VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV RADIOELEKTRONIKY

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF RADIO ELECTRONICS

SIGNÁLOVÁ ANALÝZA MLUVENÝCH SOUHLÁSEK

SIGNAL ANALYSIS OF SPOKEN CONSONANTS

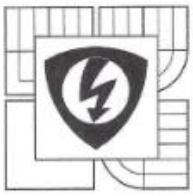
SEMESTRÁLNÍ PRÁCE

SEMESTRAL THESIS

AUTOR PRÁCE — PETR JURÁK
AUTHOR

VEDOUCÍ PRÁCE — prof. Ing. MILAN SIGMUND, CSc.
SUPERVISOR

BRNO 2014



VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ
Fakulta elektrotechniky
a komunikačních technologií
Ústav radioelektroniky

Bakalářská práce

bakalářský studijní obor
Elektronika a sdělovací technika

Student: Petr Jurák ***ID:*** 144850
Ročník: 3 ***Akademický rok:*** 2013/14

NÁZEV TÉMATU:

Signálová analýza mluvených souhlásek

POKYNY PRO VYPRACOVÁNÍ:

Seznamte se s fonetickými vlastnostmi českých souhlásek a s problematikou rozpoznávání řečových signálů. Vytvořte přehledovou studii o nejčastějších metodách na automatickou detekci a rozpoznávání souhlásek. Naprogramujte vybrané algoritmy na detekci skupin souhlásek, jako jsou frikativy, explozivy a nosovky. Zaměřte se na efektivní implementaci algoritmů. Vytvořte ucelený program na extrakci vybraných českých souhlásek z akustických záznamů plynulé řeči. Program detailně testujte pro více řečníků a různou kvalitu řeči; využijte přitom dostupné databáze. Navrhněte adaptaci programu na konkrétní hlas. Analyzujte spolehlivost a úspěšnost vašeho řešení.

DOPORUČENÁ LITERATURA:

[1] PSUTKA, J., MÜLLER, Z., MATOUŠEK, J., RADOVÁ, V. Mluvíme s počítačem česky. Praha: Academia, 2006.

[2] SIGMUND, M. Rozpoznávání řečových signálů. Skriptum FEKT VUT v Brně. Brno: MJ servis, 2007.

Termín zadání: 10.2.2014 ***Termín odevzdání:*** 30.5.2014

Vedoucí práce: prof. Ing. Milan Sigmund, CSc.
Konzultanti bakalářské práce:

doc. Ing. Tomáš Kratochvíl, Ph.D.
předseda oborové rady

ABSTRAKT

V této práci jsou nejprve prozkoumány vlastnosti českých souhlásek a problematiku rozpoznávání řeči. Následně je vytvořena studie o metodách používaných při automatické detekci souhlásek. Souhlásky se hlavně podílí na srozumitelnosti řeči. Tato práce se zabývá především segmentálním popisem řeči. Prozodické vlastnosti nebudou zatím uvažovány. Cílem této práce je vytvořit program na klasifikaci českých souhlásek jako jsou frikativy, explozivy a nosovky a také na detekci vybraných souhlásek.

KLÍČOVÁ SLOVA

Rozpoznávání řeči, souhlásky, detekce souhlásek, frikativy, explozivy, nosovky, SVM

ABSTRACT

In this work is at first to explore the characteristics and problems of the Czech consonants and the speech recognition. Then is create a study of the methods used in automatic consonants. The consonants are mainly involved in speech intelligibility. This work mainly deals with segmental description of the speech. Prosodic features are not yet considered. The aim of this work is create an program to classification of the Czech consonants like fricatives, plozives and nasals and also for detection of chosen consonants.

KEYWORDS

Recognition of speech, consonants, detection of consonants, fricatives, plosives, nasals, SVM

JURÁK, P. *Signálová analýza mluvených souhlásek*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií. Ústav radioelektroniky, 2014. 30 s., 1 příl. Bakalářská práce. Vedoucí práce: prof. ing. Milan Sigmund, CSc.

PROHLÁŠENÍ

Prohlašuji, že svou semestrální práci na téma „Signálová analýza mluvených souhlásek“ jsem vypracoval samostatně pod vedením vedoucího semestrální práce s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny uvedeny v seznamu literatury na konci práce.

Jako autor uvedené semestrální práce dále prohlašuji, že v souvislosti s vytvořením této semestrální práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne ....

PODĚKOVÁNÍ

Děkuji vedoucímu bakalářské práce prof. ing. Milan Sigmund, CSc. za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce.

V Brně dne ....

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

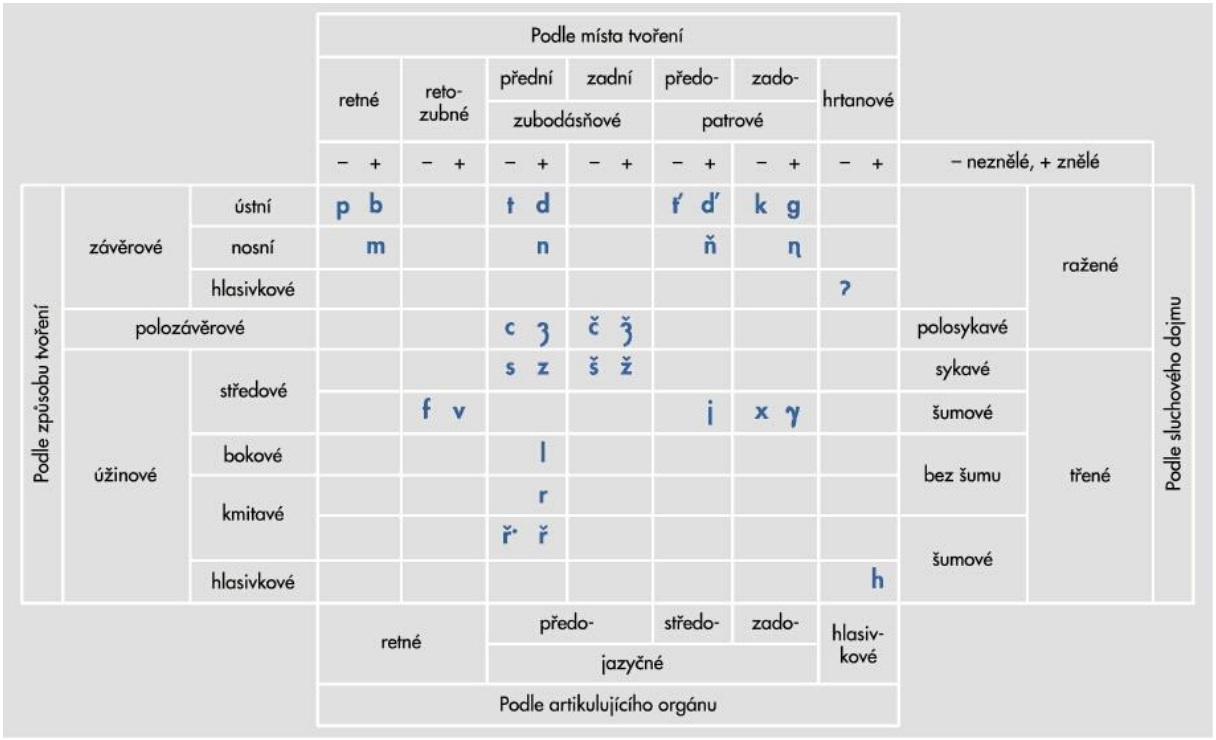
Pojem jazyk se označuje jako schopnost vyjádřit myšlenky pomocí skupiny určitých symbolů – nejčastěji grafických (latinka, azbuka, atd.…) nebo akustických (řeč). Jazyk je velmi složitý systém komunikace, který je vlastní pouze lidským bytostem. Díky němu je možné, mluvenou a psanou formou, sdělovat myšlenky, komunikovat mezi lidmi.

Mluvená řeč se přenáší komunikačním kanálem (prostředím), pomocí akustických vln, tzn. akustického signálu. Takovýto signál se přenáší elastickým prostředím v oblasti slyšitelných frekvencí. Velmi důležitou informací vedle akustické složky je informace lingvistická, daná například fonetickou, morfologickou, syntaktickou, sémantickou nebo pragmatickou strukturou. Akustický signál také obsahuje specifické informace o mluvčím dané vlastnostmi hlasového traktu řečníka, způsobem artikulace nebo případnými vady řeči.

I přes celou řadu nedořešených problémů, s nimiž se systémy hlasové komunikace člověka s počítačem potýkají, dochází stále častěji k jejich praktickému nasazování v průmyslové a společenské praxi.

1 SOUHLÁSKY V ČEŠTINĚ

V českém jazyce se vyskytuje 27 souhláskových fonémů. Na rozdíl od samohlásek se souhlásky projevují charakteristickým šumem a menší amplitudou a jsou stěžejní pro srozumitelnost řeči. V důsledku šumu a menší amplitudy je tak těžší rozlišit souhlásky mezi sebou než od samohlásek.



Obr. 1: Klasifikace českých souhlásek podle místa a způsobu artikulace, znělosti a sonornosti (převzato z [3]).

1.1 Artikulační vlastnosti souhlásek

Při artikulaci souhlásek dochází ke ztížení průchodu vzduchu hlasovým ústrojím. Toto ztížení může být vytvořeno buď vytvořením překážky, nebo zúžením hlasového traktu v daném místě hlasového ústrojí. Vznikají tak různé zvuky s charakteristickým šumem. Při artikulaci českých souhlásek se využívají čtyři charakteristiky: způsob artikulace, místo artikulace, postavení měkkého patra a činnost hlasivek.

1.1.1 Způsob artikulace

Podle způsobu artikulace se souhlásky dělí na závěrové (úplná překážka), úžinové (částečná překážka) a polozávěrové.

Závěrové souhlásky (okluzivy), také nazývané jako souhlásky výbuchové (explozivy) se vytváří tak, že artikulační orgány uzavřou cestu výdechovému proudu vzduchu. Tento závěr se projeví jako krátká pauza (okluze). Následně se uvolní překážka s nahromaděným vzduchem (exploze) a současně vzniká charakteristický šum. Délka trvání exploze trvá přibližně 10 ms. Patří sem /p/, /t/, /ť/, /k/, /b/, /d/, /ď/, /g/, /m/, /n/ a /ň/.

Úžinové souhlásky (konstriktivy). Tyto souhlásky se nazývají také jako třené souhlásky (frikativy) a vznikají tak, že se artikulační orgány těsně přiblíží, ale úplně nesevřou. Při průchodu vzduchu touto úžinou vzniká třecí šum, který je podstatou zvuku. Dalšími variantami úžinové artikulace jsou např. frikativy kmitavé, bokové, atd. Patří sem /f/, /v/, /s/, /š/, /z/, /ž/, /j/, /ch/, /h/, /l/, /r/ a /ř/.

Polozávěrové souhlásky (semiokluzivy) spojují oba typy překážek a to tak, že se v daném místě artikulačního ústrojí vytvoří krátký závěr, který se nezruší, ale přejde do úžiny a ta se postupně otevře. Vzniklé souhlásky se nazývají také jako souhlásky polotřené (afrikáty). Patří sem /c/, /č/, /dz/ a /dž/.

Frikativy je možné udržet určitou dobu v daném stavu a produkovat tak stále stejný zvuk. Naopak pro explozivy jsou typické krátkodobé přechodové zvuky.

1.1.2 Místo artikulace

Dané místo je charakterizováno zúžením nebo úplným uzavřením hlasového traktu. U českých souhlásek se využívá sedmi takových míst hlasového traktu. Souhlásky se tak dále dělí na Retné souhlásky (labiály) se dále dělí na souhlásky obouretné, kde dochází k uzavření obou rtů (/p/, /b/, /m/) a souhlásky retozubné, kdy artikuluje dolní ret proti horním řezákům (/f/, /v/).

Dásňové souhlásky (alveoláry) lze také rozdělit na předodásňové souhlásky, kde špička jazyka artikuluje svou horní plochou proti přední části alveoláry (/t/, /d/, /n/, /s/, /z/, /c/, /dz/, /l/, /r/ a /ř/) a zadodásňové souhlásky, kdy hřeben jazyka artikuluje proti zadní části alveoláry (/č/, /dž/, /š/, /ž/).

Patrové souhlásky se dělí na tvrdopatrové souhlásky, kde artikuluje střední část hřbetu jazyka proti tvrdému patru (/ť/, /ď/, /ň/, /j/) a měkkopatrové souhlásky, kdy artikuluje zadní část hřbetu jazyka proti měkkému patru (/k/, /g/, /ch/).

Hrtanové souhlásky (laryngály) vznikají artikulací hlasivek (/h/).

1.1.3 Postavení měkkého patra

Některé souhlásky využívají pro artikulaci dutiny nosní, zvuk je pak obohacen o silnou tónovou složku nosní rezonance. Dochází k tomu, když měkké patro uvolní průchod vzduchu do dutiny nosní. Vznikají tak nosové explozivy (nazály), kam patří /m/, /n/ a /ň/.

1.1.4 Činnost hlasivek

Pokud je foném tvořen základním hlasivkovým tónem (znělé souhlásky), vznikají pak další tónové složky zvuku hlásky, způsobené rezonancí tónu v nadhrtanových dutinách. Při volném průchodu vzduchu hlasivkami (bez jejich účasti) vzniká zvuk pouze překážkami v nadhrtanových dutinách (neznělé souhlásky). Souhlásky tak lze rozdělit do párů, které se liší pouze přítomností základního tónu. Souhlásky nepárové jsou vždy znělé.

Párové	neznělé	/p/	/t/	/ť/	/k/	/f/	/s/	/š/	/ch/	/c/	/č/	/ř/
	znělé	/b/	/d/	/ď/	/g/	/v/	/z/	/ž/	/h/	/dz/	/dž/	/ř/
Nepárové	znělé	/m/, /n/, /ň/, /l/, /j/, /r/										

Tab. 1: Rozdělení českých souhlásek podle znělosti [5].

1.2 Akustické vlastnosti souhlásek

Akustické vlastnosti souhlásek jsou určeny především podílem tónové a šumové složky.
Sonorní souhlásky (sonory). Zde převládá tónová složka, proto jsou vždy znělé. Mezi sonory patří nazály /m/, /n/, /ň/, dále pak /l/, /r/ a /j/.

Šumové (pravé) souhlásky. Jak už název napovídá, převládá v jejich spektru šumová složka, kromě tónové složky a formantových frekvencí. Řadí se sem všechny párové souhlásky.

Dále lze samohlásky rozdělit podle podobných akustických vlastností.
Likvidy /l/, /r/ a approximanta /j/ jsou velmi podobné samohláskám. Jejich časové průběhy jsou také periodické, ale na rozdíl od samohlásek mají menší amplitudu. Jejich spektra jsou si také velmi podobná, u samohlásek bývají o několik decibelů slabší. Šumová složka ve spektru je prakticky nezřetelná.

Nazály, jsou opět svými akustickými vlastnostmi podobné samohláskám. Amplituda signálů nazál je ovšem menší. Na rozdíl od všech hlásek se zde objevují kromě formantů, také antiformanty (spektrální nuly). Ve spektru lze také pozorovat skokové změny amplitudy a frekvence způsoben okluzí. Okluze se ovšem neprojeví pauzou v důsledku postupného průchodu vzduchu nosní dutinou.

Šumové frikativy a /ř/. Akustické signály šumových frikativ jsou spíše neperiodické mají menší amplitudu než sonory a jejich energie je soustředěna do vyšších frekvencí. V spektrech neznělých frikativ se nenacházejí formanty a jejich tvar je podobný horní propusti. Pro znělé frikativy je naopak typická slabá formantová struktura. V případě /ř/ dochází k modulaci šumové složky periodickou frekvencí od 40-60Hz. U všech znělých šumových frikativ a /ř/ předbíhá harmonická složka šumovou o 30-100 ms.

Explozivy. Jde o přechodové zvuky, kde se okluzní část (závěr) projevuje jako ticho (pauza) s délkou trvání kolem 80 ms. Uvolnění překážky způsobí krátký výbuch šumu (ve spektrogramu podobný vertikálnímu pruhu), který vybudí všechny frekvence. Výbuch přejde ve frikativní šum a následně v artikulovaný foném. Pro explozivy „neuvolněné“ platí, že okluze přechází postupně v akustický signál fonému.

Afrikáty. K jejich vytvoření se využívá všech typů artikulace – okluze, exploze a frikce. Délka trvání afrikáty ovšem není tak dlouhá jako spojení explozivy a frikativy a ani exploze není tak výrazná.

Pro souhlásky je také důležitá ta vlastnost, že jejích dálka není tak proměnlivá jako u samohlásek.

1.3 Fonetická transkripce souhlásek

Při souvisle mluvené řeči dochází k ovlivňování realizací fonémů a to tím způsobem, že může docházet ke změně realizací jednotlivých fonémů, vzniku nového fonému nebo vypuštění daného fonému.

Spojením souhlásky a samohlásky dochází ke změně výslovnosti v případě fonémů [d], [t], [n] ke změkčování hlásek ve spojení se samohláskou [i] a českým písmenem *ě*. Ve spojení *di, ti, ni* se změní předodásňové závěrové hlásky [d], [t], [n] na tvrdopatrové [ď], [ť], [ň]. Ke stejné změně dochází pro spojení *dě, tě, ně, mě.*

1.3.1 Asimilace (spodoba) znělosti

Projevuje se u skupin souhlásek v češtině poměrně často. A to jak uvnitř a na hranicích slov, tak i na hranici předložky a slova nebo předpony a kmene. Ke spodobě dochází pouze u tzv. pravých souhlásek a znělého a neznělého fonému /ř/. Dochází k tomu, že se změní znělost souhlásek v souhláskové skupině, a tím dojde k vyrovnání znělosti skupiny. K vyrovnávání znělosti dochází dvěma způsoby. První je asimilace regresivní (zpětná), kdy je výsledná znělost dána poslední souhláskou, např. sbor [zbor], nůžky [núšky]. Druhá je asimilace progresivní (postupná), kdy první souhláska určuje znělost skupiny – v češtině spíše ojedinělá.

U spodoby znělosti existují výjimky. Patří sem skupina *sh*, která se může vyslovovat buď zněle nebo nezněle, např. shoda [schoda] nebo [zhoda]. Dále pak foném /v/ podléhá asimilaci, aniž by způsobil asimilaci předchozí souhlásky, ovšem vyvolává spodobu znělosti poslední znělé párové souhlásky na konci předchozího slova, např. vtip [fťip], hned vedle [hnet vedle]. Neznělé [ch] můžeme vyslovit jako znělé [h] nebo jako „znělé ch“ (alofon fonému [ch]), např. hroch dovede [hroh dovede].

1.3.2 Asimilace (spodoba) artikulace

Opět má vliv na celou skupinu souhlásek, přičemž dochází ke spodobě artikulace. Ta se projevuje vyrovnáváním artikulačních rozdílů skupiny souhlásek. Tato spodoba se ale vyskytuje většinou fakultativně. Předodásňová hlásky [n] se uvnitř slova před [k] a [g] projevuje jako měkkopatrová [ŋ], např. banka [baŋka]. Podobně pak obouretná hláska [m] uvnitř slova před [v] a [f] se vyslovuje jako [ɱ]. Předodásňové hlásky [t], [d] se mohou uvnitř slova před [ň] vyslovit jako [ť], [ď], např. špatně [špaťňe]. Obdobně se může změnit [n] před [ť], [ď] na [ň], např. anděl [aňďel].

Spodoba artikulace se dále uplatňuje na změnu způsobu tvoření souhlásek. Spojením závěrových hlásek se může [t], [d] s úžinovou hláskou [s] může vzniknout hláska [c], např. dětský [ďecký]. Podobně pak spojením závěrových souhlásek [t] a [d] s úžinovou hláskou [š]

může dojít k vytvoření polozávěrové souhlásky [č], např. většina [vječina] i [vjetšina]. Tato situace nastává v jistých případech i pro spojení závěrových hlásek [t] a [d] se [z] a [ž], kdy mohou vzniknout polozávěrové hlásky [tz], [dz] a [tž] a [dž].

V češtině může dále docházet u některých souhláskových skupin ke zjednodušování výslovnosti, např. francouzští [francousští] i [francouští]. Také jsou-li vedle sebe dvě foneticky (zvukově) stejné souhlásky, může dojít ke zjednodušení, např. vyšší [viší] i [vyšší] nebo panna [pana].

2 ANALÝZA ŘEČOVÉHO SIGNÁLU

Zpracování řečového signálu je velmi důležitou součástí systémů pro rozpoznávání řečových signálů, počítačovou syntézu řeči, identifikaci a verifikaci člověka na základě jeho hlasu nebo pro přenos řečového signálu.

Zvuk (řečový signál) je ve své podstatě mechanické vlnění hmotných částic šířící se prostředím (plynné, kapalné nebo tuhé). Za zvuk se ovšem z fyziologického hlediska považuje pouze oblast slyšitelného vlnění, jehož frekvence je schopno vnímat sluchové ústrojí. Tyto hranice jsou pro jednotlivé osoby individuální a mění se s věkem. Často se uvádí, že zdraví člověk ve věku od 20-25 let dokáže vnímat frekvence od 16 Hz - 20 kHz. Nižší frekvence než 16 Hz se nazývají infrazvuk, vyšší frekvence než 20 kHz se nazývají ultrazvuk.

Většina metod analýzy akustického signálu řeči využívá toho, že vlastnosti řeči se v průběhu času pomalu mění. Pro zpracování se tak využívá metoda krátkodobé analýzy, kdy se zpracovávají oddělené krátké zvuky. Tyto tzv. segmenty mají délku úseku nejčastěji 10 ms, což vyplívá z vlastností artikulačních orgánů. Daný segment pak popisuje číslo nebo soubor čísel. Díky návaznosti segmentů pak dostáváme časové posloupnosti čísel, které popisují daný řečový úsek. Tyto metody spíše zpracovávají signál získaný základní digitalizací, tzv. pulzní kódovou modulací (PCM).

2.1 Zpracování v časové oblasti

Pro krátkodobou analýzu řečového signálu se nejčastěji využívá vztahu [5]

$$Q_n = \sum_{k=-\infty}^{\infty} \tau(s(k))w(n-k) \tag{2.1}$$

kde Q_n je krátkodobá charakteristika, $s(k)$ je vzorek akustického signálu získaný PCM v čase k, $\tau(.)$ zastupuje danou transformační funkci a $w(n)$ je váhová posloupnost tzv. okénko kterým se vybírají vzorky $s(k)$. Tento vztah je diskrétní konvolucí posloupnosti vzorků $\tau(s(k))$ a impulsní funkce $h(n)=w(n)$. Při zpracování v časové oblasti se nejčastěji využívá pravoúhlé nebo Hammingovo okénko. Pravoúhlé okénko je definováno jako [5]

$$w(n) = \begin{cases} 1 & pro\ 0 \le n \le L-1 \\ 0 & pro \quad ostatní\ n, \end{cases} \tag{2.2}$$

kde L je počet vzorků vybraných okénkem. Pro potlačení vzorků na krajích okénka se používá Hammingovo okénko [5]

$$w(n) = \begin{cases} 0{,}54 - 0{,}46\cos(2\pi n/(L-1)) & pro\ 0 \le n \le L-1 \\ 0 & pro \quad ostatní\ n. \end{cases} \tag{2.3}$$

U obou typů okének jde o filtr typu dolní propust. Pro analýzu řečového signálu má n nejčastěji tvar $n=Ni$-1, kde i reprezentuje pořadové číslo segmentu a N počet vzorků segmentu. Při působení okénka lze uvažovat, že všechny vzorky vně okénka jsou nulové, proto lze přepsat předchozí vztah (2.1) do tvaru [5]

$$Q_{N-1} = \sum_{k=0}^{N-1} \tau(s(k)) w(N-1-k), \tag{2.4}$$

kde *τ(.)* zastupuje danou transformační funkci, *s(k)* je vzorek akustického signálu získaný PCM v čase *k*, *w(n)* je váhová posloupnost, N počet vzorků segmentu.Pro zachycení periodických vlastností znělých úseků řeči je ovšem třeba, aby obdélníkové okénko obsahovalo alespoň jednu periodu a Hammingovo okénko alespoň dvě periody základního hlasivkového tónu. Tím se ovšem prodlouží délka okénka. Pokud tedy nebude uvnitř okénka signál stacionární, získá se spíše signál průměrný než ,,aktuální''. Proto se může zdát, že lepší je pravoúhlé okénko, které poskytuje lepší časové rozlišení. Ovšem toto okénko má na vyšších frekvencích horší útlum, proto se častěji volí Hammingovo okénko.

2.1.1 Krátkodobá energie

Funkce krátkodobé energie je popsána vztahem [5]

$$E = \sum_{k=-\infty}^{\infty} \left[s(k) w(n-k) \right]^2, \tag{2.5}$$

kde *s(k)* je vzorek signálu v čase *k* a *w(n)* je příslušným typem okénka. Pro každý mikrosegment tak lze získat průměrnou hodnotu energie. Nevýhodou této charakteristiky je velká citlivost na rychlé změny úrovně signálu. Hodnoty této charakteristiky lze využít pro oddělení segmentů ticha od segmentů řeči.

2.1.2 Krátkodobá funkce středního počtu průchodů nulou

Charakteristika počtu průchodů signálu nulovou úrovní popisuje spektrální vlastnosti signálu, ovšem pouze zjednodušenou formou. Krátkodobá funkce středního počtu průchodů nulou je popsána jako [5]

$$ZCR = \sum_{k=-8}^{\infty} \left| \operatorname{sgn}\left[s(k) \right] - \operatorname{sgn}\left[s(k-1) \right] \right| w(n-k), \tag{2.6}$$

kde *w(n)* je pravoúhlé okénko, *s(k)* je vzorek signálu v čase *k*. Tato funkce se využívá pro zjištění začátku nebo konce promluvy, pro určení základního hlasivkového tónu nebo pro přibližné určení frekvence nejsilnějšího formantu.

2.1.3 Krátkodobá autokorelační funkce

Krátkodobá autokorelační funkce je popsána vztahem [5]

$$R_n = \sum_{k=-\infty}^{\infty} s(k) w(n-k) s(k+m) w(n-k-m), \tag{2.7}$$

kde *w(n)* je váhovací okénko, *s(k)* je vzorek signálu v čase *k*. Využívá se jí pro zjištění periodičnosti signálu. Pro periodický signál s periodou P nabývá autokorelační funkce

maximálních hodnot pro *m=0, P, 2P,* … .Funkce dobře určuje periodu základního hlasivkového tónu. Použité okénko však musí být dostatečně dlouhé, nejméně dvě periody signálu.

2.2 Zpracování ve frekvenční oblasti

Ve frekvenční oblasti může být řeč reprezentována spektrální obálkou, respektující vlastnosti hlasového ústrojí. Opět se zde pracuje se stacionárními signály, které charakterizuje tzv. krátkodobá spektrální analýza.

2.2.1 Krátkodobá Fourierova transformace

Krátkodobá Fourierova transformace je nejčastěji využívaná transformace. Transformace je popsána vztahem [5]

$$S(\omega,n)=\sum_{k=-\infty}^{\infty}s(k)h(n-k)e^{-j\omega k}\ , \tag{2.8}$$

kde *h(n)* je funkce blíže nespecifikovaného okénka, *s(k)* jsou vzorky řečového signálu v čase *k*. Fourierův obraz odpovídá konvoluci okénka *h(n)* a vzorku *s(n)* modulovaného $e^{-j\omega n}$, *n* je diskrétně proměnný čas. Po úpravě lze vztah zapsat jako [5]

$$S(\omega,n)=e^{-j\omega n}\left\{s(n)*\left[h(n)e^{j\omega n}\right]\right\}. \tag{2.9}$$

Při zafixování času *n*, kdy se signál řeči zpracovává krátkodobou disktrétní Fourierovou transformací, jsou čas i frekvence disktrétní. Takto získané koeficienty se používají ve spektrálních analyzátorech řeči nebo v systémech rozpoznávání řeči. Při zafixování frekvence ω se Fourierova transformace využívá pro lineární filtraci. Lze tak navrhnou vhodnou volbou okénka pásmový filtr nebo soustavu (banku) pásmových filtrů.

Při krátkodobé diskrétní Fourierově transformaci je důležitý výběr frekvence diskretizace *S(w, n)* v čase i ve frekvenci. Musí být dodržena podmínka vzorkovacího teorému, kdy jsou vzorky odebírány alespoň s frekvencí $2B_w$, kde B_w je šířka pásma filtru.

2.2.2 Pásmová filtrace

Vhodnými pásmovými filtry lze detekovat fonetickou strukturu řeči a výsledné informace využít pro klasifikaci jednotlivých hlásek či slov. Pro odvození vzorce pro banku pásmových filtrů lze vyjít ze vzorce (2.10). Po úpravě lze dostat vzorec [5]

$$y_q(n)=\sum_{k=-\infty}^{\infty}s(n-k)h_q^{\ *}(k)\ , \tag{2.10}$$

kde [5]

$$y_q(n)=S_q(q)e^{j\omega_q n}. \tag{2.11}$$

Pokud má okénko $h_q(n)$ typu dolní propust, pak vzorec (2.10) představuje pásmovou filtraci na střední frekvenci ω_q a $h_q^*(n)$ je impulzní odezvu daného pásmového filtru. Šířka pásma je pak $2\omega_{pq}$, kde ω_{pq} je mezní frekvence ideální dolní propusti.

Pro návrh banky pásmových propustí je třeba znát celou šířku pásma, kterou bude banka filtrovat, počet těchto pásem Q, šířku ω_{pq} každého pásma a tvar každého filtru. Pro návrh banky pásmových propustí se nejvíce hodí číslicové filtry s konečnou impulzní charakteristikou (FIR), protože se dají snadno navrhnout. Umístění jednotlivých pásem udává střední frekvence ω_q a šířka propustnosti daného pásma je reprezentována frekvenční charakteristikou dolní propusti $H_q(e^{j\omega})$.

3 METODY NA AUTOMATICKOU DETEKCI A ROZPOZNÁVÁNÍ SOUHLÁSEK

Ačkoliv metody pro rozpoznávání řeči prošli značným vývojem, klasifikace řečových signálů se stále potýká s určitými obtížemi.

Z důvodu odlišného hlasového ústrojí a odlišného způsobu artikulace je hlas každé osoby individuální (barva hlasu, přízvuk, tempo řeči). Proto se systémy na rozpoznávání řeči dělí na systémy závislé na řečníkovi a systémy na řečníkovy nezávislé.

Hlas daného člověka také může být v různých situacích odlišný. V podstatě je tedy nemožné, aby člověk řekl jedno slovo ve dvou situacích zcela stejně. Navíc jev koartikulace může pozměnit fonetické vlastnosti začátku a konce slova.

Obtíže také souvisí se změnou akustického prostředí a to okolní šum a rušení přenosového kanálu. Vysoká úroveň šumu například způsobuje velké potíže při detekci frikativ a identifikaci začátku a konce promluvy.

Nejobtížnější úlohou je rozpoznávání souvislé řeči, kdy řečník vybírá ze zásoby až několika desítek tisíc slov. Na schopnost správného rozpoznání má také vliv to, zda řeč je čtená nebo zda jde o spontánní promluvu. Ve spontánní promluvě se totiž často objevuje mnoho tzv. neřečových událostí (hlasité váhání, nádechy). V češtině se navíc ve spontánní promluvě objevuje mnoho hovorových slov nebo nespisovných gramatických vazeb.

Metody rozpoznávání lze dělit na ty, které pracují na principu porovnávání se vzory a na ty které využívají statistických metod.

3.1 Metoda porovnávání se vzory

Tento způsob se využíval hlavně v šedesátých a sedmdesátých letech. Slovo se zpracovává a klasifikuje jako celek a poté se přiřazuje do té třídy vzorového obrazu, ke kterému má nejmenší vzdálenost. Nejdůležitější je však určení vzdálenosti mezi dvěma obrazy slov. Vzdálenost je nejčastěji určována na základě aplikace metody dynamického programování, kde se hledá takové nelineární transformace časové osy jednoho z obrazů, při které dojde k porovnávání obou obrazů s nejmenší výslednou vzdáleností. Přičemž kolísání v časové ose je modelováno časově nelineární „bortivou“ funkcí s přesně specifickými vlastnostmi. Časové rozdíly mezi dvěma řečovými obrazy jsou přitom eliminovány „borcením“ jedné z časových os takovým způsobem, že je dosaženo maximální shody s druhým obrazem.

3.2 Statistická metoda

Tento způsob využívá toho, že slova a celé promluvy jsou modelovány pomocí tzv. skrytých Markovových modelů. Přičemž slova mohou být modelována Markovovým modelem jako celek nebo jsou vytvářeny skryté Markovovy modely subslovních jednotek a promluva je modelována zřetězením těchto subslovních modelů. Pro každou subslovní jednotku jsou pak v procesu trénování stanoveny na základě trénovací množiny promluv parametry

odpovídajícího Markovova modelu a neznámá promluva je rozpoznána na základě toho jaká posloupnost slov tvořená řetězcem odpovídajících modelů subslovních jednotek generuje promluvu s největší aposteriorní pravděpodobností.

Mezi metody určené pro detekci souhlásek patří následující.

3.3 Detekce explozivních souhlásek pomocí vlnkové transformace

Vlnková transformace je transformace časové stupnice s konstantní relativní šířkou pásma. Vlnkové koeficienty se vypočítají jako [4]

$$O(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} h^*(\frac{t-b}{a}) p(t) dt \ , \tag{3.1}$$

kde b je časový parametr, $h(t)$ základní vlnka, $h^*(t)$ komplexně sdružená hodnota k $h(t)$.

3.3.1 Použitá metoda

Tato metoda byla použita pro neznělé explozivy /p/, /k/, /p/ a pro znělé explozivy /b/, /g/, /d/. Rozklad frekvencí je přizpůsoben vlastnostem vlnkové analýzy a explozivním souhláskám. Frekvenční pásmo je rozděleno na čtyři oktávy ([372 – 5000] Hz) se čtyřmi bankovými filtry na oktávu. Je použita vzorkovací frekvence 10 kHz. Relativní šířka pásma vlnkové analýzy je mezi 1/3 a 1/4 oktávy. Pro dentály je energie koncentrovaná v oblasti 500 Hz a nad 4 kHz, pro labiály v oblasti ([500 – 1500] Hz) a pro veláry v oblasti ([1500 - 4000] Hz).

V řečovém signálu se může objevit několik typů nespojitostí. Proto je důležité před určením impulzu ověřit, zda část analyzovaného řečového segmentu souhlasí s energetickými a frekvenčními charakteristikami exploziv, které jsou pro oba typy exploziv rozdílné. Neznělé explozivy mají nižší energii na nižších frekvencích. Znělé explozivy mají vysoký index znělosti a energie se nekoncentruje na vyšších frekvencích.

Princip detekce je rozdělen do dvou částí. Funkce S_m lokalizuje lokální minimum, jehož hodnota je menší než nalezená prahová hodnota S. Poté se časová lokalizace impulzu provádí pomocí následujícího maxima funkce S_M. Pro detekci neznělých souhlásek platí, pokud časová vzdálenost dvou po sobě jdoucích detekcí (d_k a d_l) je větší než maximální časová vzdálenost dvou po sobě jdoucích exploziv, je brána v úvahu pouze detekce d_k (d_l odpovídá segmentu samohlásky následující po souhlásce). Pro znělé explozivy je brána v úvahu pouze detekce d_l (d_k může odpovídat znělým segmentům během okluze). Maximální časová vzdálenost mezi dvěma po sobě jdoucími explozivami byla změřena na nejméně 140 ms pro neznělé explozivy a 130 ms pro znělé explozivy.

3.3.2 Výsledky použité metody

Korpus C_a, který byl použit pro analýzu parametrů a pro návrh algoritmů se skládá ze sady 35 vět pronesených 11 řečníky (6 mužů a 5 žen) a souboru 263 vzorků CVC (souhláska – samohláska - souhláska) ze souboru BDSONS pronesených 9 řečníky (6 mužů a 3 ženy). Pro

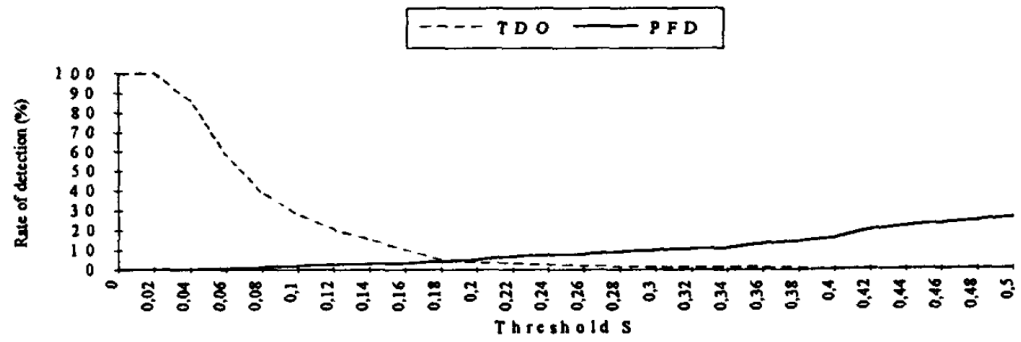
testování algoritmu byl použit korpus C_t obsahující 176 vět (ze sady SYL, BDSONS) pronesených mužskými řečníky.

Výsledky testovaných korpusů jsou na obr. (2) a obr. (3). Bylo definováno pět proměnných – nto (počet exploziv ve sledovaném korpusu), nod (počet detekovaných exploziv), nte (celkový počet detekcí), PFD (3.2) a TDO (3.3). Kde PFD má tvar [4]

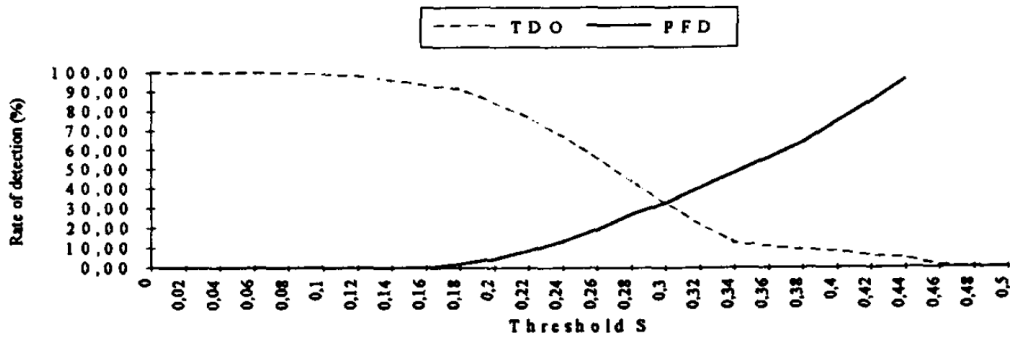
$$PFD = 100\frac{nte - nod}{nto}, \tag{3.2}$$

a TDO má tvar [4]

$$TDO = 100\frac{nto - nod}{nto}. \tag{3.3}$$



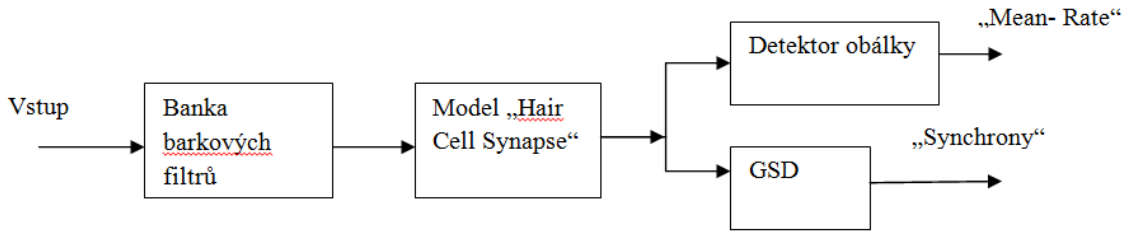
Obr. 2: Charakteristika detektoru pro neznělé explozivy [4].



Obr. 3: Charakteristika detektoru pro znělé explozivy [4].

3.4 Systém pro rozpoznávání frikativ

Tento systém využívá 36 barkových filtrů s preemfází 20dB/dek na vysokých frekvencích. Tento sytém byl vyvinut Saneffem a je popsán blíže v [3]. Tento systém má dva výstupy nazvané jako „mean-rate“ a „Generalized Synchrony Detector (GSD)“. Blokový diagram je popsán na obr. (4).



Obr. 4: Blokové schéma zpracování vstupního signálu [1].

Tento systém byl navržen na základě 220 frikativ extrahovaných z plynulé řeči z databáze TIMIT. Kde promluvy byly prováděny 6 řečníky (3 muži a 3 ženy). Systém byl testován na 500 frikativách extrahovaných z plynulé řeči z databáze TIMIT pro 22 řečníků s různými americkými dialekty.

3.4.1 Použité akustické vlastnosti

Doba trvání neznělých částí (DUP) frikativ se používá při detekci jako vlastnost znělosti. Využívá se metoda detekce absence nebo přítomnosti znělosti v signálu a tím se odhalí počáteční a koncové body. Znělost se projevuje na výstupu jako vyšší energie na nízkých frekvencích. Pro odhalení této energie se používají dvě metody. První způsob využívá toho, že se vezme celková energie na nejnižších devíti filtrech (frekvence je menší než 1 kHz) z výstupu „GSD“. Dále bude nazýván LOWG. Druhý způsob je založen na poměru mezi nízkofrekvenční (menší než 1,5 kHz) a vysokofrekvenční (větší než 3 kHz) energií na výstupu „mean-rate“. Dále bude nazýván LOWE. Výhodou použití těchto dvou metod je, že se navzájem doplňují. Pokud jeden z nich přesahuje prahovou hodnotu, pak se předpokládá, že foném je přítomen. Pokud jsou obě hodnoty LOWG a LOWE menší než prahová hodnota, pak je DUP periodický. Pokud je DUP nižší než určený práh, pak je frikativ znělý, jinak jde o neznělý frikativ. Práh určený pro DUP je kolem 60 ms. Ovšem pokud je nad 100 ms pak je s určitostí neznělý. Výsledky jsou uvedeny v tab.(2)

	Detekované znělé explozivy	Detekované neznělé explozivy
Znělé explozivy	186	17
Neznělé explozivy	9	288

Tab. 2: Detekce znělosti. Úspěšnost 95% [1].

Relativní amplituda (RA) je v literatuře označována jako vlastnost jak rozlišovat mezi sykavkami, které mají větší RA a nesykavkami, které mají menší RA. RA je definována jako [1]

$$RA = \sum_{i:všechnyfiltry} yenv_i\big|_{frikativa} / \sum_{i:všechnyfiltry} yenv_i\big|_{samohláska}, \qquad (3.4)$$

kde $yenv_i$ je výstup „mean-rate“ z i-tého filtru. Lepší výsledek se získá požitím dvou vlastností. Jednou z nich je relativně nízká amplituda a druhou je monotónní spektrum, které charakterizují nesykavky. Funkce, která by mohla být použita pro rozlišení mezi sykavkami a nesykavkami je nazývaná „Maximum Normalized Spectral Slope (MNSS)“ a má tvar [1]

$$MNSS = \max\{(yenv_i - yenv_{i-1})\big|_{frikativa}\} / \sum_{i:všechnyfiltry} yenv_i\big|_{samohláska}, \qquad (3.5)$$

Práh byl empiricky určen na 0,02 pro neznělé a 0,01 pro znělé frikativy. Pokud je MNSS větší než tento práh, tak se jedná o sykavku, jinak jde o nesykavku. Pokud je MNSS blízko prahové hodnotě použije se normalizace s respektováním energie frikativy místo nejbližší souhlásky. Výsledky jsou uvedeny v tab. (3).

	Detekované sykavky	Detekované nesykavky
/s/ a /z/	89	0
/f/, /v/, /th/ a /dh/	8	83
/sh/ a /zh/	34	6

Tab. 3: Detekce sykavek a nesykavek. Úspěšnost 94% [1].

Velmi důležitý je spektrální tvar pro detekci místa artikulace frikativy. Alveoláry (předodásňovky) jsou charakteristické nižšími spektrálními vrcholy než palatály (předopatrovky). Nejdůležitější vlastností je nalezení nejdominantnějšího vrcholu. Palatály jsou charakteristické monotónním spektrem, které má dominantní vrchol na relativně nízké frekvenci ve srovnání aveoláry, jejichž vrchol je na vyšší frekvenci a nesykavkami, které nemají významný vrchol na vyšší nebo mnohem menší frekvenci. Nejlepší výsledky se získají použitím výstupu GSD. Výsledky jsou uvedeny v tab.(4).

	Detekované palatály	Detekované nepalatály
/s/ a /z/	0	89
/f/, /v/, /th/ a /dh/	0	91
/sh/ a /zh/	37	3

Tab. 4: Detekce palatál. Úspěšnost 98,5% [1].

Jiná vlastnost, která hraje významnou roli při rozlišení alveolár a papatál je Spectral Center of Gravity (SCG). Ta popisuje vlastnosti tvaru spektra, které nejsou popsány MNSS nebo umístěním MDP.

Získané vlastnosti jsou dále sloučeny a podle různých vlastností je rozhodnuto o místu artikulace. K rozlišení se používají tři vlastnosti - „Maximum Normalized Spectral Slope (MNSS)“, Spectral Center of Gravity (SCG), umístění nejdominantnějšího vrcholu (MDP). Výsledky jsou uvedeny v tab.(5).

	Detekované alveoláry	Detekované dentály	Detekované palatály
Alveoláry: /s/ a /z/	85	4	0
Dentály: /f/, /v/, /th/ a /dh/	0	91	0
Palatály:/sh/ a /zh/	2	0	38

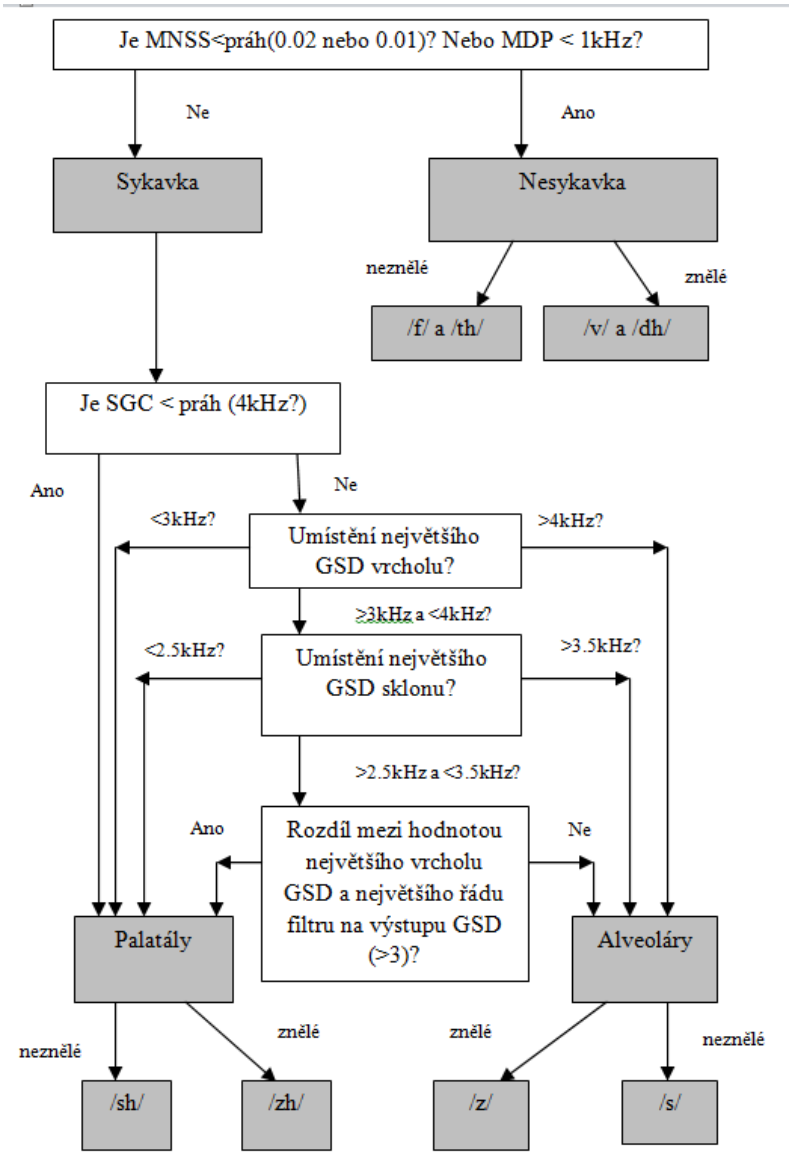
Tab. 5: Detekce místa artikulace pro šest řečníků. Úspěšnost 97% [1].

Výsledky detekce znělosti a místa artikulace jsou uvedeny v tab.(6)

	/s/	/f/, /th/	/sh/	/z/	/v/, /dh/	/zh/
/s/	90	8	2	0	0	0
/f/, /th/	4	87	4	0	5	0
/sh/	4	1	92	0	0	3
/z/	8	1	1	85	2	3
/v/, /dh/	0	0	0	0	100	0
/zh/	0	0	9	7	1	83

Tab. 6: Detekce frikativ pro dvaadvacet řečníků. Úspěšnost 90% [1].

Použitý algoritmus je zobrazen na obr. (5).



Obr. 5: Algoritmus pro detekci místa artikulace [1].

3.5 Detekce frikativ na základě použití algoritmu SVM

Tato metoda využívá třístupňové klasifikace pro zařazení fonémů mezi neznělé frikativy. V první fázi probíhá předzpracování, tzn. že signál je nejprve nasegmentován. Z těchto segmentů jsou poté extrahovány vlastnosti pomocí vytvořených funkcí. V druhé fázi jsou jsou vytvořeny modely pomocí Support Vector Machine (SVM). Třetí fáze je pak fáze klasifikace, kde jsou fonémy nejprve rozděleny do dvou podskupin na sykavky (/ s / a / sh /) a neskyvky (/ f / a / th /). Každá skupina je dále klasifikována pomocí jiného SVM modelu.

Pro vytvoření této metody bylo prostudováno více než 35 parametrů a funkcí z nichž bylo vybráno 15 podle jejich vlastností na základě, kterých se dají jednotlivé frikativy rozlišit.

Patří sem detekce spektrální špičky, spektrální rolloff, spektrální těžiště, poměr energií dvou frekvenčních pásem dále pak počet průchodů nulou, směrodatná odchylka, sešikmenost a špičatost intervalů mezi průchody nulou v časové oblasti. Ve frekvenční oblasti pak melovské frekvenční kepstrální koeficienty, lacunarity β parametr popsán v [14], frekvence nejvyššího vrcholu podle gammatonové banky filtrů, spektrální deformace a spektrální šířka. Každá tato funkce byla následně testována z důvodu jejího rozlišení pro jednotlivé fonémy. Byla využita databáze TIMIT.

3.5.1 Klasifikace

Při klasifikaci je možné postupovat dvěma způsoby. Při prvním způsobu se pomocí trénovací sady vytvoří model, který je dále používán pro klasifikaci. Druhý způsob využívá dvou stupňů. V prvním stupni jsou fonémy rozděleny na sykavky a nesykavky pomocí 12 vlastností a jednoho modelu SVM. V druhé fázi jsou sykavky pomocí dalšího SVM modelu a vektoru 11 vlastností rozděleny na jednotlivé fonémy. U nesykavek se využívá jiný vektor funkcí.

3.5.2 Trénování a testování

Pro trénování a testování je využit stažený balíček libsvm [9]. Pro tvorbu modelu vybrána klasifikace C-SVC a RFB jádro. Data pro trénovací proces se skládají ze 104 fonémů čtyř typů z databáze TIMIT. Každý foném byl rozdělen do 8 ms po sobě jdoucích segmentů s 50% překrytím. Pro každý segment byl vytvořen vektor vlastností. Všechny vektory vlastností byly normovány mezi hodnoty [-1,+1]. Parametry C a gamma byly vybrány za použití vyhledávací mříže a křížové validace.

3.5.3 Výsledky

Testování s využitím všech neznělých frikativ z databáze TIMIT probíhalo tak, že bylo vybráno 150 fonémů *s* a *sh*, 170 fonémů *f* a 270 fonémů *th* pro trénování. Zatímco 11848 fonémů bylo použito pro testování. Výsledky jsou v Tab.(7)

	s	sh	f	th
s	88,68±0,74	0,51±0,08	3,35±0,58	7,46±0,54
sh	0,60±0,09	79,99±3,20	17,69±3,38	1,72±0,36
f	4,64±0,79	14,84±3,64	77,92±3,59	2,60±0,39
th	4,26±0,53	3,93±0,70	0,18±0,06	91,64±0,90

Tab. 7: Úspěšnost detekce neznělých frikativ [8]

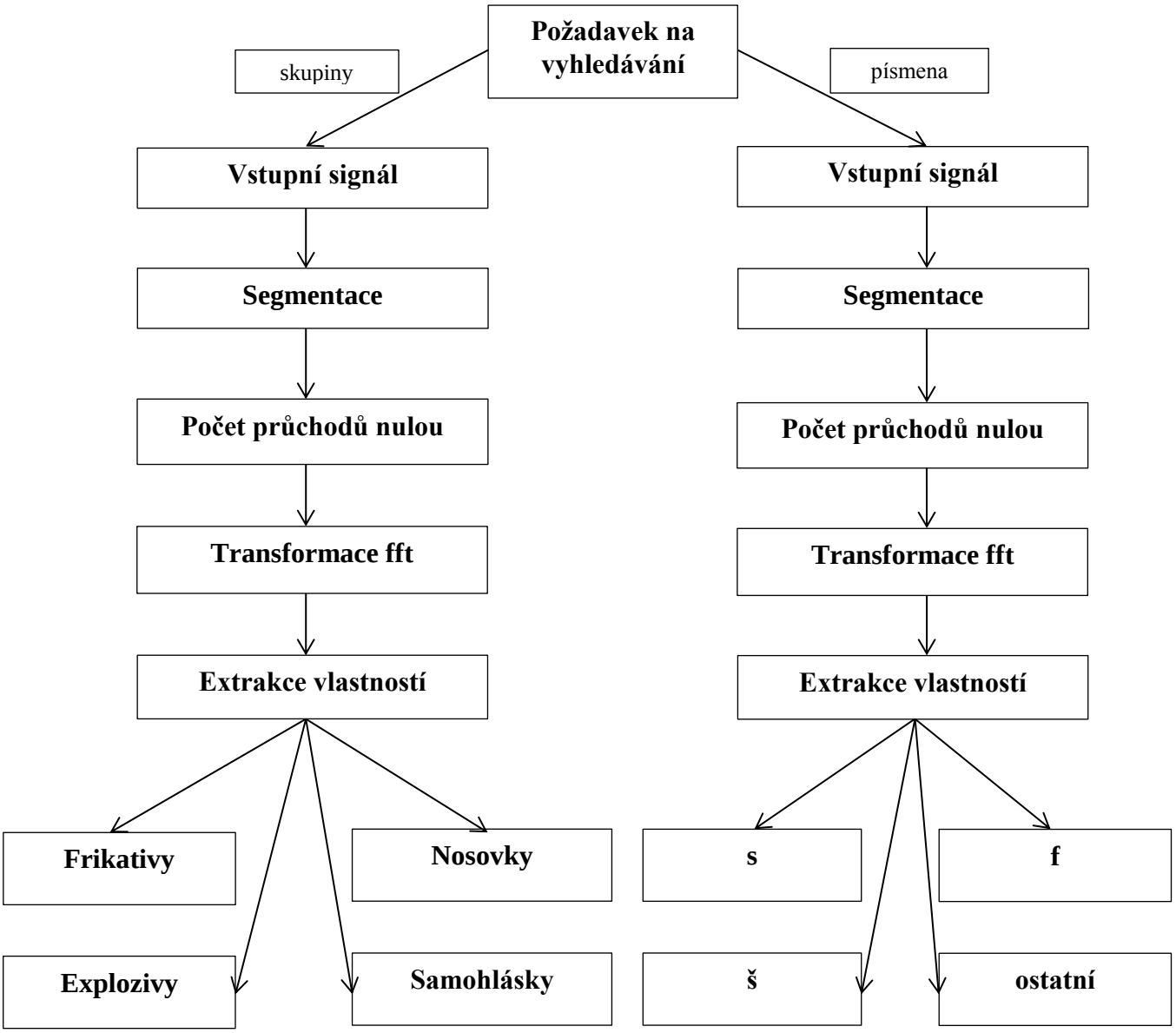
4 DETEKČNÍ ALGORITMUS

Na základě studia o metodách na automatickou detekci a rozpoznávání souhlásek byla pro vytvoření detekčního algoritmu vybrána jednoduchá metoda založená na strojovém učení tzv. Support vector machines (SVM). Pro vývoj algoritmu bylo využito programového prostředí MATLAB verze R2010a a vývojového prostředí Microsoft Visual Studio 2010. Byl vytvořen ucelený program pro klasifikaci skupin souhlásek jako jsou frikativy, explozivy a nosovky a současně pro detekci jednotlivých souhlásek *s*,*f* a *š*. Vývojový diagram je na Obr.(6).

Procesem pulzní kódové modulace byl signál nejprve převeden z analogové do digitální formy. Na začátku byl akustický signál z mikrofonu navzorkován se vzorkovací frekvencí f_{vz} = 22050 Hz a 16-ti bitovým převodem. Signál byl dále načten příkazem *wavread* do Matlabu. Načtený signál byl rozdělen do jednotlivých segmentů pomocí funkce okna. Pro účely této práce bylo dostačující využití pravoúhlého okna. Pro zachycení znělých úseků byla délka okna stanovena na 20 ms s překrytím 5 ms, což odpovídalo počtu vzorků N = 442 pro jeden segment. Z něj pak 112 vzorků tvořilo překrytí. Výsledkem segmentace byla dvourozměrná matice hodnot jednotlivých segmentů. Tyto segmenty byly dále jednotlivě podrobeny rychlé Fourierově transformaci, požitím funkce *fft*. Následně bylo pro každý segment vybráno pouze spektrum v rozsahu 0 Hz až 11025 Hz, které bylo dále zpracováváno, podle způsobu klasifikace. A to buď pomocí funkcí určených pro klasifikaci skupin souhlásek nebo funkcí určených pro detekci jednotlivých souhlásek. Pro každý segment, tak byl vytvořen vektor vlastností, který byl uložen do společné matice vlastností pro všechny segmenty.

Trénovací data byla ze signálů vybírána ručně pomocí programu Audacity. Jde o velice jednoduchou, ale poměrně pracnou metodu získávání dat. Tato metoda je nevhodná pro přesné určení hranic jednotlivých hlásek. Trénovací data obsahují několik skupin, které mají být odděleny a navíc jednu skupinu, která obsahuje co nejvíce ostatních hlásek, které nepatří, ani mezi jednu skupinu. Každé této skupině je přiřazena určitá celočíselná hodnota, pomocí níž jsou následně identifikována testovaná data.

Pro trénování a testování je využit balíček stažený z libsvm [9]. Tento balíček v sobě zahrnuje algoritmus SVM vytvořený v prostředí Visual Studio a také rozhraní pro Matlab, které umožňuje využití tohoto algoritmu v Matlabu. Pro práci s tímto algoritmem byl také využit průvodce [10], který poskytuje návod, jak algoritmus správně používat. Vektor vlastností byl tedy podle tohoto návodu normován mezi hodnoty [-1,+1]. Normovaný vektor byl uložen do matice vlastností, která byla následně převedena do formátu SVM pomocí funkce *libsvmwrite*, která byla byla součástí balíčku libsvm. Podle [8] byl použit C-SVC typ pro klasifikaci a RFB jádro pro tvorbu následného modelu. Poté byl nalezen nejlepší parametr C. K tomuto účelu byl využit algoritmus [15], který postupně porovnával jednotlivé hodnoty C pro trénovací data a automaticky vygeneroval vhodné C. Po nalezení ideálního parametru C byl pomocí funkce *svmtrain* nalezen model pro klasifikaci nebo detekci.



Obr. 6: Vývojový diagram algoritmu

4.1 Klasifikace skupin souhlásek

Pro klasifikaci jednotlivých skupin souhlásek, jako jsou frikativy, explozivy a nosovky bylo využito 13 funkcí a to podle jejich vlastností, na základě kterých bylo možno jednotlivé skupiny rozlišit. Tyto vlastnosti byly následně pro každý segment uloženy do vektoru. Vznikla, tak matice n vektorů reprezentujících n analyzovaných segmentů. Pro získání těchto vlastností byly využity následující funkce.

1. Spektrální Rolloff. Jedná se o frekvenci f_r, která reprezentuje ve spektru hranici, pod kterou je koncentrováno p procent velikostí spektrálních složek. Na základě

experimentů byla použita hodnota p = 50%. Vypočítá se jako [8]

$$\sum_{j=0}^{f_r} M_i(j) = p\sum_{j=0}^{J-1} M_i(j), \tag{4.1}$$

kde i je daný segment, j představuje spektrální složku a $M_i\ (j)$ je velikost spektrální složky. Byl využit algoritmus [11].

2. Poměr energií dvou frekvenčních pásem. Byly využity dva poměry energií E_1 a E_2. Kde E_1 je vhodné pro oddělení samohlásek od souhlásek podle [7]. E_2 je naopak vhodné pro oddělení exploziv od ostatních souhlásek [7]. Pro E_1 platí

$$E_1 = 10\log_{10}\left(\frac{B_1}{B_2}\right), \tag{4.2}$$

kde B_1 = 0 – 800 Hz a B_2 = 7000 – 8000 Hz. Pro E_2 platí

$$E_2 = 10\log_{10}\left(\frac{B_3}{B_4}\right), \tag{4.3}$$

kde B_3 = 1200 – 4000 Hz a B_4 = 5000 – 8000 Hz.

3. Melovské frekvenční kepstrální koeficienty. Jsou počítány jako logaritmus velikostí spektrálních složek, kde jsou jednotlivé spektrální složky seskupeny podle Melovské frekvenční škály. Následně jsou podrobeny diskrétní kosinově transformaci (DCT). Na základě experimentů bylo využito osm koeficientů, pro vektor vlastností. Byl využit algoritmus [11].
4. Počet průchodů nulou. Počet průchodů nulou je počítán pro každý segment v časové oblasti. Tato funkce je vhodná pro oddělení frikativ od ostatních souhlásek. Vypočítá se jako [8]

$$ZCR = \frac{1}{2}\sum_{n=1}^{N-1}\left|\mathrm{sgn}(s[k]) - \mathrm{sgn}(s[k-1])\right|, \tag{4.4}$$

kde $s(k)$ je signál v časové oblasti pro daný segment a k je daný vzorek.

5. Frekvence nejvyššího vrcholu podle gammatonové banky filtrů. Tyto gammatonové spektrogramy jsou podobné běžným spektrogramům. Norozdíl od nich ale nemají konstantní šířku všech pásem. Gammatonové banky filtrů respektují vlastnosti lidského ucha a jejich pásma jsou na vyšších frekvencích širší. Byl využit algoritmus [12].

Trénovací data pro vytvoření SMV modelu sestávala z 60 frikativ, 60 exploziv, 60 nosovek a 60 samohlásek. Tato trénovací data byla získána z audio záznamů rádia stažených z internetu [13] od šesti různých osob. Na základě těchto dat byl vytvořen obecný model pro klasifikaci souhlásek. Algoritmus byl pomocí vytvořeného modelu dále detailně testován pro více osob a různou kvalitu řeči. Testování probíhalo nejprve na audio záznamech rádia tří osob a to jiných osob než, které byly využity pro trénování dat. Promluva těchto řečníků byla plynulá a v pozadí nebyl šum. Tab.(8) ukazuje úspěšnost klasifikace, která je 74,6%. K častým záměnám samohlásek za nosovky došlo pouze u jedné této osoby, jejíž vlastnosti samohlásek byly relativně podobné trénovací sadě nosovek.

	frikativy	explozivy	nosovky	samohlásky
frikativy	25	1	4	1
explozivy	8	22	1	2
nosovky	0	1	17	0
samohlásky	2	2	14	42

Tab. 8: Úspěšnost klasifikace samohlásek pro plynulou promluvu bez šumu

Algoritmus byl dále testován na skupině dvou osob s plynulou promluvou bez šumu, ale s mírným cizím přízvukem (němčina, angličtina). Testovací data byla opět získána z audio záznamů rádia stažených z internetu. Tab.(9) znázorňuje tyto výsledky. Úspěšnost klasifikace byla 71,9%.

	frikativy	explozivy	nosovky	samohlásky
frikativy	13	1	2	3
explozivy	4	8	1	3
nosovky	0	1	7	0
samohlásky	1	1	4	21

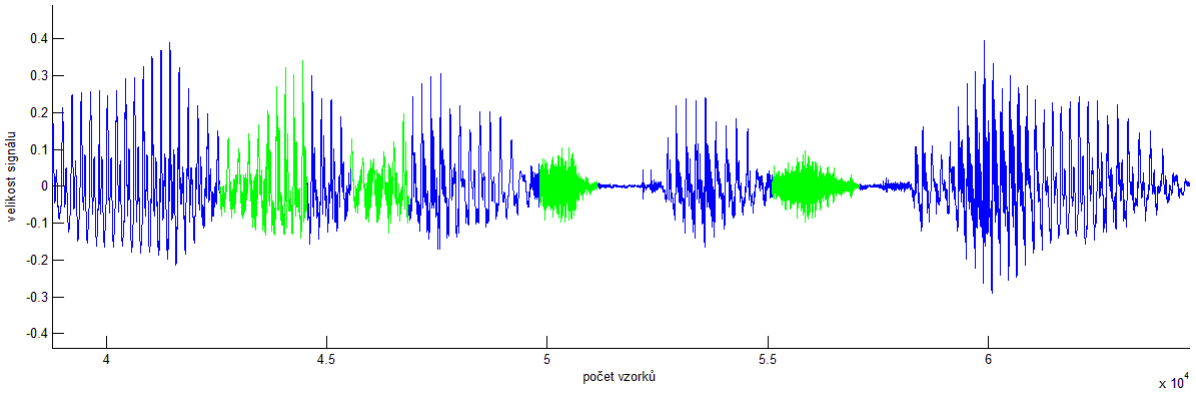
Tab. 9: Úspěšnost klasifikace samohlásek pro plynulou řeč s cizím přízvukem

Nakonec byl algoritmus testován pro telefonický hovor, který byl postižen kontinuálním šumem. Testování probíhalo na skupině dvou osob, které mluvily mírně zrychleným tempem. Testovací data byla opět získána z audio záznamů rádia stažených z internetu. Výsledky jsou v Tab.(10), pro kterou platí velmi nízká úspěšnost 48,8%. Z toho tedy vyplývá, že algoritmus je pro tento případ nepoužitelný.

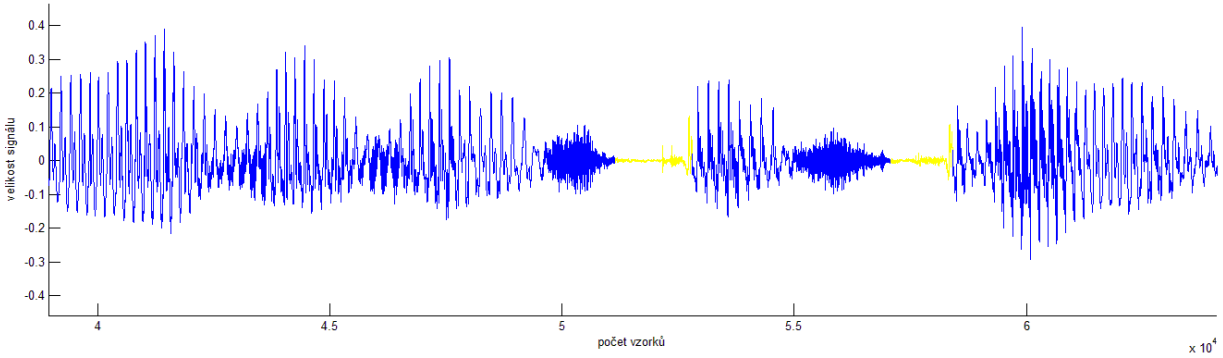
	frikativy	explozivy	nosovky	samohlásky
frikativy	6	3	2	7
explozivy	4	8	2	2
nosovky	2	3	2	9
samohlásky	2	3	2	23

Tab. 10: Úspěšnost klasifikace samohlásek pro telefonický hovor narušený šumem

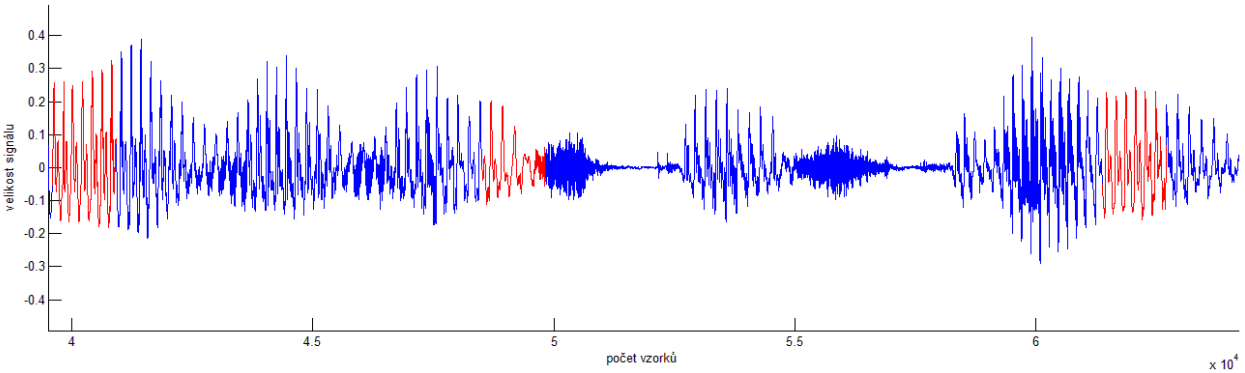
Na obrázcích níže je ukázka výsledků klasifikace souhlásek do jednotlivých skupin. Jedná se slovní spojení *Nizozemské strany*. Na Obr.(7) jsou zelenou barvou zvýrazněny oblasti výskytu frikativ. Frikativy jsou zde identifikovány správně. Oddělení mezi hláskami *zo* však není zcela přesné a část samohlásky byla tedy identifikována jako frikativa. Na Obr.(8) jsou explozivy zvýrazněny žlutou barvou. Zvýrazněné explozivy jsou zde identifikovány správně ovšem mírně zde zasahují do sousedních hlásek. Poslední Obr.(9) znázorňuje červenou barvou nosovky, které opět mírně zasahují do sousedních hlásek. Jsou ale identifikovány správně.



Obr. 7: Zobrazení frikativ



Obr. 8: Zobrazení exploziv



Obr. 9: Zobrazení nosovek

Algoritmus byl dále přizpůsoben pro konkrétní osobu. Toto přizpůsobení spočívalo ve vytvoření modelu pro klasifikaci samohlásek pro danou osobu. V programu Audacity byly namluveny úseky řeči, ze kterých byly následně ručně extrahovány hlásky. Bylo extrahováno 30 frikativ, 30 exploziv, 30 nosovek a 30 samohlásek. Následně byl stejnou procedurou vytvořen model pro konkrétní osobu. Poté byl algoritmus pomocí tohoto nového modelu otestován na plynulých úsecích řeči bez šumu. Výsledky klasifikace jsou v Tab.(11). Úspěšnost klasifikace byla 81,3%.

	frikativy	explozivy	nosovky	samohlásky
frikativy	29	1	3	3
explozivy	5	25	0	0
nosovky	0	1	17	0
samohlásky	8	0	5	42

Tab. 11: Úspěšnost klasifikace samohlásek pro konkrétní osobu

Úspěšnost klasifikace na základě vytvořeného modelu pro konkrétní osobu je tedy vyšší, než úspěšnost vytvořeného obecného modelu. Nevýhodou modelu vytvořeného pro konkrétní osobu ovšem je, že nemůže být použit pro jiné osoby, neboť jeho úspěšnost velice výrazně klesá.

4.2 Detekce frikativ *f,s* a *š*

Pro detekci jednotlivých frikativ *f,s* a *š* byly využity funkce na základě článku [8]. Na základě tohoto článku bylo vybráno 9 funkcí podle jejich vlastností, na základě kterých bylo možno jednotlivé frikativy rozlišit. Tyto vlastnosti byly následně pro každý segment uloženy do vektoru. Vznikla, tak matice n vektorů reprezentujících n analyzovaných segmentů. Pro získání těchto vlastností byly využity následující funkce.

1. Spektrální plochost. Určuje plochost spektra, která je vhodná pro oddělení znělých a neznělých úseků. Neznělé úseky (*f,s* a *š*) jsou reprezentovány hodnotami blízkými k 1. Naopak znělé části reprezentují hodnoty blíže k 0. Byl využit algoritmus [11].
2. Spektrální Rolloff. Jedná se o frekvenci f_r, která reprezentuje ve spektru hranici, pod kterou je koncentrováno p procent velikostí spektrálních složek. Na základě experimentů byla použita hodnota p = 50%. Byl využit algoritmus [11]. Podle [8]

$$\sum_{j=0}^{f_r} M_i(j) = p \sum_{j=0}^{J-1} M_i(j), \tag{4.5}$$

kde i je daný segment, j představuje spektrální složku a $M_i\ (j)$ je velikost spektrální složky. Byl využit algoritmus [11].

3. Spektrální těžiště. Spektrální těžiště je definováno jako těžiště velikostí spektrálních složek. Byl využit algoritmus [11]. Podle [8]

$$S_i = \frac{\sum_{k=0}^{K-1} M_i(j) \cdot j}{\sum_{k=0}^{K-1} M_i(j)}, \tag{4.6}$$

kde *i* je daný segment, *j* představuje spektrální složku, $M_i\ (j)$ je velikost spektrální složky a S_i reprezentuje danou frekvenci.

4. Poměr energií dvou frekvenčních pásem. Podle [8]

$$E_t = 10\log_{10}\left(\frac{B_1}{B_2}\right), \tag{4.7}$$

kde B_1 = 4000 – 8000 Hz, B_2 = 2000 – 4000 Hz, E_t je energie daného segmentu.

5. Počet průchodů nulou. Počet průchodů nulou je počítán pro každý segment v časové oblasti. Tato funkce je vhodná pro oddělení frikativ od ostatních souhlásek. Vypočítá se jako [8]

$$ZCR = \frac{1}{2}\sum_{n=1}^{N-1}\left|\mathrm{sgn}(s[k]) - \mathrm{sgn}(s[k-1])\right|, \tag{4.8}$$

kde s(k) je signál v časové oblasti pro daný segment a *k* je daný vzorek.

6. Melovské frekvenční kepstrální koeficienty. Jsou počítány jako logaritmus velikostí spektrálních složek, kde jsou jednotlivé spektrální složky seskupeny podle Melovské frekvenční škály. Následně jsou podrobeny diskrétní kosinově transformaci (DCT). Na základě článku [8] byli využity pouze první tři koeficienty, pro vektor vlastností. Byl využit algoritmus [11].
7. Frekvence nejvyššího vrcholu podle gammatonové banky filtrů. Tyto gammatonové spektrogramy jsou podobné běžným spektrogramům. Narozdíl od nich ale nemají konstantní šířku všech pásem. Gammatonové banky filtrů respektují vlastnosti lidského ucha a jejich pásma jsou na vyšších frekvencích širší. Byl využit algoritmus [12].

Trénovací data pro vytvoření SMV modelu byla vytvořena ze skupiny 30 písmen *f*, 30 písmen *s*, 30 písmen *š* a 120 ostatních písmen. Tato trénovací data byla získána z audio záznamů rádia, která byla stažena z internetu [13] od šesti osob a to stejných, kteří byly využiti pro trénovací data pro klasifikaci souhlásek. Na základě těchto dat byl vytvořen obecný model pro detekci frikativ. Algoritmus byl pomocí vytvořeného modelu dále detailně testován pro více osob a různou kvalitu řeči. Testování probíhalo nejprve na audio záznamech

rádia tří osob a to jiných osob než, které byly využity pro trénování dat. Promluva těchto řečníků byla plynulá a v pozadí nebyl šum. Tab.(12) ukazuje úspěšnost klasifikace, která je 80,8%. Docházelo zde k častým záměnám *ř* a *ž* se *s*.

	s	**š**	**f**	**ostatní**
s	25	0	0	3
š	4	7	0	2
f	1	0	5	3
ostatní	10	1	3	68

Tab. 12: Úspěšnost detekce frikativ pro plynulou promluvu bez šumu

Algoritmus byl dále testován na skupině dvou osob s plynulou promluvou bez šumu, ale s mírným cizím přízvukem (němčina, angličtina). Testovací data byla opět získána z audio záznamů rádia stažených z internetu. Tyto osoby byly stejné jako při testování klasifikace frikativ. Tab.(13) znázorňuje tyto výsledky. Úspěšnost klasifikace byla 78,4%.

	s	**š**	**f**	**ostatní**
s	12	0	0	6
š	0	7	0	4
f	0	0	2	1
ostatní	0	0	5	37

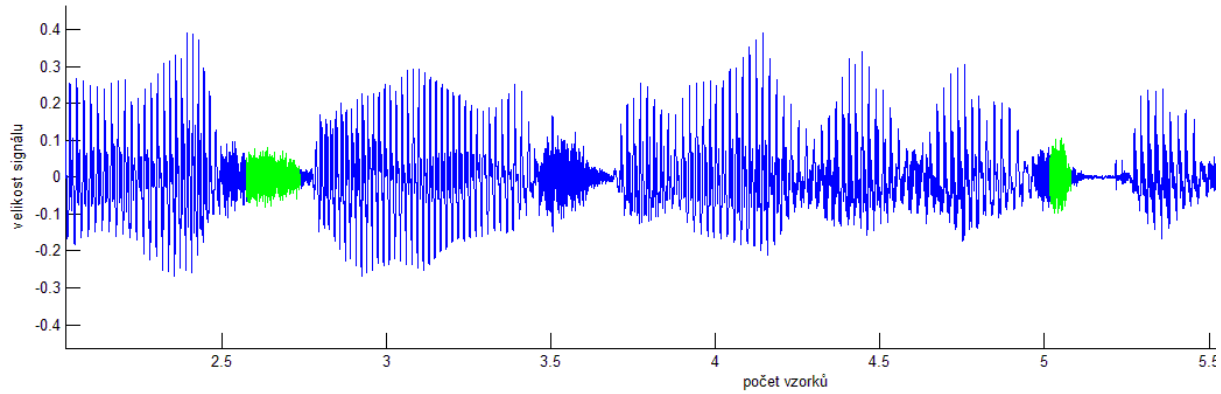
Tab. 13: Úspěšnost detekce frikativ pro plynulou řeč s cizím přízvukem

Nakonec byl algoritmus testován pro telefonický hovor, který byl postižen kontinuálním šumem. Testování probíhalo na skupině dvou osob, které mluvily mírně zrychleným tempem. Testovací data byla opět získána z audio záznamů rádia stažených z internetu. Výsledky jsou v Tab.(14), pro kterou platí velmi nízká úspěšnost 40%. Z tabulky je také patrné, že nebylo identifikováno žádné písmeno *f*. Z toho tedy vyplývá, že algoritmus je pro tento případ opět nepoužitelný.

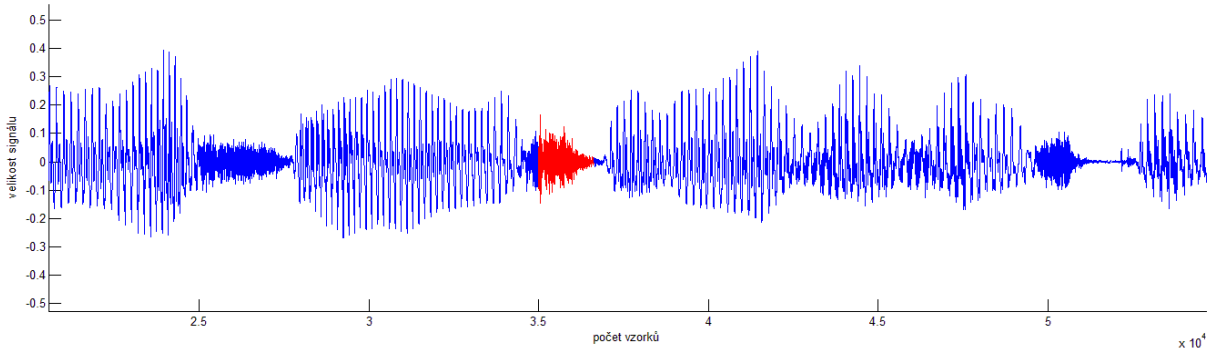
	s	**š**	**f**	**ostatní**
s	4	4	0	6
š	0	5	0	4
f	0	0	0	0
ostatní	0	14	7	31

Tab. 14: Úspěšnost detekce frikativ pro telefonický hovor narušený šumem

Na obrázcích níže je ukázka výsledků detekce frikativ a jejich přiřazení do jednotlivých skupin. Jedná se slovní spojení *nejsilnější Nizozemské*. Na Obr.(13) jsou zelenou barvou zvýrazněny oblasti výskytu frikativy *s*. Frikativy jsou zde identifikovány správně ovšem jejich zvýraznění není zcela přesné a části frikativ tak nejsou plně zvýrazněny. Na Obr.(14) jsou frikativy *š* zvýrazněny červenou barvou. Frikativa š je zde správně identifikována ovšem její zvýraznění není zcela přesné a část frikativy tak není zvýrazněna.



Obr. 10: Zobrazení frikativy *s*



Obr. 11: Zobrazení frikativy *š*

Algoritmus byl dále přizpůsoben pro konkrétní osobu. Toto přizpůsobení spočívalo ve vytvoření modelu pro detekci frikativ pro danou osobu. V programu Audacity byly namluveny úseky řeči, ze kterých byly následně ručně extrahovány hlásky. Bylo extrahováno 15 fonémů *f*, 15 fonémů *s*, 15 fonémů *š* a 60 ostatních fonémů. Následně byl stejnou procedurou vytvořen model pro konkrétní osobu. Poté byl algoritmus pomocí tohoto nového modelu otestován na plynulých úsecích řeči bez šumu. Výsledky klasifikace jsou v Tab.(15). Úspěšnost klasifikace byla 81,5%.

	s	**š**	**f**	**ostatní**
s	21	0	0	1
š	1	14	0	0
f	0	0	8	3
ostatní	5	1	3	54

Tab. 15: Úspěšnost detekce frikativ pro konkrétní osobu

Úspěšnost detekce na základě vytvořeného modelu pro konkrétní osobu je tedy opět vyšší, než úspěšnost vytvořeného obecného modelu. Nevýhodou modelu vytvořeného pro konkrétní osobu je, že nemůže být použit pro jiné osoby, neboť jeho úspěšnost velice výrazně klesá.

5 ZÁVĚR

První část práce se zabývá převážně teorií. Jsou zde popsány podrobněji popsány akustické vlastnosti souhlásek, způsoby pro analýzu signálů a nakonec je zde shrnuta problematika automatického rozpoznávání řeči.

Další část práce se zabývá používanými metodami pro automatickou detekci jednotlivých souhlásek. Je zde popsána metoda na detekci explozivních souhlásek pomocí vlnkové transformace, dále dvě metody na rozpoznávání frikativ, z nichž jedna byla využita pro návrh programu.

Poslední částí práce je popis vlastního algoritmu pro jehož vývoj bylo využity metody strojového učení tzv. Support vector machines (SVM). Tato metoda byla vybrána z důvodu své jednoduchosti. Základem této metody je vytvoření modelu pomocí matice trénovacích dat. Pomocí tohoto modelu pak mohou být odděleny jak skupiny souhlásek, tak jednotlivé souhlásky. Trénovací data tedy musí obsahovat několik skupin, které mají být odděleny. Každé této skupině je přiřazena určitá celočíselná hodnota, pomocí níž je následně identifikována. Cílem této práce bylo klasifikovat skupiny souhlásek na frikativy, explozivy a nosovky a také detekovat jednotlivé souhlásky. Pro tuto práci byly vybrány souhlásky *s*,*f* a *š*. Byl tedy vytvořen jednotný ucelený program. Cílem této práce bylo také přizpůsobit program pro konkrétní hlas. Ve výsledku tedy byly vytvořeny čtyři SVM modely. Dva byly vytvořeny pro obecnou klasifikaci a detekci souhlásek. A další dva byly vytvořeny pro konkrétní hlas. Na základě testování algoritmu pomocí těchto modelů bylo zjištěno, že modely vytvořené pro konkrétní hlas dokáží lépe odhalovat, jak skupiny souhlásek, tak jednotlivé souhlásky. Největší úspěšnosti detekce bylo dosaženo při detekci jednotlivých souhlásek *s*,*f* a *š*.

Nevýhodou této metody je, že trénovací data musí obsahovat také, kromě skupin, které mají být odděleny, navíc jednu skupinu, která obsahuje co nejvíce ostatních hlásek, které nepatří, ani mezi jednu skupinu. A to proto, aby je bylo možné oddělit od testovaných úseků řeči.

6 LITERATURA A INTERNETOVÉ ZDROJE

[1] AHMED, M. A. A., van der SPIEGEL, J., MUELLER, P. An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants.Philadelphia: Department of Electrical Engineering, 1998. Dostupné na www: http://ieeexplore.ieee.org

[2] IEEE [online]. [cit. 2014-05-30]. Dostupné na www: http://ieeexplore.ieee.org/Xplore/home.jsp

[3] Leccos [online]. [cit. 2014-05-30]. Dostupné na www: http://leccos.com/index.php/clanky/souhlaska

[4] MALBOS, F., BAUDRY, M., MONTRESOR, S. Detection of stop consonants with the wavelet transform. France: Laboratoire d´Informatique de l´Université du Maine, 1994. Dostupné na www: http://ieeexplore.ieee.org

[5] PSUTKA, J., MÜLLER, Z., MATOUŠEK, J., RADOVÁ, V. Mluvíme s počítačem česky. Praha: Academia, 2006.

[6] SENEFF, S., Joint Synchrony/Mean Rate Model of Auditory Speech Processing. J. Phonetics, 1988.

[7] SIGMUND, M. Rozpoznávání řečových signálů. Skriptum FEKT VUT v Brně. Brno: MJ servis, 2007.

[8] FRID, Alex a Yizhar LAVNER. Acoustic-phonetic analysis of fricatives for classification using SVM based algorithm. *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel* [online]. IEEE, 2010, s. 000751-000755 [cit. 2014-05-30]. DOI: 10.1109/EEEI.2010.5662110. Dostupné z: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5662110

[9] CHANG, Chih-Chung a Chih-Jen LIN. LIBSVM -- A Library for Support Vector Machines. [online]. [cit. 2014-05-30]. Dostupné z: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[10] HSU, Chih-Wei, Chih-Chung CHANG a Chih-Jen LIN. A Practical Guide to Support Vector Classification. [online]. [cit. 2014-05-30]. Dostupné z: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[11] Audio Content Analysis: Machine Listening & Music Information Retrieval. [online]. [cit. 2014-05-30]. Dostupné z:http://www.audiocontentanalysis.org/code/

[12] ELLIS, D. P. W. Gammatone-like spectrograms. [online]. [cit. 2014-05-30]. Dostupné z: http://www.ee.columbia.edu/ln/rosa/matlab/gammatonegram/

[13] IRadio. [online]. [cit. 2014-05-30]. Dostupné z: http://hledani.rozhlas.cz/iradio/

[14] HADJILEONTIADIS, L.J. A Texture-Based Classification of Crackles and Squawks Using Lacunarity. *IEEE Transactions on Biomedical Engineering* [online]. 2009, vol. 56, issue 3, s. 718-732 [cit. 2014-05-30]. DOI: 10.1109/TBME.2008.2011747. Dostupné z: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4760231

[15] REBER, Jonas. Matlab Central. [online]. [cit. 2014-05-30]. Dostupné z: http://www.mathworks.co.uk/matlabcentral/answers/8704-i-can-not-use-libsvm

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

a	Měřítko vlnky
b	Časový posun
BDSONS	Base de données des sons du français - Databáze francouzské zvuky
B_w	Šířka pásma filtru
CVC	Consonant – Vowel – Conconant, Souhláska – Samohláska – Souhláska
DUP	Duration of the unvoiced portion, Doba trvání neznělých částí
E	Energie
GSD	Generalized Synchrony Detector
h(n)	Funkce blíže nespecifikovaného okénka
h(t)	Základní vlnka
$h_q(n)$	Okénko typu dolní propust
$h_{q*}(n)$	Impulzní odezva daného pásmového filtru
i	Číslo segmentu
j	Spektrální složka
L	Počet vzorků okénka
LOWE	Low Energy
LOWG	Low GSD
MDP	Most Dominant Peak, Nejvýznamnější vrchol
MNSS	Maximum Normalized Spectral Slope
$M_i(j)$	Velikost spektrální složky
P	Perioda
p	Procento
p(t)	Fourierova transformace
PCM	Pulse Code Modulation, Pulzní kódová modulace
Q_n	Krátkodobá charakteristika řečového signálu v čase
RA	Relative Amplitude, Relativní amplituda
R_n	Krátkodobá autokorelační funkce
s(k)	Vzorek akustického signálu v čase k
N	Počet vzorků segmentu
n	Diskrétní proměnný čas
nod	Počet detekovaných exploziv

nte	Celkový počet detekcí
nto	Počet exploziv ve sledovaném korpusu
SCG	Spectral Center of Gravity
S_i	Spektrální těžiště
TIMIT, C_a	Řečové databáze
w(n)	Funkce okénka
$yenv_i$	Výstup „mean-rate“ z i-tého filtru
ZCR	Počet průchodů nulou
$\tau(.)$	Transformační funkce
ω_q	Střední frekvence pásmového filtru