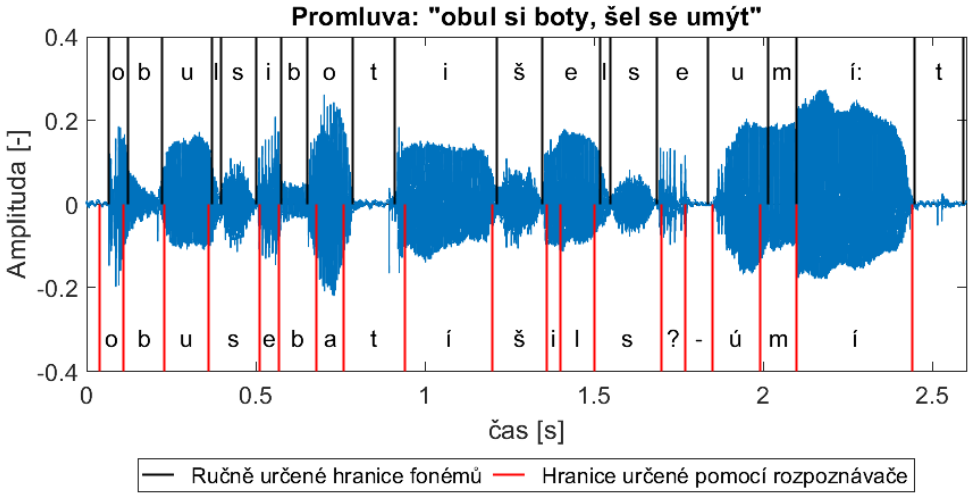
1. Pokud jsou vedle sebe nalezeny dvě stejné samohlásky, jsou spojeny do jedné. A to bez ohledu na to zda se jedná o dlouhou či krátkou variantu samohlásek.

2. Pokud je nalezena dvojice samohlásek /au/ nebo /ou/, jsou spojeny do jedné dvojhlásky /a_u/, resp. /o_u/.

3. Pokud se souhláska /l/ nebo /r/ nachází mezi dvěma souhláskami, nebo jí předchází souhláska a následuje řečová pauza, je nahrazena slabikotvornou variantou /l=/, resp. /r=/.

4. Pokud jsou nalezeny 3 souhlásky v řadě (kromě souhlásek /l/ a /r/) je prostřední z nich nahrazena samohláskou.

První dvě pravidla mají za úkol redukovat chybné rozdělení některých fonémů na dvě části a tudíž zkreslení výsledného počtu samohlásek, resp. dvojhlásek v promluvě. Pravidlo č. 3 je zavedeno z toho důvodu, že rozpoznávač nedokáže odlišit slabikotvorné varianty českých souhlásek /l/ a /r/. Poslední pravidlo bylo zavedeno ve snaze nalézt samohlásky i v úsecích, kde nebyly zaznamenány rozpoznávačem. Toto pravidlo by mohlo



Obrázek 4.3: Ukázka výstupu fonémového rozpoznávače

způsobit problém u slov, která skutečně obsahují tři souhlásky v řadě, které ale netvoří slabiku. Takových slov však v použité databázi nebylo mnoho, a když už takové slovo v nahrávce bylo, např. slovo „vstávat", bylo často vysloveno zjednodušeně, spíše jako „stávat".

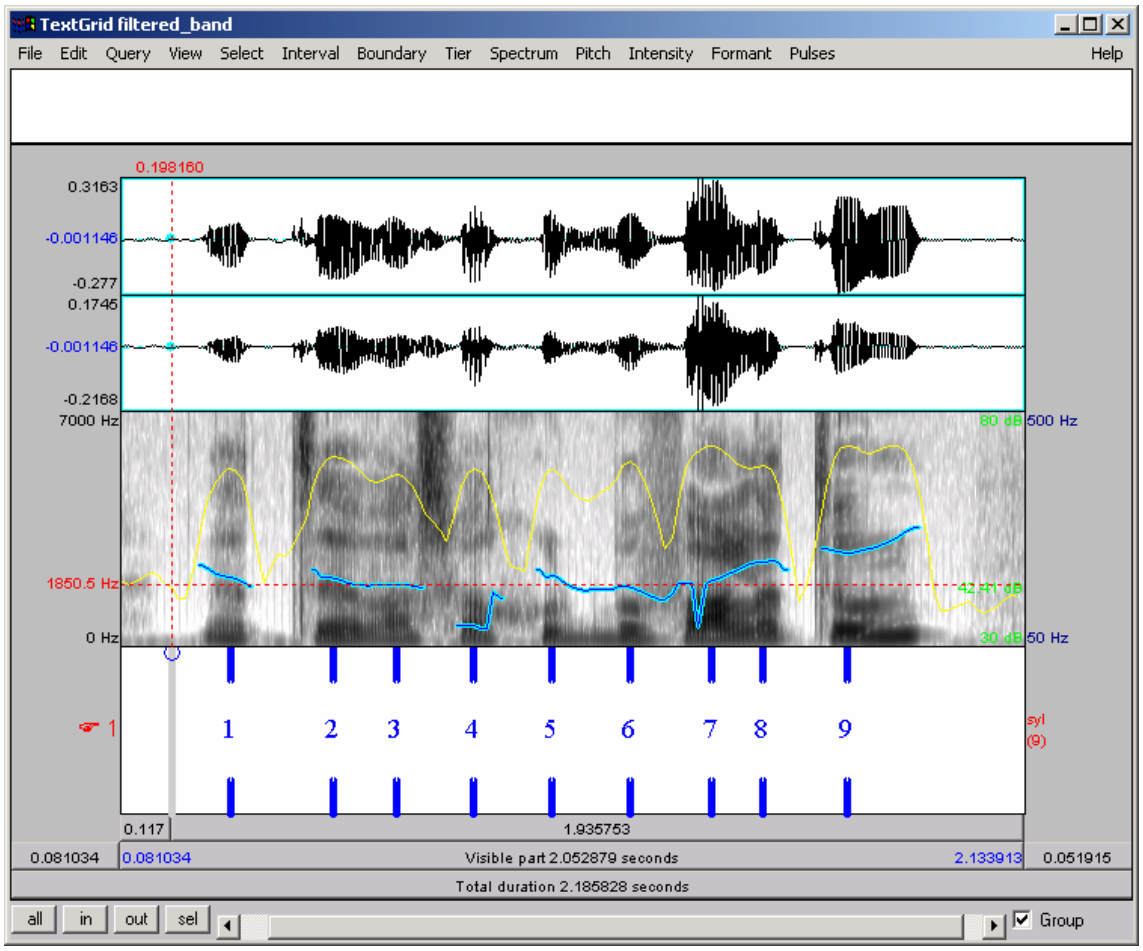
Výsledný počet slabik v promluvě je určován z výstupu rozpoznávače, upraveného zavedenými lingvistickými pravidly. Na takto přepsané promluvy je použit skript napsaný v prostředí MATLAB. Stejně jako u manuálně přepsaných promluv je výsledný počet slabik stanoven jako součet všech samohlásek, dvojhlásek a slabikotvorných souhlásek v promluvě.

4.3.2 Praat skript

Tato metoda je volně dostupná ve formě skriptu [16] pro detekci slabičných jader v promluvě a byla vyvinuta v programu Praat [27]. Motivací autorů byla automatizace úlohy výpočtu tempa řeči, která bývá ve výzkumech, zabývajících se poruchami řeči, apod., často vynechávána z důvodu časové náročnosti měření, pokud je prováděno ručně. To bývá problém zejména u velkých databází s několika hodinami záznamů promluv. Cílem této metody je nalezení jednotlivých slabičných jader a následné určení jejich celkového počtu v promluvě. Z toho lze následně vypočítat tempo řeči jako podíl počtu slabik a délky promluvy.

Slabičná jádra jsou nejčastěji tvořena samohláskami, pro které je typická přítomnost základní hlasivkové frekvence a vyšší energie. Proto je v rámci této metody pro detekci slabičných jader využito parametrů intenzita signálu a znělost signálu. Ukázka funkce skriptu je naznačena na obrázku 4.4, kde je v horní části zobrazen průběh signálu, v prostřední části je spektrogram, modře vyznačený průběh f_0 a žlutě vyznačená intenzita signálu a v dolní části jsou označena detekovaná slabičná jádra. Autoři doporučují pro lepší

funkčnost skriptu u zašuměných signálů aplikovat filtraci omezující spektrum signálu na řečové pásmo.



Obrázek 4.4: Ukázka funkce Praat skriptu. Převzato z [28].

Postup samotného algoritmu je následující:

- Výpočet průběhu intenzity v segmentech o délce 64 ms, s krokem 16 ms
- Nalezení lokálních maxim vyšších než medián intenzity v celém signálu. Pokud byl signál filtrován na řečové pásmo, je možné hranici zvýšit o 2 dB. Nalezené vrcholy vyšší než stanovený práh tvoří potenciální kandidáty na přítomnost slabičného jádra.
- Určení zda v úseku mezi dvěma maximy došlo k poklesu intenzity alespoň 2 dB (ve filtrovaném signálu 4 dB). Pokud k poklesu o zvolenou hodnotu nedojde, je vrchol jako kandidát vyloučen.
- Výpočet průběhu základní hlasivkové frekvence v segmentech o délce 100 ms, s krokem 20 ms. K tomu je použit autokorelační estimátor, který je součástí programu

Praat [27]. Na základě toho jsou vyloučeny vrcholy, které se nacházejí v neznělých úsecích, kde nebyla zjištěna přítomnost základní hlasivkové frekvence.

- Zbylé vrcholy jsou považovány za slabičná jádra.

Výstup skriptu je uložen do souboru ve formátu TextGrid. Tam jsou zaznamenány časové značky nalezených slabičných jader. Pro určení tempa řeči v promluvě stačí spočítat celkový počet nalezených jader a podělit ho délkou promluvy v sekundách.

4.3.3 Theta Oscilátor

Algoritmus [17] je založený na principu oscilátorů a rytmické segmentaci. Z nahrávek řeči je vypočítán časový průběh veličiny zvané sonorita neboli zvučnost. Její hodnota by měla být vyšší v částech promluvy, kde převládá tónová složka nad šumovou. Takže největších hodnot by měla dosahovat u samohlásek a naopak nejnižších u sykavek. Jedná se tedy o vhodný příznak pro detekci slabik, které jsou nejčastěji tvořené právě samohláskami.

Prvním krokem algoritmu je převzorkování signálu na vzorkovací kmitočet 16 kHz. Převzorkovaný signál poté prochází bankou 20 filtrů logaritmicky rozložených v pásmu od 50 do 7500 Hz. V každém pásmu je pak signál podvzorkován na vzorkovací frekvenci 1 kHz. Potom jsou jednotlivé signály přivedeny na vstupy oscilátorů, které jsou popsány následujícími rovnicemi:

$$F_c(t) = e_c(t) - ky_c(t-1) - dv_c(t-1) \tag{4.1}$$

$$v_c(t) = v_c(t-1) + F_c(t)/(f_s m) \tag{4.2}$$

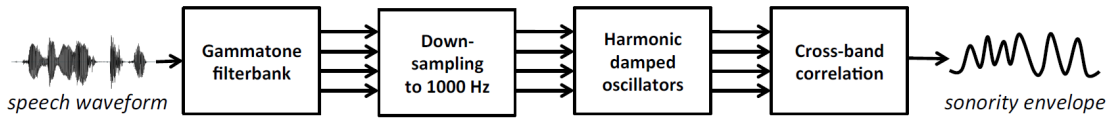
$$y_c(t) = y_c(t-1) + v_c(t)/f_s \tag{4.3}$$

kde c je index jednotlivých pásem, který nabývá hodnot c = 1, 2, ... 20, e_c jsou jednotlivé obálky, F, v a y jsou síla, rychlost a amplituda oscilátoru v čase t, m reprezentuje hmotnost, k tuhost a d koeficient tlumení. Zafixováním hodnoty $k = 1$ je umožněno nastavením m a d stanovit centrální kmitočet f_0 a šířku pásma Δf. Standardní hodnoty těchto parametrů jsou $f_0 = 5$ Hz a Δf = 6 Hz. Tyto harmonicky tlumené oscilátory odpovídají elektrickému rezonančnímu obvodu druhého řádu. Všechny oscilátory jsou naladěny stejně. V každém segmentu je následně proveden logaritmický součet amplitud v N pásmech s nejvyšší energií podle vzorce:

$$S(t) = \sum_{i=1}^{N} \log_{10} \vec{y}_i(t) \tag{4.4}$$

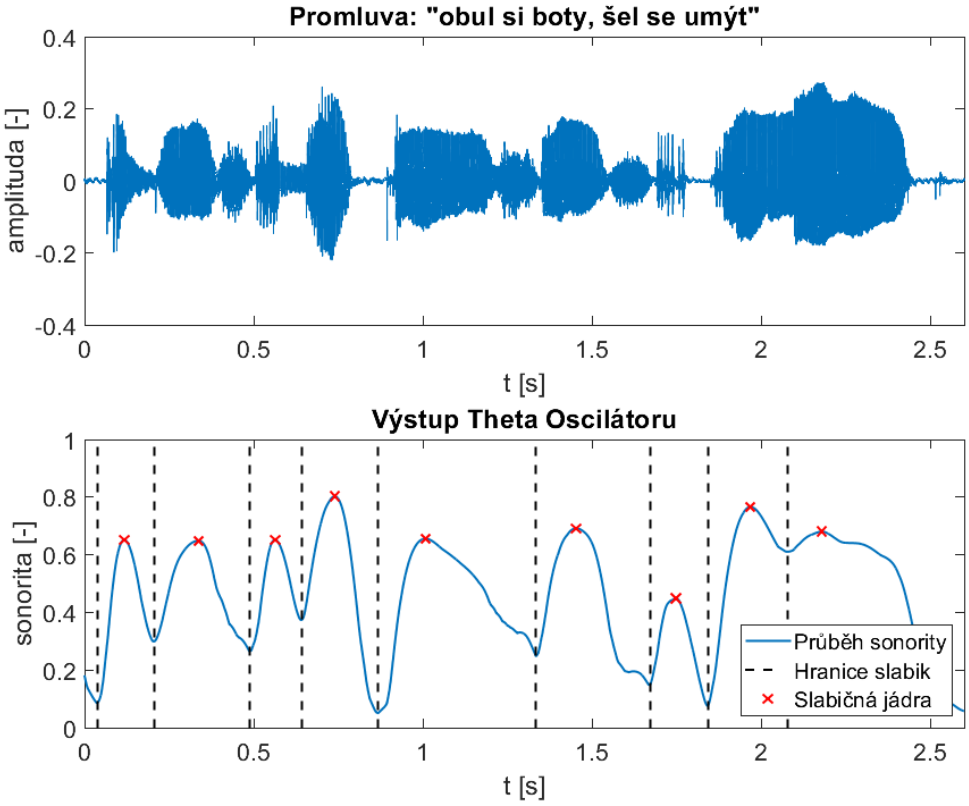
kde $\vec{y}_i$ je vektor amplitud oscilátorů v daném čase t, seřazených od největší po nejmenší. Tím vznikne časový průběh sonority. Před výpočtem logaritmu je k hodnotám přičtena konstanta, aby amplitudy nenabývaly záporných hodnot. Hodnota N byla autory experimentálně stanovena na $N = 8$. Následujícím krokem algoritmu je normování průběhu sonority tak, aby nabývala hodnot od 0 do 1. V normovaném průběhu jsou hledána lokální

minima, kterým předchází lokální maximum větší o zvolenou konstantu δ než dané minimum. Takto nalezené body jsou pak považovány za hranice slabik. Pozice každé slabiky je pak určena jako maximum mezi jejími hranicemi. Blokové schéma celého algoritmu je zobrazeno na obrázku 4.5.



Obrázek 4.5: Blokové schéma Theta Oscilátoru. Převzato z [17].

Na základě experimentů byla v této práci nastavena hodnota δ = 0,05, u které vycházely nejpřesnější výsledky. Testovány byly hodnoty parametru $delta$ = 0,01, 0,02, 0,05, 0,1, 0,2 a 0,5. Hodnota parametru f_0 byla ponechána na autory doporučené hodnotě f_0 = 5 Hz a parametr Q, který odpovídá podílu f_0 a Δf, byl nastaven tak, aby Δf = 6, což je rovněž autory doporučená hodnota. Ukázka funkce algoritmu na části nahrávky je zobrazena na obrázku 4.6, kde jsou naznačeny nalezené hranice slabik i pozice slabičných jader v průběhu sonority.



Obrázek 4.6: Ukázka výstupu Theta Oscilátoru

4.3.4 Vlastní algoritmus

Tento algoritmus byl navržen v rámci diplomové práce jako alternativa k ostatním volně dostupným algoritmům. Jeho úkolem je detekovat slabičná jádra v promluvě. Byl realizován v prostředí MATLAB a vychází z některých principů často uváděných v literatuře. Je využita krátkodobá energie pásmově filtrovaného signálu, podobně jako v [18], [16], dále detekce základní hlasivkové frekvence f_0, která byla využita např. v [21], [16] a dále počet průchodů nulou ZCR, jehož je použito mimo jiné v [19]. V následujících odstavcích je algoritmus podrobně popsán.

Signál je filtrován pásmovou propustí s mezními kmitočty f_1 = 300 Hz a f_2 = 1,1 kHz, aby bylo zvýrazněno pásmo, ve kterém se typicky nachází první formanty českých samohlásek [6]. Je použit filtr typu Butterworth 5. řádu. Dále se pracuje jak s filtrovaným tak s nefiltrovaným signálem. Oba signály jsou segmentovány na úseky o délce 20 ms s krokem 10 ms. V každém segmentu filtrovaného signálu je vypočítán výkon P_{dB} podle vzorce 4.5 a v každém segmentu nefiltrovaného signálu je vypočten vychýlený odhad autokorelační funkce $R_s[k]$ podle vzorce 4.6 a počet průchodů nulou Z podle vzorce 4.7.

$$P_{dB} = 10 \log \left(\frac{1}{L} \sum_{n=1}^{L} s^2[n] \right) \tag{4.5}$$

$$R_s[k] = \frac{1}{L} \sum_{k=0}^{L-|k|-1} s[n]s[n+k] \tag{4.6}$$

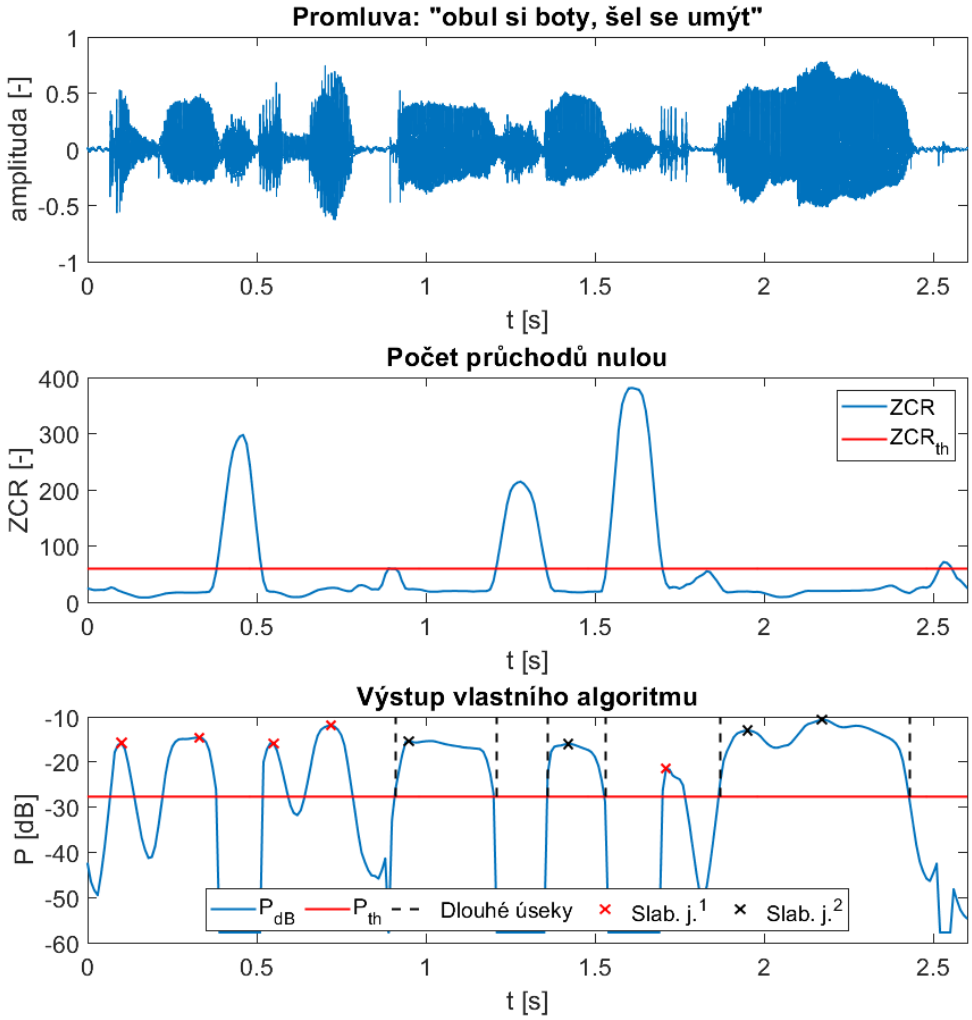
$$Z = \frac{1}{L} \sum_{n=1}^{L-1} |\operatorname{sgn}(s[n+1]) - \operatorname{sgn}(s[n])| \tag{4.7}$$

kde s je segment o délce L.

Z odhadu autokorelační funkce je vyhodnocena přítomnost základní hlasivkové frekvence podle následujícího postupu. V průběhu autokorelační funkce je vyhledáno maximum v intervalu odpovídajícímu frekvenčnímu pásmu 60 až 400 Hz, ve kterém se základní hlasivkový tón typicky nachází [6]. V každém segmentu je uložena jak výška maxima M, tak pozice maxima, přepočtená na frekvenci f_0. Dále je v celém signálu vypočtena prahová hodnota M_{th}, a to jako useknutá střední hodnota (trimmed mean) s vyloučením 10 % krajních hodnot ze všech hodnot M. Segmenty, ve kterých je maximum vyšší než prahová hodnota, jsou označeny jako ty, které obsahují základní hlasivkovou frekvenci. Průběh M je následně vyhlazen klouzavým mediánem s délkou okna 5 prvků.

Průběhy krátkodobého výkonu P_{dB} a počtu průchodů nulou Z jsou vyhlazeny klouzavým průměrem, s délkou okna 5 prvků. Potom jsou vypočteny jejich prahové hodnoty. P_{th} jako useknutá střední hodnota s vyloučením krajních 10 % krajních hodnot průběhu P_{dB} a hodnota Z_{th} je určena jako střední hodnota průběhu Z. Segmenty kde je $Z > Z_{th}$ nejsou dále uvažovány, proto jsou v těchto úsecích hodnoty M nahrazeny nulou a hodnoty P_{dB} nahrazeny minimem P_{dB}, jelikož P_{dB} je udáváno v decibelech a většinou nabývá záporných hodnot. To je mimo jiné naznačeno na obrázku 4.7.

V průběhu P_{dB} jsou následně vyhledány úseky, kde je $P_{dB} > P_{th}$ a je vypočtena délka těchto úseků. Vyšší výkon signálu je typicky ukazatelem přítomnosti samohlásek, což je



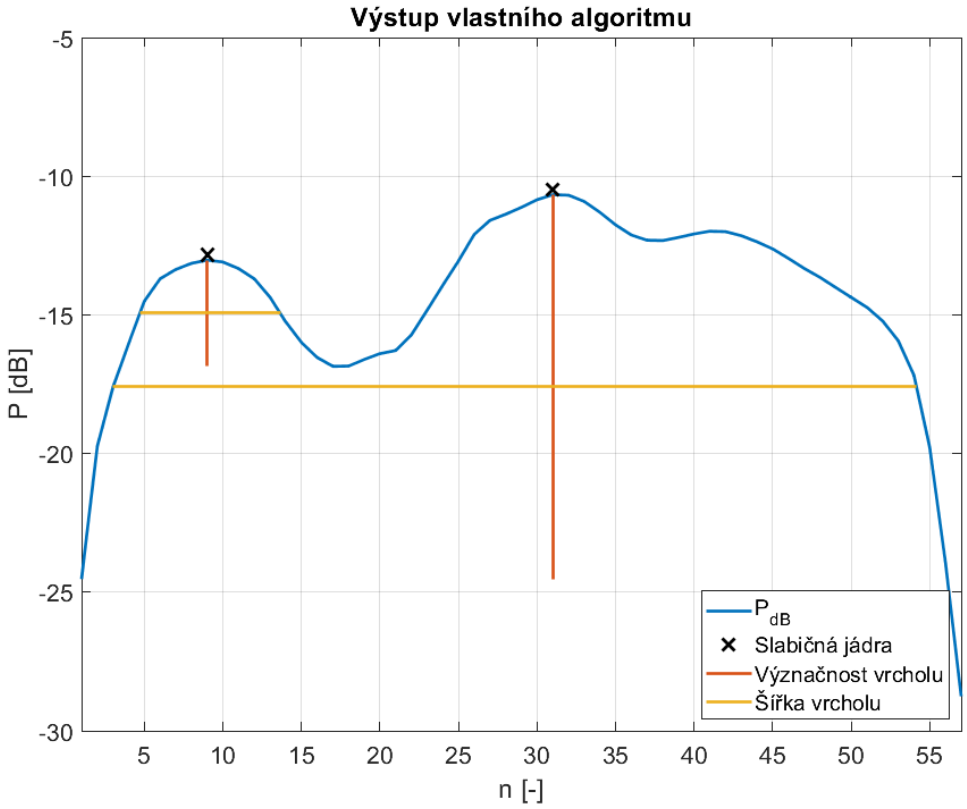
Obrázek 4.7: Ukázka funkce vlastního algoritmu

ještě zvýrazněno tím, že v tomto případě je výkon počítán z pásmově filtrovaného signálu. Proto jsou tyto úseky označené jako obsahující samohlásky. Jsou rozděleny na delší a kratší úseky podle hraniční hodnoty, stanovené jako medián délek všech úseků. Pokud je medián menší než 60 ms, potom je hraniční hodnota zvýšena na tuto hodnotu. Kratší úseky jsou pak brány jako právě jedna samohláska a delší úseky jsou dále zkoumány, aby byly ošetřeny případy, kdy se v úseku nachází více samohlásek.

V úsecích delších než hraniční hodnoty jsou vyhledány vrcholy v průběhu krátkodobého výkonu, které splňují následující parametry. Jejich minimální význačnost (prominence) musí být minimálně 0,5 dB, jejich minimální vzdálenost musí být 50 ms, jejich

[1] Slabičná jádra detekovaná v krátkých úsecích
[2] Slabičná jádra detekovaná v dlouhých úsecích

šířka musí být minimálně 25 ms a musí se nacházet v úseku, kde byla detekována základní hlasivková frekvence. Význačnost vrcholu je definována jako minimální výška, o kterou musí hodnota signálu klesnout na obou stranách vrcholu než se v signálu objeví hodnota vyšší než daný vrchol. Šířka vrcholu je počítána v polovině jeho význačnosti. Určování význačnosti a šířky vrcholů je naznačeno na obrázku 4.8, kde je zobrazen v detailu úsek promluvy, kde se nachází dvě slabičná jádra.



Obrázek 4.8: Určování význačnosti a šířky vrcholů v průběhu P_{dB}

Celkový počet slabičných jader v promluvě je ve výsledku stanoven jako počet úseků s krátkodobým výkonem vyšším než prahová hodnota P_{th}, kde kratší úseky než hraniční hodnota jsou počítány jako jedna samohláska a delší úseky jsou počítány jako tolik samohlásek, kolik v nich bylo nalezeno vrcholů.

Na obrázku 4.7 je mimo jiné naznačeno, že úseky, kde je $Z > Z_{th}$ jsou v průběhu P_{dB} nahrazeny minimální hodnotou. Dále je zde naznačeno hledání potenciálních slabičných jader v krátkých úsecích signálu, kde je $P > P_{th}$, kde je pouze vyhledáno maximum v daném úseku. V dlouhých úsecích jsou potenciálními kandidáty všechny vrcholy, které splňují výše zmíněné podmínky.

Kapitola 5

Výsledky

V následující kapitole jsou popsány jednak výsledky manuálního vyhodnocení tempa řeči a jednak porovnání automatických metod na odhad řeči v promluvě. Z manuálně vyhodnocených nahrávek jsou stanoveny typické hodnoty tempa řeči v jednotlivých věkových kategoriích a je zkoumána závislost tohoto parametru na věku dítěte. Metody pro automatický odhad tempa řeči jsou porovnány na základě přesnosti určení počtu slabik v různých typech promluv.

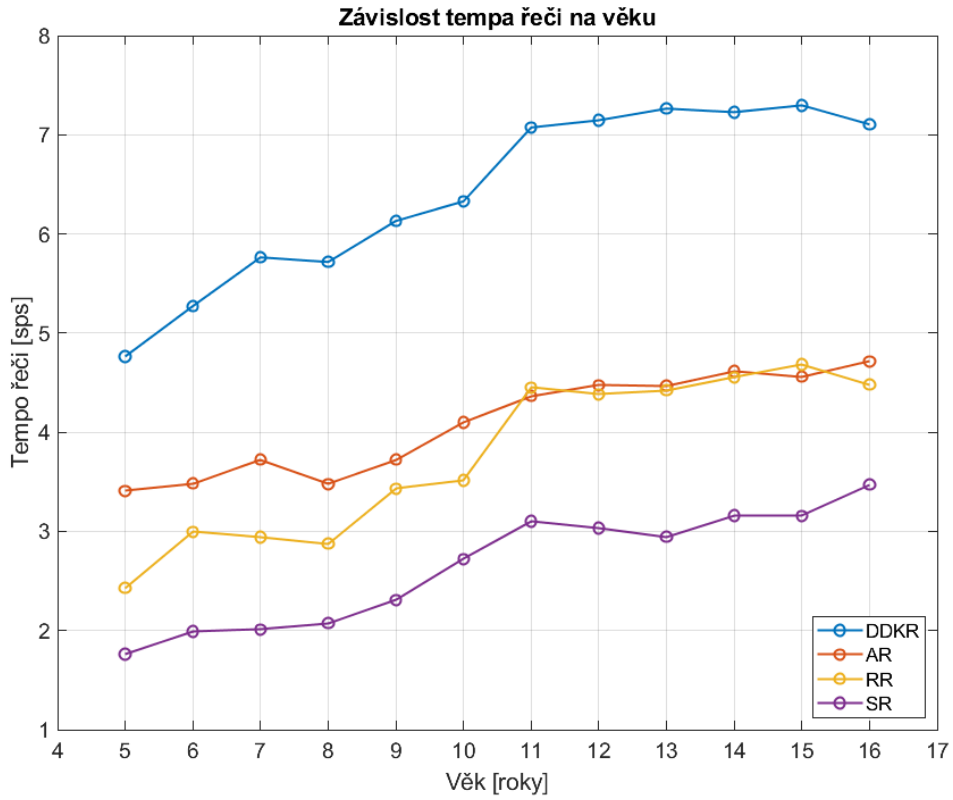
5.1 Manuální vyhodnocení tempa řeči

V této práci byly z dostupných promluv vyhodnoceny čtyři veličiny spojené s tempem řeči. Jednak bylo vyhodnoceno celkové tempo řeči ve volné promluvě – SR, dále artikulační tempo řeči ve volné promluvě – AR, tempo řeči v DDK promluvě – DDKR, a tempo řeči v říkance - RR. Všechny veličiny jsou udávány v jednotce počet slabik za sekundu – SPS (syllable per second) a jejich střední hodnoty v jednotlivých věkových kategoriích jsou porovnány na obrázku 5.1.

Za účelem výpočtu středních hodnot všech zkoumaných veličin v jednotlivých věkových kategoriích byl věk dětí zaokrouhlen na celé roky. Vzhledem k tomu, že se v kategorii čtyřletých a sedmnáctiletých účastníků studie nachází jen po jedné nahrávce, byly tyto nahrávky pro výpočet středních hodnot přidány k nejbližší věkové kategorii, která obsahuje více vzorků, tedy k pětiletým, resp. šestnáctiletým. Při výpočtu korelací tempa řeči s věkem už bylo počítáno se skutečným věkem, známým s přesností na měsíce.

5.1.1 Porovnání s výsledky vybraných studií

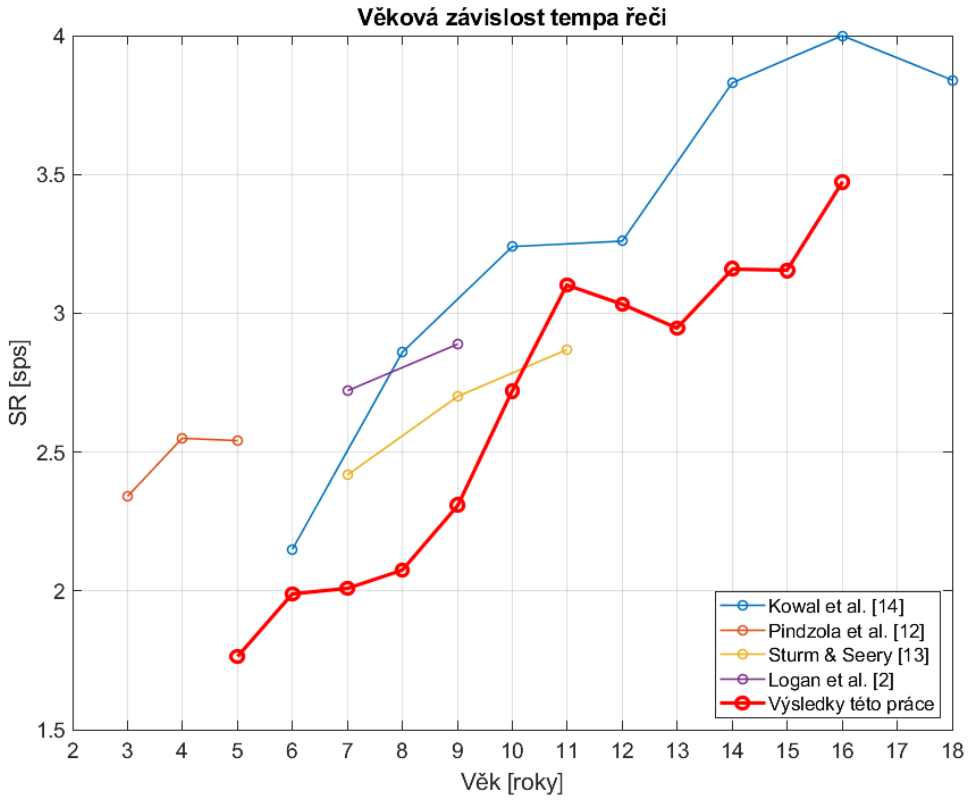
Hodnoty parametru SR ve volné promluvě získané v této práci jsou na obrázku 5.2 porovnány s výsledky vybraných studií, popsaných v podkapitole 3.1. Z obrázku 5.2 je patrné, že i v této práci je výrazná stoupající tendence celkového tempa řeči s věkem dítěte. Hodnoty v této práci vycházejí většinou nižší než ve vybraných studiích, což může být zapříčiněno mnoha faktory. Předně se jedná o studie, prováděné na promluvách v jiných jazycích než je čeština. V některých studiích nebyla analyzována úloha vypravování podle obrázku a proto byla pro porovnání vybrána jiná volná promluva. Dále může mít vliv i



Obrázek 5.1: Porovnání zkoumaných veličin

konkrétní zadání řečové úlohy nebo prostředí, ve kterém se dítě při nahrávání nacházelo. V neposlední řadě byla většina studií včetně této práce prováděna na poměrně malých počtech účastníků, kdy se většinou počet subjektů v každé věkové kategorii pohyboval od 10 do 30, viz tabulka 3.2. Některé studie pak pracovaly s nahrávkami, které byly pořízeny v jedné místní škole, nebo například v několika školách v jedné oblasti, což by mohlo zvýšit vliv místních dialektů apod.

Podobné srovnání bylo provedeno i pro studie vyhodnocující parametr AR, tedy artikulační rychlost. Hodnoty AR získané v rámci této práce opět vychází spíše nižší než ve vybraných studiích a to zejména pro starší děti a mladistvé, viz obrázek 5.3. Příčiny tohoto výsledku mohou být již zmíněné rozdíly v jazyce, nízký počet účastníků studií, konkrétní zadání řečových úloh, apod. Zde se navíc může daleko výrazněji projevit i definice artikulačního tempa. Například zda bylo artikulační tempo vyhodnocováno pouze ve vybraných plynulých úsecích nahrávky nebo zda bylo vyhodnoceno v celé promluvě s odstraněnými pauzami, jako tomu bylo v této práci. Dále se mohou lišit definice plynulých úseků, například zda jsou odstraněny všechny nespojitosti, jako např. přeřeknutí, nebo zda jsou ponechány. A konečně velký vliv na výsledek mohla mít také definovaná maximální délka pauzy, která může být v promluvě ponechána.

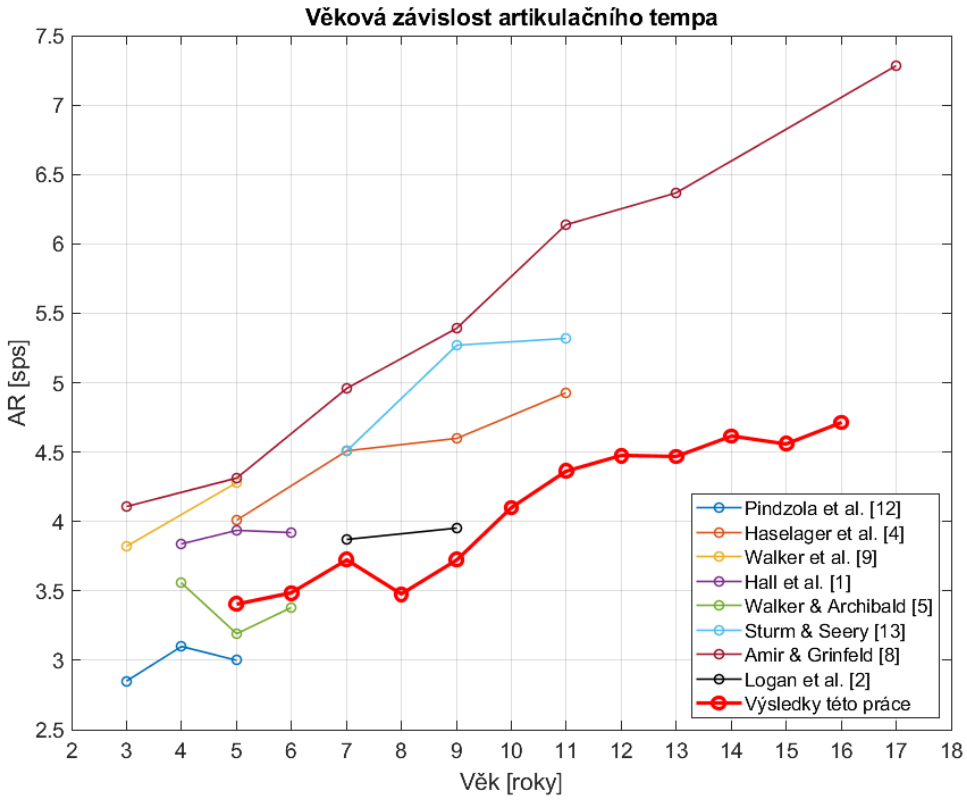


Obrázek 5.2: Porovnání s výsledky vybraných studií - SR

5.1.2 Věková závislost tempa řeči

V databázi 245 dětských promluv bylo manuálně vyhodnoceno tempo řeči ve třech typech promluv. Ve volné promluvě byly hodnoceny veličiny SR a AR, tedy průměrné tempo řeči v celé promluvě a artikulační tempo v úsecích řečové aktivity. Dále bylo vyhodnoceno tempo řeči v minimálně třech plynulých opakováních trojice slabik /pa/-/ta/-/ka/ v DDK promluvě a průměrné tempo řeči v říkance. Jelikož ne všechny zkoumané veličiny vyhovují Chí-kvadrát testu dobré shody na normalitu rozdělení dat, byly věkové závislosti vyhodnoceny pomocí výpočtu Spearmanova koeficientu pořadové korelace. Tabulka 5.1 zachycuje vypočtené korelační koeficienty dané veličiny s věkem kontrolované na pohlaví dítěte a v dalším sloupečku korelace s parametrem pohlaví dítěte, kontrolované na věk.

Všechny čtyři hodnocené veličiny vykazují silnou korelaci s věkem, jelikož Spearmanův korelační koeficient vychází ve všech případech v rozmezí 0,6 až 0,7. Nejsilnější korelace s věkem vychází u říkanky a naopak nejslabší oproti ostatním u DDK promluvy. Dále byly analyzovány vzájemné korelace jednotlivých veličin, opět kontrolované na faktor pohlaví dítěte, viz tabulka 5.2. Výsledky vzájemných korelací naznačují, že jednotlivé veličiny jsou navzájem svázány. Nejvyšší korelace vychází mezi AR a SR, což je s největší pravděpodobností způsobeno zejména tím, že byly vyhodnocovány ve stejné promluvě. Silná korelace vychází i mezi těmito veličinami a tempem řeči v říkance. Mírná

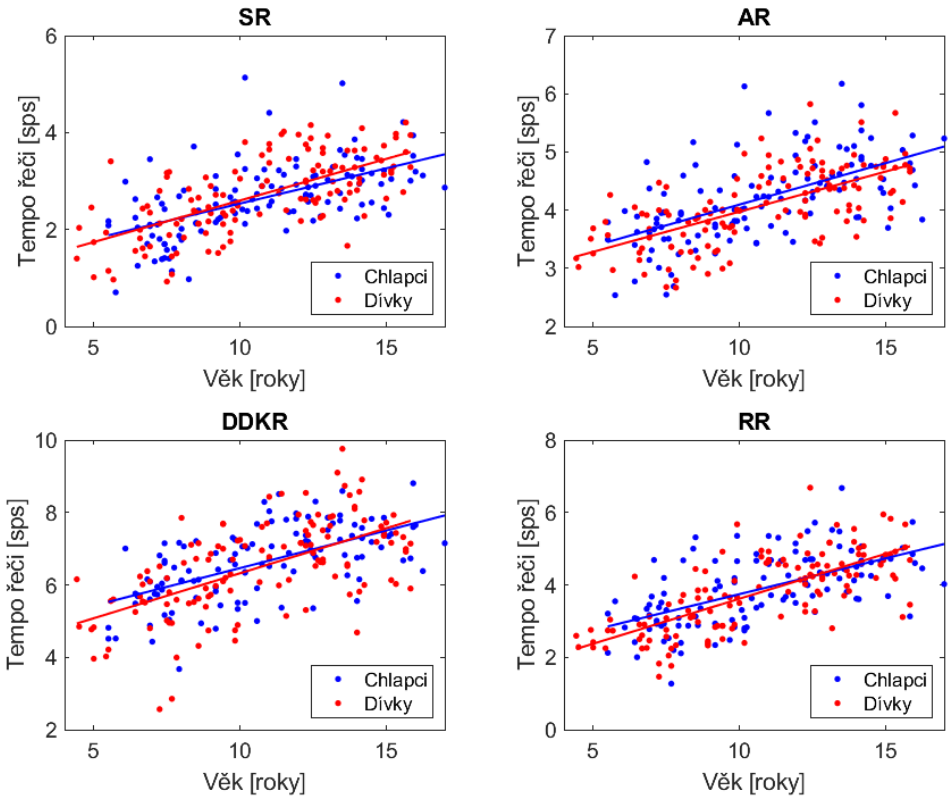


Obrázek 5.3: Porovnání s výsledky vybraných studií - AR

Tabulka 5.1: Parciální korelace zkoumaných veličin s věkem a pohlavím dítěte

	Věk	Pohlaví
AR	0,645 ($p > 0,001$)	-0,084 ($p = 0,19$)
SR	0,636 ($p > 0,001$)	0,098 ($p = 0,13$)
DDKR	0,609 ($p > 0,001$)	0,061 ($p = 0,34$)
RR	0,675 ($p > 0,001$)	0,073 ($p = 0,26$)

korelace vychází mezi tempem řeči v DDK promluvě a ostatními veličinami. Nižší korelace může být způsobena jednak rozdílným způsobem vyhodnocení veličiny DDKR, kde je vyhodnocena pouze část promluvy a nikoliv celá, jako je to u ostatních veličin, a jednak rozdílným zadáním úlohy, kdy v ostatních promluvách není tempo řeči nijak specifikováno, zatímco v DDK promluvě je úkolem opakovat slabiky co nejrychleji.



Obrázek 5.4: Korelace zkoumaných veličin s věkem

Tabulka 5.2: Vzájemné korelace veličin

	RR	DDKR	AR
SR	0,600[1]	0,482[1]	0,768[1]
AR	0,629[1]	0,536[1]	
DDKR	0,518[1]		

[1] $p > 0,001$

5.2 Porovnání metod pro automatický odhad tempa řeči

Jelikož je v této práci tempo řeči počítáno jako počet slabik za sekundu, bylo možné úlohu automatického odhadu tempa řeči zjednodušit na určení počtu slabik v promluvě. Metody, které byly porovnávány, jsou blíže rozepsány v podkapitole 4.3. Jedná se o Rozpoznávač VUT, Praat skript, Theta Oscilátor a vlastní algoritmus. V následující podkapitole jsou vyhodnoceny odchylky počtu nalezených slabik v promluvách pomocí těchto metod od skutečných hodnot určených manuálně. Pro porovnání jsou použita tři kritéria střední chyba ME, druhá odmocnina ze střední kvadratické chyby RMSE a Spearmanův

koeficient pořadové korelace r_s. První dvě kritéria jsou definována jako

$$\text{ME} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{x_i} - x_i) \tag{5.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{x_i} - x_i)^2} \tag{5.2}$$

kde N značí počet promluv, $\widehat{x_i}$ je odhad počtu slabik v dané promluvě a x_i je referenční hodnota počtu slabik v promluvě určená manuálně. Střední chyba ME indikuje, zda má algoritmus tendenci počet slabik spíše nadhodnotit, nebo podhodnotit. Pokud je hodnota kladná, znamená to, že daný algoritmus typicky odhaduje spíše vyšší počet slabik, než se v promluvě skutečně nachází, a pokud je hodnota záporná, pak odhaduje spíše nižší počet. Kritérium RMSE nabývá pouze kladných hodnot, jelikož se jedná o odmocninu z průměru kvadratických chyb. Čím je jeho hodnota menší, tím větší přesnosti dosahuje testovaný algoritmus. Spearmanův korelační koeficient se pohybuje v rozmezí -1 až 1. Čím přesnější je algoritmus, tím je jeho hodnota vyšší.

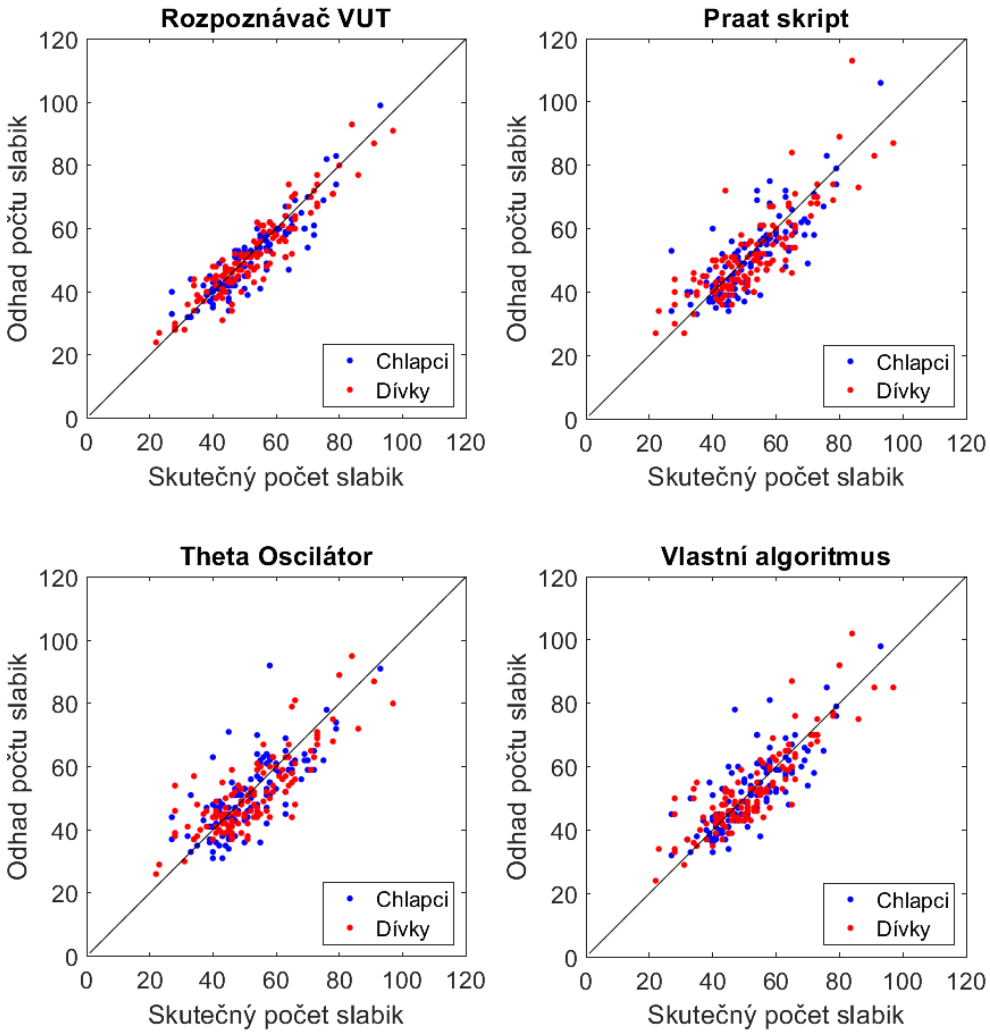
5.2.1 Odhad počtu slabik ve volné promluvě

Odhad počtu slabik ve volné promluvě by měl algoritmům působit největší problémy. Zaprvé jsou promluvy delší než ve zbylých dvou úlohách, viz tabulka 4.1. Dalším faktorem je, že obsahovaly velké množství pauz, hezitací, přeřeknutí a často také těžko srozumitelné úseky, ve kterých bylo těžké rozpoznat slabiky i při manuálním hodnocení. Výsledky porovnávaných metod jsou uvedeny v tabulce 5.3. Porovnání výsledků metod je naznačeno také na obrázku 5.5, kde jsou na vodorovné ose uvedeny manuálně určené počty slabik a na svislé ose počty slabik nalezené jednotlivými algoritmy. Úhlopříčka v grafech naznačuje kde by se měly nacházet hodnoty počtu detekovaných slabik, pokud by algoritmy fungovaly bezchybně. Z důvodu lepší čitelnosti grafů, je na obrázcích uveden pouze výřez grafů, do kterého se nevešly data ze tří promluv, které obsahovaly více než 120 slabik.

Tabulka 5.3: Porovnání přesnosti metod ve volné promluvě

Metoda	ME	RMSE	r_s	p
Rozpoznávač VUT	-1,306	5,348	0,903	$> 0,001$
Praat skript	-0,992	7,999	0,813	$> 0,001$
Theta Oscilátor	-1,163	8,703	0,765	$> 0,001$
Vlastní algoritmus	0,151	7,694	0,822	$> 0,001$

Nejlepších výsledků ve volné promluvě dosahuje jednoznačně metoda využívající fonémový rozpoznávač, u kterého vychází nejnižší hodnota RMSE a naopak nejvyšší korelační koeficient. Druhé nejvyšší skóre bylo dosaženo pomocí vlastního algoritmu, třetí nejlepší je Praat skript a nejnižší skóre má Theta Oscilátor.



Obrázek 5.5: Porovnání přesnosti metod ve volné promluvě

5.2.2 Odhad počtu slabik v říkance

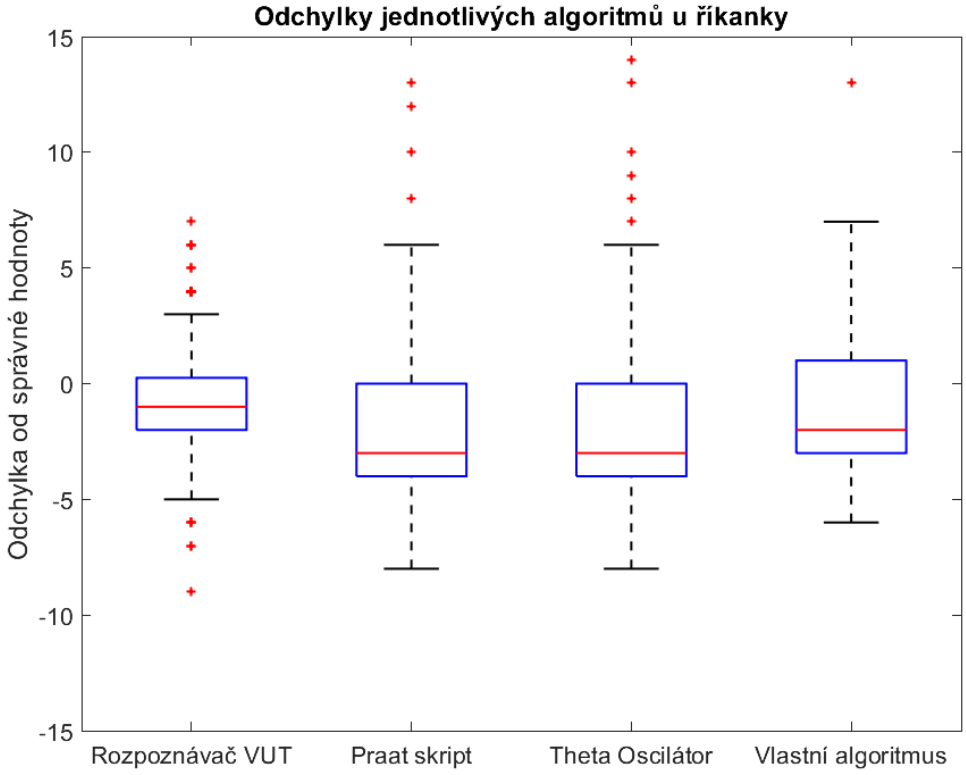
V říkance bylo ve většině případů přesně 26 slabik, proto nemá význam hodnotit algoritmy pomocí korelačního koeficientu. Nejlepším ukazatelem přesnosti je tak RMSE. Porovnání metod je uvedeno v tabulce 5.4. Nejlepších výsledků bylo dosaženo opět pomocí fonémového rozpoznávače. Další pořadí je pak stejné jako v předchozí úloze, tedy druhý nejlepší je vlastní algoritmus, poté Praat skript a nakonec Theta Oscilátor.

Výsledky jsou navíc uvedeny také na obrázku 5.6, kde je pomocí boxplotu naznačeno rozdělení jednotlivých odchylek kolem referenčních hodnot. Vyneseny jsou rozdíly $\widehat{x_i} - x_i$, kde $\widehat{x_i}$ značí odhad počtu slabik pomocí dané metody a x_i je manuálně určená referenční hodnota. Čím je algoritmus přesnější, tím je jeho boxplot průměrnou hodnotou blíže k hodnotě 0 a zároveň je co nejužší. Z obrázku je stejně jako z tabulky 5.4 jasně

patrné, že nejpřesnější metodou je v této úloze Rozpoznávač VUT.

Tabulka 5.4: Porovnání přesnosti metod v říkance

Metoda	ME	RMSE
Rozpoznávač VUT	-0,759	2,575
Praat skript	-1,963	3,815
Theta Oscilátor	-1,922	3,910
Vlastní algoritmus	-1,078	3,063



Obrázek 5.6: Porovnání přesnosti metod v říkance

5.2.3 Odhad počtu slabik v DDK promluvě

V DDK promluvách se nacházelo vždy buď 9, nebo 12 slabik což je nejméně ze všech řečových úloh, které jsou v této práci testovány a v porovnání s ostatními úlohami jsou nahrávky DDK promluv výrazně kratší, viz tabulka 4.1. Srovnání metod pomocí zvolených kritérií je uvedeno v tabulce 5.5. Jako nejpřesnější metoda vychází v této úloze Theta Oscilátor, který ve zbylých úlohách dosahoval nejslabších výsledků. Zde byl téměř bezchybný. Druhý nejlepší podle parametru RMSE je vlastní algoritmus, a na třetím místě je

Praat skript, jehož výsledky mají sice vyšší korelační koeficient než u vlastního algoritmu, ale detekuje průměrně o téměř 3 slabiky méně, než kolik se jich v promluvě nachází.

Tabulka 5.5: Porovnání přesnosti metod v DDK promluvě

Metoda	ME	RMSE	r_s	p
Rozpoznávač VUT	-3,290	4,242	0,396	$> 0,001$
Praat skript	-2,906	2,979	0,865	$> 0,001$
Theta Oscilátor	-0,012	0,143	0,986	$> 0,001$
Vlastní algoritmus	0,176	0,992	0,793	$> 0,001$

Rozpoznávač VUT se pro tuto úlohu neukázal jako vhodný nástroj, jelikož se mu v některých promluvách nepodařilo detekovat žádné slabiky. Zásadní problém byl v tom, že rozpoznávač řadu slabik označil jako vyplněné pauzy, tedy hezitace, jako je např. „ehm". Proto byla otestována i varianta, kdy byly úseky označené rozpoznávačem jako hezitace započítány do celkového počtu slabik. Tím bylo dosaženo zlepšení u DDK promluv, které by podle metriky RMSE znamenalo posun před Praat skript. U ostatních typů promluv se toto pravidlo projevilo zhoršením výsledků, a proto není v práci uvedeno. Výsledky pro Upravený rozpoznávač VUT jsou uvedeny v tabulce 5.6 společně s výsledky neupraveného rozpoznávače.

Tabulka 5.6: Výsledky upraveného rozpoznávače VUT v DDK promluvě

Metoda	ME	RMSE	r_s	p
Rozpoznávač VUT	-3,290	4,242	0,396	$> 0,001$
Upravený rozpoznávač VUT	0,192	2,369	0,521	$> 0,001$

Kapitola 6

Závěr

V rámci této práce bylo manuálně vyhodnoceno tempo řeči v databázi dětských promluv, obsahující 3 různé řečové úlohy od celkem 245 dětí. Byly vyhodnoceny celkem 4 veličiny související s tempem řeči, a to celkové tempo řeči SR a artikulační tempo AR ve volné promluvě, dále tempo řeči v recitaci říkanky RR a tempo řeči diadochokinetické promluvě DDKR. Všechny tyto veličiny vykazují silnou věkovou závislost, která byla ohodnocena pomocí Spearmanova korelačního koeficientu, viz tabulka 5.1. Naopak vliv pohlaví dítěte na tempo řeči nalezen nebyl. Dále byly vyhodnoceny vzájemné korelace uvedených veličin, viz tabulka 5.2, z čehož vychází, že jednotlivé veličiny jsou spolu vzájemně svázány, i když korelace s DDKR vychází nižší než u zbylých veličin.

Dalším výsledkem jsou normativní data tempa řeči a artikulačního tempa v dětských promluvách pro češtinu. Střední hodnoty všech zkoumaných veličin v jednotlivých věkových kategoriích jsou vyneseny v grafu na obrázku 5.1. Z výsledků jasně vyplývá, že tempo řeči je nejvyšší v DDK promluvách, což je mimo jiné dáno i zadáním úlohy, kdy subjekty mají za úkol opakovat slabiky /pa/-/ta/-/ka/ co nejrychleji. Dále se projeví i způsob, jakým je tempo měřeno, kdy v rámci této práce byly vybrány pouze ty úseky DDK promluvy, kde jsou slabiky opakovány bez přerušení a zbytek nahrávky není použit. V neposlední řadě je to dáno tím, že subjekty nemusí přemýšlet o tom co říkají, jedná se tedy o automatickou promluvu a reflektuje tak pouze artikulační schopnosti dítěte. Dalším příkladem automatické řečové úlohy je recitace říkanky. V ní je tempo nižší než u DDK promluvy zejména proto, že děti měly za úkol říkanku recitovat v libovolném tempu, nikoliv co nejrychleji jako v případě DDK promluvy. Dále mohou mít vliv občasné nespojitosti, které při zpracování promluv z této úlohy nebyly odstraněny.

Ve volné promluvě byly analyzovány dva parametry, a to artikulační tempo a celkové tempo řeči. Artikulační tempo ve volné promluvě vychází pro děti přibližně od 11 let velmi podobně jako tempo řeči v říkance. Naproti tomu ve věku od 5 do 10 let jsou hodnoty o něco vyšší než v říkance. Celkové tempo řeči ve volné promluvě vychází ze všech zkoumaných veličin nejnižší. Největší podíl na tom bude mít zejména fakt, že se v promluvách nachází velké množství řečových pauz, kdy si subjekt nejprve rozmýšlí, co chce říci.

Normativní data všech 4 řečových parametrů naznačují, že k jejich největšímu nárůstu dochází ve věku přibližně do 11 let. Dále je nárůst výrazně pomalejší, v některých pří-

padech téměř nepatrný. To naznačuje, že závislost tempa řeči na věku není lineární, ale v určitých fázích vývoje dítěte se mění. Nelinearitu závislosti tempa řeči na věku dítěte naznačují i výsledky některých studií, zabývajících se dětmi v předškolním věku [1], [12].

V další části práce byl navržen algoritmus pro detekci slabičných jader v promluvě, což je jeden z přístupů umožňujících odhad tempa řeči v promluvě. Společně s dvěma volně dostupnými algoritmy a metodou založenou na využití fonémového rozpoznávače, byl navržený algoritmus otestován na databázi dětských promluv. Výsledky naznačují, že nejpřesnější výsledky detekce slabik v promluvě jsou dosažitelné za použití fonémového rozpoznávače v případě volné promluvy a recitace říkanky. Naopak v DDK promluvě je fonémový rozpoznávač mezi hodnocenými algoritmy až na posledním místě. To může být způsobeno tím, že se jedná o velmi atypický druh promluvy, která neobsahuje žádná slova, ale pouze jednotlivé slabiky. Úkolem v této promluvě bylo opakovat slabiky co nejrychleji, což mohlo způsobit, že samohlásky nebyly artikulovány dostatečně jednoznačně, aby je rozpoznávač dokázal detekovat. To by vysvětlovalo, proč rozpoznávač v těchto promluvách samohlásku často označil za hezitaci, která může svými akustickými vlastnostmi samohlásku připomínat. U DDK promluv vychází jako nejpřesnější metoda Theta Oscilátor, který v dalších typech promluv patřil mezi slabší metody.

Literatura

[1] K. D. Hall, O. Amir a E. Yairi. "A Longitudinal Investigation of Speaking Rate in Preschool Children Who Stutter". In: *Journal of Speech, Language, and Hearing Research* 42.6 (1999), s. 1367–1377.

[2] K. J. Logan et al. "Speaking rate characteristics of elementary-school-aged children who do and do not stutter". In: *Journal of Communication Disorders* 44 (2006), s. 130–147.

[3] J. Janda. "Posuzování logopedického věku dítěte". Disertační práce. Fakulta elektrotechnická, ČVUT v Praze, 2012.

[4] G. J. T. Haselager, I. H. Slis a A. C. M. Rietveld. "An alternative method of studying the development of speech rate". In: *Clinical Linguistics & Phonetics* 5.1 (1991), s. 53–63.

[5] J. F. Walker a L. M. D. Archibald. "Articulation rate in preschool children: a 3-year longitudinal study". In: *International Journal of Language & Communication Disorders* 41.5 (2006), s. 541–565.

[6] J. Psutka et al. *Mluvíme s počítačem česky*. Academia, 2006. ISBN: 80-200-1309-1.

[7] J. Uhlíř et al. *Technologie hlasových komunikací*. Nakladatelství ČVUT, 2007. ISBN: 978-80-01-03888-8.

[8] O. Amir a D. Grinfeld. "Articulation rate in childhood and adolescence: Hebrew speakers". In: *Language and Speech* 54 (2011), s. 225–240.

[9] J. F. Walker et al. "Articulation Rate in 3- and 5-Year-Old Children". In: *Journal of Speech, Language, and Hearing Research* 35.1 (1992), s. 4–13.

[10] S. Lee, A. Potamianos a S. Narayanan. "Acoustics of children's speech: Developmental changes of temporal and spectral parameters". In: *The Journal of the Acoustical Society of America* 105.3 (1999), s. 1455–1468.

[11] M. Icht a B. M. Ben-David. "Oral-diadochokinesis rates across languages: English and Hebrew norms". In: *Journal of Communication Disorders* 48 (2014), s. 27–37. ISSN: 0021-9924.

[12] R. H. Pindzola, M. M. Jenkins a K. J. Lokken. "Speaking Rates of Young Children". In: *Language, Speech, and Hearing Services in Schools* 20.2 (1989), s. 133–138.

[13] J. A. Sturm a C. H. Seery. “Speech and Articulatory Rates of School-Age Children in Conversation and Narrative Contexts”. In: *Language, Speech, and Hearing Services in Schools* 38.1 (2007), s. 47–59.

[14] S. Kowal, D. C. O’Connell a E. J. Sabin. “Development of temporal patterning and vocal hesitations in spontaneous narratives”. In: *Journal of Psycholinguistic Research* 4.3 (1975), s. 195–207.

[15] N. Tomashenko a Y. Khokhlov. “Speaking Rate Estimation Based on Deep Neural Networks”. In: *Speech and Computer (Lecture Notes in Computer Science)* (2014).

[16] N. H. de Jong a T. Wempe. “Praat script to detect syllable nuclei and measure speech rate automatically”. In: *Behavior research methods* 41 (2009), s. 385–90.

[17] O. Räsänen, G. Doyle a M. C. Frank. “Pre-linguistic segmentation of speech into syllable-like units”. In: *Cognition* 171 (2018), s. 130–150.

[18] H. R Pfitzinger. “Two approaches to speech rate estimation”. In: *Proceedings of SST* 96 (1996), s. 421–426.

[19] T. Pfau a G. Ruske. “Estimating the speaking rate by vowel detection”. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2 (1998), 945–948 vol.2.

[20] R. Faltlhauser, T. Pfau a G. Ruske. “On-line speaking rate estimation using Gaussian mixture models”. In: *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* 3 (2000), s. 1355–1358.

[21] D. Wang a S. S. Narayanan. “Robust speech rate estimation for spontaneous speech”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.8 (2007), s. 2190–2201.

[22] C. Yarra, O. D. Deshmukh a P. K. Ghosh. “A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection”. In: *Speech Communication* 78 (2016), s. 62–71.

[23] T. Dekens et al. “Speech rate determination by vowel detection on the modulated energy envelope”. In: *2014 22nd European Signal Processing Conference (EUSIPCO)* (2014), s. 1252–1256.

[24] S. Nayak, S. Bhati a K. S. Rama Murty. “Zero Resource Speaking Rate Estimation from Change Point Detection of Syllable-like Units”. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), s. 6590–6594.

[25] P. Schwarz. “Phoneme Recognition based on Long Temporal Context”. Disertační práce. Brno University of Technology, 2009.

[26] ELRA. *Czech SpeechDat(E) Database*. 2018. URL: `http://catalog.elra.info/en-us/repository/browse/ELRA-S0094/` (cit. 03. 07. 2020).

[27] P. Boersma a D. Weenink. *Praat: doing phonetics by computer*. [Computer program]. Ver. 6.0.49. 10. břez. 2019. URL: `http://www.praat.org`.

[28] N. H. de Jong a T. Wempe. “Automatic measurement of speech rate in spoken Dutch”. In: *ACLC Working Papers* 2.2 (2007), s. 51–60.