Univerzita Karlova v Praze
Filozofická fakulta

Fonetický ústav

Studijní program Filologie
Studijní obor Fonetika

Teze disertační práce

Lenka Weingartová

Identifikace mluvčího v temporální doméně řeči

Speaker identification in the temporal domain of speech

školitel: doc. PhDr. Jan Volín, Ph.D.

2015

Obsah

1 Cíle disertační práce 3
2 Současný stav poznání 4
2.1 Temporální charakteristiky v rozpoznávání mluvčího 5
2.2 Modelování temporální struktury 5
2.3 Temporální charakteristiky češtiny: přehled dosavadního výzkumu 6
3 Metoda a materiál 8
4 Průměrné trvání českých hlásek 9
5 Rozpoznání mluvčího: tempo řeči a globální temporální ukazatele 11
6 Model temporálních charakteristik 11
Reference 13

1 Cíle disertační práce

Disertační práce má dva hlavní cíle: popsat temporální vlastnosti českých hlásek a faktory, které se podílejí na organizaci temporální struktury češtiny, a na tomto základě vytvořit model, který popisuje její referenční stav. Tento model následně poslouží jako srovnávací základ k prozkoumání individuálních odchylek jednotlivých mluvčích, které by se daly využít pro jejich identifikaci ze zvukového záznamu.

Podrobnější popis temporálních vlastností českých hlásek v deskripci češtiny, ke kterému by se mohly vztahovat nejen deskriptivní příručky, ale i odborné studie zabývající se touto problematikou, také prozatím chybí.. Poskytnutí reprezentativních hodnot trvání českých hlásek poslouží jako odrazový můstek k budoucímu zkoumání (například tempa řeči, akustických korelátů slovního přízvuku, cizineckého přízvuku v češtině apod.).

Tato studie si také klade za cíl přispět k objasnění vlivu hlavních faktorů, segmentálních i prozodických, na temporální strukturu řeči. Mnohé vlivy jsou již známy z jiných jazyků, ale výsledky pro češtinu jsou prozatím kusé či těžko zobecnitelné. Temporální model vytvořený na základě těchto výsledků by se následně mohl stát prvním krokem v popisu rytmu češtiny.

Předpokládáme, že výsledky této práce budou také přímo aplikovatelné – naším primárním cílem je postihnout individuální rozdíly, užitečné pro potřeby forenzní analýzy, sekundárně předpokládáme také využitelnost pro účely syntézy a rozpoznávání řeči. Návrhy, jak a kde hledat individuální temporální rozdíly mezi mluvčími, by měly pomoci forenzním expertům.

Využitím některých temporálních charakteristik pro rozpoznávání mluvčího se již zabýval článek *Rhythm metrics for speaker identification in Czech* (Weingartová, 2013), kde bylo prozkoumáno devět globálních temporálních ukazatelů a jeden lokální. Ukazatele globální se neukázaly jako příliš vhodné pro postižení individuálních rozdílů, naopak lokální ukazatel LAR (Volín, 2009), který zachycuje vzdálenosti slabičných jader, byl výrazně úspěšnější. Při aplikaci na koncové slabiky úseků dokázal rozpoznat různé způsoby závěrového zpomalování u tří mluvčích. Možnosti tohoto ukazatele – a lokálního pojetí rytmu vůbec – jsou však mnohem širší. LAR lze využít nejen k celkovému popisu průběhu temporálních modulací řeči u jednotlivých mluvčích, ale také k vytvoření průměrných vzorců pro temporální průběhy vět v češtině (podobně jako např. u Volína a Skarnitzla, 2007). Odchylky od těchto průměrů by pak mohly sloužit jako charakteristiky jednotlivých mluvčích (či jejich skupin).

Disertační práce je strukturována následovně: Kapitola 1 obsahuje úvod do problematiky. V kapitole 2 je čtenář seznámen se zásadními otázkami současného výzkumu rytmické struktury řeči, kapitola 3 se zabývá využitím temporálních charakteristik při rozpoznávání

mluvčího. V kapitole 4 jsou představeny přístupy k modelování temporálních charakteristik promluv. Kapitola 5 obsahuje zevrubné shrnutí dosavadních studií zabývajících se trváním českých hlásek. V kapitole 6 jsou nastoleny výzkumné otázky a hypotézy a kapitola 7 představuje využitou metodu a řečová data. Deskriptivní statistiky českých hlásek jsou obsahem kapitoly 8, kapitola 9 pak na základě těchto popisů shrnuje jevy specifické pro mluvčího. Modelu trvání českých hlásek, jeho výstavbě, ladění a aplikaci na reálné promluvy je věnována kapitola 10. Kapitola 11 pak shrnuje závěry práce, její omezení a možné směry budoucího výzkumu. V těchto tezích jsou uvedeny pouze vybrané výsledky.

2 Současný stav poznání

Snaha o zachycení rozdílů mezi taktově a slabičně izochronními jazyky podnítila mnoho studií usilujících o empirickou kvantifikaci rytmu (např. Ramus et al., 1999; Grabe & Low, 2002; Asu & Nolan, 2006; Dellwo, 2006 nebo Arvaniti, 2009). Metody zde využité bývají souhrnně označovány jako „rytmické ukazatele“. Na základě trvání vokalických a konsonantických intervalů v řeči se snaží nějakým způsobem jedinou hodnotou postihnout globální temporální organizaci promluv (či celých jazyků). Výhodou těchto ukazatelů je relativní snadnost a přímočarost jejich získávání ze zvukového materiálu, nevýhodou nesnadnost lingvistické interpretace. Tyto ukazatele nicméně mohou být úspěšně využity pro rozpoznání mluvčího, viz např. Dellwo & Koreman (2008), Dellwo et al. (2012), nebo Leemann et al. (2014); pro češtinu Weingartová (2013).

Spojitý vztah mezi vnořenými prozodickými jednotkami se snaží modelovat teorie spřažených oscilátorů, k osvětlení vztahů mezi základními temporálními jednotkami rytmu je využíváno také experimentální paradigma nazvané cyklická řeč (*speech cycling*), tj. opakování jednotek do rytmu metronomu. Kooperativní funkcí rytmu se zabývají experimenty se synchronizací řeči, kromě toho existuje též široká oblast výzkumu senzorimotorické synchronizace, tj. synchronizace vnímaných stimulů s pohybem.

Kohler (2009a: 35) vyzdvihuje a zdůrazňuje percepční roli rytmu, která bývá ve výzkumu rytmické struktury často zanedbávána. Jedním z prvních průkopníků byla Lehiste (1973, 1977, citováno z Lehiste, 1979), která etablovala izochronii jako subjektivní, nikoliv objektivní jev. Posluchači neslyší objektivní trvání akustických jevů a promluvy se jim zdají pravidelnější, než doopravdy jsou (Lehiste, 1979: 244). Tento závěr potvrdili také Donovan a Darwin (1979).

V disertační práci je hlavní pozornost věnována kvantitativní analýze temporální struktury řeči, s plným vědomím toho, že nejde o rytmus jako takový. Nicméně načasování a trvání řečových jevů zcela jistě je základní a neodmyslitelnou součástí rytmu – a bez informací o temporální doméně nemůžeme na jeho komplexní popis vůbec aspirovat. V souladu s Kohlerovým novým paradigmatem pro výzkum rytmu (Kohler, 2009a) by měl následovat další krok ve formě popisu řečové prominence, jejích fyzikálních korelátů a následně též percepce. Tímto směrem výzkum v pražském Fonetickém ústavu také postupuje, především v souvislosti s popisem české angličtiny (viz Volín & Weingartová, 2014, či Weingartová et al., 2014), nicméně není předmětem práce.

2.1 Temporální charakteristiky v rozpoznávání mluvčího

V řeči existují rysy, které jsou pro každého mluvčího specifické a umožňují nám více či méně spolehlivě rozpoznat osoby pouze podle jejich hlasu. Většina výzkumu v oblasti identifikace mluvčího se soustředí na spektrální informace v řečovém proudu, využití temporálních charakteristik je spíše vzácností. Nicméně jejich přitažlivost je založena na dvou hypotézách, za prvé mluvčí mohou mít svůj vlastní naučený rytmus řeči, za druhé, dynamika pohybů mluvidel je pro mluvčího jedinečná díky fyziologickým odlišnostem.

Temporální charakteristiky mají oproti charakteristikám spektrálním tu výhodu, že jsou odolné vůči distorzím signálu šumem nebo pásmovými filtry (například telefonního přenosu). Několik zdrojů také ověřilo, že mluvčí zásadním způsobem nemění své temporální charakteristiky, když chtějí maskovat svůj hlas (Eriksson & Wretling, 1997, Wretling & Eriksson, 1998 nebo Dellwo et al., 2009), což může být ve forenzní praxi velmi výhodné.

2.2 Modelování temporální struktury

Základem pro modelování temporální struktury je popis a model trvání segmentů (zejména hlásek). Jde o klíčový parametr pro kvalitní rozpoznávání i syntézu řeči (O'Shaughnessy, 1995: 600; Lazaridis et al., 2010: 175). Důležitým mezníkem v modelování trvání hlásek byl model Dennise Klatta (1976), který jednoduchým způsobem kombinuje inherentní (tj. průměrné) a minimální, „nestlačitelné" trvání (tj. nejmenší možná doba pro uspokojivé provedení artikulačního gesta) jednotlivých hlásek s faktory, jež mají vliv na trvání výsledné. Tento model byl v Klattově (poměrně malém) korpusu schopen predikovat trvání vokálů v různých pseudoslovech. Carlson et al. (1979, citováno z Carlson, 1991) ověřovali percepční adekvátnost modelu experimentem se syntetizovanou řečí, dále byl upraven a rozšířen pro švédštinu.

Klattův model je příkladem modelu založeného na pravidlech (*rule-based*). Apriorním úsudkem nebo na základě analýzy dat formulujeme pravidla, která pak model vytvářejí. Ten je jen tak dobrý, jak přesná a vyčerpávající jsou námi stanovená pravidla a jaké množství faktorů ovlivňujících trvání jsme schopni postihnout. Takový druh výzkumu však může trvat velmi dlouho a vyžaduje fonetické znalosti a zkušenosti, což technicky založené studie považují za hlavní nevýhodu tohoto postupu. Technická větev temporálního modelování se proto uchyluje k modelům založeným na datech (*data-driven*). S pomocí statistického modelování jsou tyto systémy také schopny zachytit variabilitu v datech, někdy dokonce i úspěšněji. Nevýhodou pro naše účely je neprůhlednost těchto modelů – nedozvíme se, které řečové parametry a do jaké míry mají na trvání hlásek vliv (Kohler, 2009b: 5).

Nezbytným základem výše uvedených postupů jsou velké řečové korpusy – které je ale vzhledem k jejich obsáhlosti často nemožné manuálně zpracovávat, a tedy je nutné hranice hlásek stanovovat automaticky, což může vnést do modelu vyšší chybovost. Naopak modely založené na pravidlech jsou obvykle postaveny na materiálu ručně segmentovaném fonetiky, kterého je ale na druhou stranu řádově méně, protože jde o práci značně časově náročnou. Ideální cestou by jistě byl kompromis, nebo spíše spojení obou těchto přístupů a jejich vzájemné obohacení, k čemuž mnohé studie již léta vybízejí (např. Doddington, 1985: 1663, Carlson, 1991: 246 nebo Bourland et al., 1996). Algoritmy by bylo možné například trénovat na řečovém materiálu ručně anotovaném fonetiky – tento postup se začal úspěšně využívat již i v českém prostředí (Pollák et al., 2007).

2.3 Temporální charakteristiky češtiny: přehled dosavadního výzkumu

Temporálním vlastnostem češtiny se v průběhu historie fonetického výzkumu věnovala řada odborníků. Prvním, kdo se pokusil o systematický popis trvání hlásek v češtině, byl průkopník české fonetiky, Josef Chlumský. Ve svém *Pokusu o měření českých zvuků a slabik v řeči souvislé* (1911) uvádí hodnoty trvání hlásek u dvou mluvčích češtiny, z nichž jedním byl on sám. Tyto poznatky dále rozvádí v rozsáhlejší práci *Kvantita, melodie a přízvuk* (1928), kdy již měl k dispozici mluvčí tři. Chlumský se zabýval jednotlivými mluvčími i temporálními jevy postupně a velmi podrobně. Jeho celkové výsledky přehledně shrnuje Hála v *Uvedení do fonetiky češtiny na obecně fonetickém základě* (1962). I přes stáří těchto studií (tehdy například nebylo možné měřit s přesností na milisekundy, jak je to obvyklé dnes) se stále jedná o počin zásadní, který snese srovnání s výsledky novějších kvantitativních studií.

V šedesátých letech se trváním hlásek kromě drobné studie o českých a holandských vokálech L. Kaiserové s využitím výzkumu Přemysla Janoty (Kaiser, 1964) zabývala ještě práce

Borovičkové a Maláče *The Spectral Analysis of Czech Sound Combinations* (1967), která se (pravděpodobně) stala východiskem též pro hodnoty trvání hlásek uváděné v Mluvnici češtiny (Petr et al., 1986). Přehledné shrnutí všech těchto výsledků lze nalézt v Palkové (1994).

Na tento směr výzkumu bylo po dlouhé odmlce navázáno teprve po obnovení pražského Fonetického ústavu v 90. letech – pro novější empirický výzkum trvání českých hlásek je třeba nahlédnout do kvalifikačních prací absolventů, plošná studie, která by replikovala Chlumského výsledky na bohatším materiálu, zatím provedena nebyla.

Trvání slabičných likvid se věnovala Vernerová (2006); Machač (2006) a Šimek (2010) popisovali trvání českých exploziv a Homolková (2009) frikativ. Několik relevantních údajů můžeme najít též u Ondruškové (2011). Hodnoty týkající se vybraných vokalických elementů nalezneme v disertační práci Studenovského (2012) o českých diftonzích.

Geografickou výjimkou je výzkum V. J. Podlipského z Univerzity Palackého v Olomouci, který se zabývá kontrastem kvantity českých vokálů, také v kontextu osvojování angličtiny jako druhého jazyka (viz např. Podlipský, 2009 nebo Podlipský et al., 2009).

Jediným novějším zdrojem, kde je možné nalézt referenční hodnoty pro všechny české hlásky, je kniha J. Psutky a jeho spolupracovníků *Mluvíme s počítačem česky* (Psutka et al., 2006), kteří údaje o trvání hlásek extrahovali z rozsáhlého strojově zpracovaného korpusu pro potřeby řečových technologií, zejména syntézy řeči.

Co se týče typu využitého materiálu, je v těchto pracích zřetelný přechod od regulovaného laboratorního materiálu a pseudoslov přes čtené projevy až k semispontánním a spontánním promluvám.

Spolehlivost údajů o trvání hlásek se nutně opírá o spolehlivost označení jejich hranic, což není problém triviální. Při manuální segmentaci, která je využívána ve fonetickém výzkumu, může hrát velkou roli osoba anotátora, jeho zkušenosti, přesnost a konzistentnost – segmentace různých osob nemusí být přímo srovnatelné a mohly by do výsledků zavádět artefakty. Z důvodu potřeby jednotné koncepce vznikla v roce 2009 práce *Fonetická segmentace hlásek* (Machač & Skarnitzl, 2009), jež poskytuje souhrnná pravidla pro označování hranic segmentů. Novější studie z Fonetického ústavu a studie Podlipského a jeho kolegů (2009) se těmito pravidly řídí, a tedy můžeme s rozumnou mírou jistoty prohlásit tyto údaje za přímo srovnatelné.

Naproti tomu Psutka a kolegové (Psutka et al., 2006) vzhledem k velikosti korpusu určovali trvání hlásek automaticky na základě tzv. nuceného zarovnávání (*forced alignment*) textu ke zvuku, což vedlo k vysoké variabilitě hodnot trvání. Směrodatné odchylky jejich trvání jsou řádově dvoj- až trojnásobné než u materiálu segmentovaného ručně.

3 Metoda a materiál

Zvukový materiál využitý v této práci pochází z korpusu *Minidialogy*, subkorpusu *H*, který je v současné době stále rozvíjen ve Fonetickém ústavu FF UK v Praze. Korpus obsahuje nahrávky čtených dialogů od 34 mluvčích, z toho 26 žen a 8 mužů ve věku 20–25 let. Přečtených dialogů je celkem 24, text korpusu obsahuje celkem 119 unikátních replik. Celkový počet replik analyzovaných v této práci je tedy 4046; analyzovaných slov je přes 27 000. Řečový materiál trvá celkem déle než dvě hodiny.

Nahrávky byly ručně označkovány v programu *Praat* (Boersma & Weenink, 2014) na úrovni prozodických frází, slov, slabik, typů hlásek, hlásek a fonémů. Vrstva slov a hlásek byla vytvořena automaticky programem *Prague Labeller* (Pollák et al., 2007), hranice následně byly manuálně opraveny, stejně jako transkripce hlásek.

Oproti zvyklostem u jiných korpusů budovaných na Fonetickém ústavu jsme se u tohoto korpusu rozhodli nezarovnávat hranice na průchod zvukové vlny v oscilogramu nulou, protože by tato změna mohla deformovat trvání hlásek. ze stejného důvodu jsme se rozhodli pauzy na prozodických předělech označovat i v případech, kdy jejich trvání nedosahovalo 120 ms. I velmi krátká pauza rozdělená mezi okolní hlásky by kontaminovala jejich trvání.

Celkově analyzujeme 112 920 hlásek. Z každé položky byly extrahovány následující informace: kód repliky, kód mluvčího, identita hlásky, typ hlásky (C, V, R), trvání hlásky (v ms), trvání hlásky vzhledem k trvání slabiky (v %), trvání hlásky vzhledem k trvání slova (v %), předcházející hláska, následující hláska, foném(y) přináležející hlásce, slabika, v níž se segment nalézá, struktura slabiky (např. CVC), pozice hlásky ve slabice (prétura/nukleus/koda), trvání slabiky (v ms), pozice slabiky ve slově (iniciální/mediální/finální/individuální), pozice hlásky ve slově (iniciální/mediální/finální/individuální), slovo, trvání slova (v ms), délka slova (ve slabikách a fonémech), pozice slova v prozodické frázi (iniciální/mediální/finální/individuální), délka prozodické fráze (v ms, ve slabikách a v grafických slovech) a hloubka následujícího prozodického předělu (3 nebo 4).

Pro potřeby kvantifikace tempa řeči používáme artikulační tempo, tedy trvání lingvistické jednotky za jednotku času s vyřazením pauz. Průměrné artikulační tempo jednoho mluvčího je spočítáno jako průměr artikulačních temp ve všech jím pronesených replikách.

Dále měříme rytmické ukazatele, pro něž zavádíme název globální temporální ukazatele, a to tyto: %V, ΔV/ΔC, VarcoV/VarcoC, rPVI-V/rPVI-C a nPVI-V/nPVI-C.

Lokální artikulační tempo kvanitifkujeme ukazatelem LAR převzatým od Volína (2009). Je spočítán jako převrácená hodnota vzdálenosti onsetů dvou po sobě následujících slabičných jader.

4 Průměrné trvání českých hlásek

V tabulce 3 jsou uvedeny reprezentativní průměrná trvání vokálů z korpusu Minidialogy-H.

	Počet	Průměr	Sm.odch.	Minimum	Maximum	Medián	10. percent.	90. percent.
Krátké vokály								
a	3151	**58**	**14,8**	12,6	131,1	56,9	39,9	77,1
e	9010	**49,9**	**14,6**	9	151,7	48,5	32	69
i	3624	**45,6**	**14,6**	7	116,3	44,1	28,4	64,5
o	4984	**47**	**13,9**	15,2	136	46	30,4	65,1
u	1632	**45**	**13,8**	8,8	105,2	43,9	28,8	63,2
Dlouhé vokály								
a:	1691	**93**	**23**	20,6	193,4	92,2	64,6	122
e:	89	**99,8**	**19**	66,4	166,3	96,6	80,2	126,4
i:	1824	**53,8**	**17,3**	7,7	145,4	51,8	33	76,6
u:	100	**60,7**	**25,5**	22,2	153,4	55,3	33,3	96,2
Diftong								
ou	549	**80,3**	**18,1**	32,5	157,7	80	56,3	102,6
Celkem	26654	**53,3**	**19,7**	7	193,4	50	32	78,2

Tabulka 3: Reprezentativní hodnoty trvání českých vokálů naměřené na vybraných částech korpusu Minidialogy-H. Hodnoty ve třetím až devátém sloupci jsou uvedeny v milisekundách.

U krátkých vokálů je zřetelně vidět tendence ke kratšímu trvání zavřených vokálů, oproti otevřenému [a], které je průměrně nejdelší. Středové vokály [e] a [o] pak spadají mezi ně. Srovnáme-li tyto údaje s dřívějšími výsledky z literatury (oddíl 5.1.1, obr. 5.1a), vidíme, že co se týče rozsahu hodnot trvání (krátké vokály trvají mezi 45 a 58 ms), nejvíce se blížíme hodnotám Podlipského et al. (2009), ostatní jsou výrazně delší. Rozdíly mezi trváním podle otevřenosti vokálu pak jsou nejpodobnější Chlumskému (1928) a Kaiserové (1964).

U dlouhých vokálů tuto tendenci porušuje [e:], které je průměrně nejdelší. Zavřené [i:] a [u:] však zůstává nejkratší. Porovnáme-li to opět se staršími daty (obr. 5.1b), rozsahem hodnot se pohybujeme opět poblíž výsledků Podlipského et al. (2009). Rozdíly mezi jednotlivými vokály jsou ovšem jiné než v citovaných studiích, nejkratší trvání [i:] a [u:] nicméně zůstává.

Průměrné trvání diftongu [o͡u] je o více než 20 ms kratší než v ostatních studiích.

Tabulka 4 zobrazuje průměrné reprezentativní hodnoty pro konsonanty.

Konsonant	Zdroj	Počet	Průměr	Sm.odch.	Minimum	Maximum	Medián	10. percent.	90. percent.
explozivy									
p	bez shluků	708	**78,1**	**18,7**	21,6	178,9	78,4	55,4	99,9
b	intervokal.	1123	**58,7**	**15,3**	8,7	128,3	58,6	39,9	77,2
t	intervokal.	2387	**68,1**	**17,5**	16,8	189,8	67,2	48,0	88,4
d	intervokal.	1089	**30,7**	**10,4**	9,0	94,2	29,3	19,1	45,5
ť	bez shluků	694	**71,9**	**20,4**	19,7	148,3	71,3	47,5	96,5
ď	bez shluků	376	**52,1**	**18,3**	13,0	120,0	50,5	30,5	76,2
k	bez shluků	1439	**64,4**	**19,6**	20,4	274,0	63,0	42,3	85,5
g	bez shluků	270	**43,4**	**15,5**	7,9	114,1	42,2	25,9	64,0
frikativy									
f	bez shluků	396	**59,6**	**18,5**	21,8	134,0	58,1	35,6	83,9
v	bez shluků	576	**41,6**	**13,1**	11,6	96,8	40,8	25,7	58,5
s	intervokal.	996	**92,7**	**19,1**	32,0	160,2	91,0	70,3	118,2
z	bez shluků	535	**52,9**	**17,8**	9,2	128,0	53,2	28,3	74,1
š	bez shluků	1358	**71,8**	**26,2**	19,7	188,8	69,8	39,7	105,2
ž	bez shluků	988	**54,9**	**19,5**	11,7	134,6	54,5	29,8	80,0
ch	bez shluků	398	**65,8**	**21,7**	22,2	135,0	63,2	38,8	95,8
h	bez shluků	151	**55,4**	**21,6**	11,3	143,3	53,0	32,6	79,8
afrikáty									
c	bez shluků	809	**87,9**	**21,2**	29,4	180,7	87,9	60,8	112,0
dz	bez shluků	82	**66,7**	**21,2**	27,1	138,6	65,9	41,1	95,0
č	bez shluků	431	**101,5**	**21,1**	46,4	173,3	101,4	76,1	127,6
dž	vše	23	**63,7**	**23,2**	33,4	127,5	52,9	37,8	98,2
nazály									
m	intervokal.	1202	**58,0**	**15,4**	11,5	159,9	58,2	39,0	75,4
ɱ	vše	27	**83,1**	**23,8**	32,0	130,1	83,5	56,2	117,3
n	bez shluků	1979	**43,4**	**18,4**	7,1	274,7	40,8	24,0	64,0
ň	bez shluků	1094	**50,5**	**19,0**	9,3	146,4	47,9	29,1	72,7
ŋ	vše	34	**86,9**	**16,4**	50,2	114,4	88,6	68,3	111,5
vibranty									
r nesl.	bez shluků	330	**41,5**	**10,6**	16,7	83,4	40,5	29,1	55,1
r slabičné	vše	273	**71,1**	**14,6**	23,9	117,8	71,9	51,6	88,4
ř	bez shluků	189	**53,8**	**16,6**	23,6	121,1	49,5	35,5	77,3
aproximanty									
l	intervokal.	1015	**37,7**	**11,2**	8,4	100,8	36,8	24,2	52,6
j	intervokal.	916	**28,6**	**12,3**	7,2	146,3	26,4	15,7	43,7

Tabulka 4: Reprezentativní trvání českých konsonantů naměřené na vybraných částech korpusu Minidialogy-H. Sloupec Zdroj uvádí, ze kterých hláskových okolí byly průměry vypočítány. Hodnoty ve čtvrtém až desátém sloupci jsou uvedeny v milisekundách.

5 Rozpoznání mluvčího: tempo řeči a globální temporální ukazatele

Artikulační tempo se projevilo jako dobrý ukazatel identity mluvčího. Vykazuje velkou variabilitu napříč subjekty a malou v rámci jednoho subjektu. Rovněž dle našich výsledků je slabičné tempo o něco lepším ukazetelem než tempo hláskové, které je pravděpodobně více ovlivněno hláskovou stavbou textu (jež byla pro všechny mluvčí stejná), zatímco na slabičném tempu se mohou více projevit individuální rytmické preference mluvčích.

Celkové průměrné artikulační tempo v prozkoumaném materiálu je 6,54 slabik za sekundu (sl/s) a 15,55 hlásek za sekundu (hl/s). Nejnižší naměřené artikulační tempo je 3,53 sl/s a 7,38 hl/s (pro jednotlivou repliku) a 6,03 sl/s a 14,29 hl/s (pro jednotlivého mluvčího, konkrétně mluvčí REBA). Naopak nejvyšší naměřené tempo v jedné replice bylo 9,56 sl/s a 21,77 hl/s a průměrně má nejvyšší artikulační tempo mluvčí MORC – 7,38 sl/s a 17,51 hl/s. Stejného slabičného tempa dosahuje i mluvčí VASA, ta má ale o něco nižší hláskové tempo.

Směrodatné odchylky artikulačního tempa jednotlivých mluvčích jsou poměrně nízké, pohybují se mezi 0,7 a 1,1 sl/s a mezi 1,5 a 2,6 hl/s.

Ve shodě s výsledky z literatury (viz např. Byrd, 1994; Künzel et al., 1995 nebo Jacewicz et al., 2010) mají i v našem materiálu muži průměrně vyšší artikulační tempo než ženy. Efekt mluvčího je také vysoce významný: $F(33, 3935) = 23{,}4$; $p < 0{,}001$. ANOVA se závislou proměnnou slabičné tempo od sebe odliší 248 dvojic mluvčích z 561 možných, což odpovídá 44,2 %. Zároveň je slabičné tempo stabilní v rámci mluvčího.

Co se týče globálních temporálních ukazatelů (neboli rytmických ukazatelů), ženy vykazovaly vyšší variabilitu vokalických intervalů, muži naopak vyšší variabilitu konsonantických intervalů. Forenzně nejúspěšnější ze zkoumaných ukazatelů bylo %V, rPVI-C a ΔC. Tyto tři ukazatele jednak odhalují variabilitu mezi mluvčími, na druhou stranu jsou ale poměrně stabilní vzhledem k rozdílům v rámci mluvčího. Ovšem u rPVI-C a ΔC je třeba obezřetnosti, jelikož oba ukazatele jsou nenormalizované, a tedy v sobě do určité míry přenášejí i rozdíly v artikulačním tempu.

6 Model temporálních charakteristik

Na základě získaných měření trvání hlásek a jeho modifikací v závislosti na různých faktorech byl vytvořen temporální model, který předpovídá trvání každé hlásky. Tento model je pravidlový a v principu podobný modelu Klattově (1976). Vstupem do modelu je výchozí trvání jednotlivých hlásek, které je následně násobeno osmi různými koeficienty. Jsou to

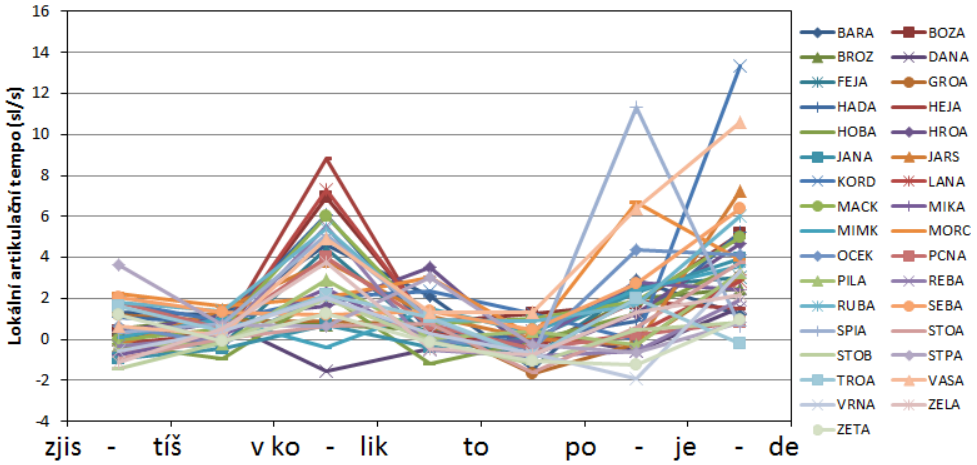
koeficienty pozice ve slabice, slova v prozodické frázi, finální slabiky ve finálním slově, hláskového okolí, trvání vokálů po rázu, délky slova ve slabikách, délky prozodické fráze a struktury slabiky. Modelová hodnota každé hlásky je spočítána podle vzorce:

$$T_m = T_v \times \prod_{f=1}^{8} K_f,$$

kde T_m je modelové trvání hlásky, T_v je výchozí trvání hlásky a K_f je koeficient faktoru f.

Odladěný model má průměrnou odchylku od reálných hodnot 0,9 ms a průměr z absolutních hodnot odchylek je 16,54 ms. Obsahuje 860 koeficientů, které se nerovnají jedné. Z toho téměř 60 % (514) spadá mezi 0,9 a 1,1 včetně. Celkově se hodnoty koeficientů pohybují mezi 0,5 a 2,1, což odpovídá zhruba polovičnímu a dvojnásobnému trvání. Lze tedy říci, že modifikace trvání (zkracování a prodlužování) je poměrově symetrická.

Modelová trvání jsou použita k výpočtu modelových kontur lokálního artikulačního tempa (LAR; Volín, 2009) a individuálních reziduí. Rezidua repliky H3b_3 jsou zobrazena na obrázku 1 níže.



Obrázek 1: Graf kontur reziduí lokálního artikulačního tempa pro větu H3b_3, *Zjistíš, v kolik to pojede?* Reziduum je vypočítáno jako rozdíl modelové a reálné hodnoty LAR. Nula zastupuje modelové hodnoty LAR, kladná čísla znamenají zrychlení a záporná zpomalení lokálního artikulačního tempa oproti modelu.

Z analýz kontur lokálního artikulačního tempa vyplývá, že mluvčí se od sebe liší i v tom, jak zacházejí s artikulačním tempem v průběhu jedné promluvy, a to dokonce i v rámci jediného slova. Tyto odlišnosti přitom nejsou náhodné – mluvčí nevariují na libovolných místech v promluvě. Je to dané primárně segmentálním obsahem promluvy (tj. identitou hlásek), sekundárně také délkou přízvukového taktu. Naopak se zdá, že kategorie slova (autosémantické vs. synsémantické) roli nehraje. K největší individuální variabilitě přitom dochází v okolí intervokalických znělých exploziv (především [d]) a intervokalických sonor

(zejména [j]). V žádné z replik se nestalo, že by stejná pozice LAR, případně stejné slabiky, byly jednou místem s velkou variabilitou a jindy místem stabilním. V těch částech čtveřic vět, které mají stejný text, jsou místa s největší variabilitou (a obvykle také místa s největší stabilitou) stejná.

Temporální model popsaný v disertační práci velmi dobře obstojí ve srovnání s dalšími modely zmíněnými v literatuře. Model Carlsonův (1991) předpovídal hodnoty trvání hlásek se směrodatnou odchylkou 20 ms. Náš model je o něco úspěšnější, průměrná odchylka rozdílů modelových hodnot od reálných činí 16,5 ms. Stejně tak lze úspěšnost porovnat i s modernějším statistickým modelem Lazaridise et al. (2010), který uvádí korelaci modelových a skutečných hodnot mezi 0,6 a 0,8. Náš temporální model dosahuje srovnatelné korelace 0,66. Tyto výsledky jsou o to cennější, že oba citované modely byly vytvořeny pouze na základě jediného mluvčího, zatímco zde je variabilita mnohem vyšší, modelujeme 34 mluvčích.

Vytvořený temporální model najde uplatnění i jinde než ve zkoumání specifik mluvčích. Užitečným může být například při syntéze nebo rozpoznávání řeči, také může sloužit jako srovnávací měřítko při různých dalších výzkumech prozodické struktury, cizineckého přízvuku, řečových vad, a podobně – díky němu je možné si vytvořit představu o tom, co je v češtině obvyklé a co je již vybočující.

Reference

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica, 66,* 46–63.

Asu, E. L. & Nolan, F. (2006). Estonian and English rhythm: a twodimensional quantification based on syllables and feet. In: *Proceedings of Speech Prosody 2006*, Dresden, Germany.

Boersma, P. & Weenink, D. (2014): *Praat: doing phonetics by computer* [počítačový program], verze 5.3.71. Získáno z http://www.praat.org/.

Borovičková, B. & Maláč, M. (1967): *The Spectral Analysis of Czech Sound Combinations.* Praha: Academia.

Bourland, H., Hermansky, H. & Morgan, N. (1996): Towards increasing speech recognition error rates. *Speech Communication, 18/3,* 205–231.

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication, 15,* 39–54.

Carlson, R. (1991): Duration models in use. *Proceedings of the XIIth ICPhS,* 278–281, Aix-en-Provence.

Carlson, R., Granström, B., & Klatt, D. H. (1979): Some notes on the perception of temporal patterns in speech. In Lindblom, B., & Öhman, S. (eds.), *Frontiers of Speech Communication Research,* 223–243. London: Academic Press.

Dellwo, V. (2006). Rhythm and Speech Rate: A Variation Coefficient for C. In: P. Karnowski & I. Szigeti (eds.), *Language and Language Processing*, 231–241. Frankfurt am Main: Peter Lang.

Dellwo, V. & Koreman, J. (2008). How speaker idiosyncratic is measurable speech rhythm? *Proceedings of IAFPA 2008*. Lausanne: IAFPA.

Dellwo, V., Ramyead, S. & Dankovičová, J. (2009): The influence of voice disguise on temporal characteristics of speech. *Proceedings of IAFPA 2009*. Cambridge: IAFPA.

Dellwo, V., Kolly, M. & Leemann, A. (2012). Speaker identification based on temporal information: A forensic phonetic study of speech rhythm and timing in the Zurich variety of Swiss German. *Proceedings of IAFPA 2012*. Santander: IAFPA.

Donovan, A. & Darwin C. J. (1979): The perceived rhythm of speech. *Proc. 9th ICPhS, vol. II*, 268–274, Copenhagen.

Eriksson, A. & Wretling, P. (1997). How flexible is the human voice? A case study of mimicry. *Proceedings of Eurospeech, 97*, 1043–1046.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In: N. Warner, & C. Gussenhoven (eds.), *Papers in laboratory phonology 7*, 515–546. Berlin: Mouton de Gruyter.

Hála, B. (1962): *Uvedení do fonetiky češtiny na obecně fonetickém základě.* Praha: Československá akademie věd.

Homolková, V. (2009): *Temporální vlastnosti českých frikativ.* Diplomová práce. Praha: Fonetický ústav FF UK.

Chlumský, J. (1911): *Pokus o měření českých zvuků a slabik v řeči souvislé.* Praha: Česká akademie.

Chlumský, J. (1928): *Česká kvantita, melodie a přízvuk.* Praha: Česká akademie věd a umění.

Jacewicz, E., Fox, R. A. & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *Journal of the Acoustical Society of America, 128*, 839–850.

Kaiser, L. (1964): Phonetic similarity apart from linguistic affinity. *Zeitschrift für Phonetik 17*: 243–249.

Klatt, D. H. (1976): Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59*, 1208–1221.

Kohler, K. J. (2009a): Rhythm in Speech and Language: A New Research Paradigm. *Phonetica, 66*, 29–45.

Kohler, K. J. (2009b): Whither speech rhythm research? *Phonetica, 66*, 5–14.

Künzel, H. J., Masthoff, H. R. & Köster, J.-P. (1995). The relation between speech tempo, loudness, and fundamental frequency: an important issue in forensic speaker recognition. *Science & Justice, 35*, 291–295.

Lazaridis, A., Ganchev, T., Kostoulas, T., Mporas, I. & Fakotakis, N. (2010): Phone duration modeling: Overview of techniques and performance optimization via feature selection in the context of emotional speech. *International Journal of Speech Technology 13/3*, 175–188.

Leemann, A., Kolly, M.-J. & Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International, 238,* 59–67.

Lehiste, I. (1973): Rhythmic units and syntactic units in production and perception. *Journal of the Acoustical Society of America, 54,* 1228–1234.

Lehiste, I. (1977): Isochrony reconsidered. *Journal of Phonetics, 5,* 253–263.

Lehiste, I. (1979): Temporal relations within speech units. *Proc. 9th ICPhS,* vol. II, 241–244.

Machač, P. (2009): *Temporální a spektrální struktura českých explozív.* Disertační práce. Praha: Fonetický ústav.

Machač P. & Skarnitzl R. (2009): *Fonetická segmentace hlásek.* Praha: Nakladatelství Epocha.

Ondrušková, L. (2011): *Zvukové vlastnosti jednoslabičných slov v semispontánním dialogu a hlasitém čtení.* Diplomová práce. Praha: Fonetický ústav.

O'Shaughnessy, D. (1995): Timing Pattersn in Fluent and Disfluent Spontaneous Speech. *Proceedings IEEE Conference on Acoustics, Speech and Signal Processing, vol. I,* 600–603.

Palková, Z. (1994): *Fonetika a fonologie češtiny.* Praha: Karolinum.

Petr, J. (ed.) et al. (1986). *Mluvnice češtiny I: Fonetika, fonologie, morfonologie a morfemika, tvoření slov.* Praha: Academia.

Podlipský, V. J. (2009): *Reevaluating perceptual cues: Native and non-native perception of Czech vowel quantity.* Disertační práce. Olomouc: Katedra anglistiky a amerikanistiky.

Podlipský V. J., Skarnitzl R. & Volín J. (2009): High Front Vowels in Czech: a Contrast in Quantity or Quality? *Proceedings of Interspeech 2009,* 132–135. Brighton: ISCA.

Pollák P., Volín J. & Skarnitzl R. (2007): HMM-Based Phonetic Segmentation in Praat Environment. *Proceedings of the XIIth International Conference "Speech and computer – SPECOM 2007",* 537–541, Moscow.

Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives. *Proceedings of Speech Prosody 2002,* 115–120, Aix-en-Provence.

Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73/3,* 265–292.

Russell, M. J. & Cook, A. E. (1987): Experimental evaluation of duration modeling techniques for automatic speech recognition. In *Proc. Int. Conf. Acoust., Speech, Signal Process.,* 2376–2379.

Studenovský, D. (2012): *Akustické vlastnosti českých diftongů.* Disertační práce. Praha: Fonetický ústav.

Šimek, J. (2010): *Explozívy v češtině: temporální vlastnosti a variabilita při realizaci.* Diplomová práce. Praha: Fonetický ústav.

Vernerová, T. (2006): *Trvání slabikotvorných likvid v češtině.* Diplomová práce. Praha: Fonetický ústav.

Volín, J. (2009): Metric warping in Czech newsreading. In: R. Vích (ed.), *Speech Processing – 19th Czech-German Workshop,* 52–55, Praha.

Volín, J. & Skarnitzl, R. (2007): Temporal downtrends in Czech read speech. *Proceedings of Interspeech 2007,* 442–445, Antwerpen.

Volín, J. & Weingartová, L. (2012): Idiosyncrasies in local articulation rate trajectories in Czech. *Proceedings of Perspectives on Rhythm and Timing,* 67. Glasgow.

Volín, J. & Weingartová, L. (2014): Acoustic correlates of word stress as a cue to accent strength. *Research in Language, 12/2,* 175–183.

Weingartová, L. (2013): Rhythm metrics for speaker identification in Czech. *AUC Philologica 1/2014, Phonetica Pragensia XIII,* 33–42.

Weingartová, L., Poesová, K. & Volín, J. (2014): Prominence contrasts in Czech English as a predictor of learner's proficiency. In: N. Campbell, D. Gibbon & D. Hirst (eds.), *Proceedings of the 7th International Conference on Speech Prosody,* 236–240. Dublin: TCD.

Wretling, P. & Eriksson, A. (1998): Is articulatory timing speaker specific? – evidence from imitated voices. *Proc. FONETIK 98,* 48–52.