VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

SOFTWAROVÁ PODPORA ANALÝZY EMOCIONÁLNÍCH STAVŮ

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

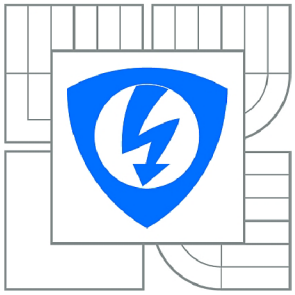
AUTOR PRÁCE
AUTHOR

JAKUB LNĚNIČKA

BRNO 2010

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF TELECOMMUNICATIONS

SOFTWAROVÁ PODPORA ANALÝZY EMOCIONÁLNÍCH STAVŮ

SW SUPPORT FOR EMOTIONAL STATE ANALYSIS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

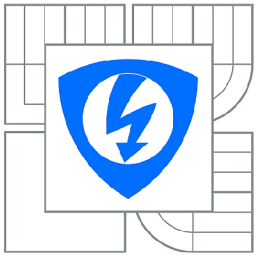
AUTOR PRÁCE — JAKUB LNĚNIČKA

AUTHOR

VEDOUCÍ PRÁCE — prof. Ing. ZDENĚK SMÉKAL, CSc.

SUPERVISOR

BRNO 2010



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Bakalářská práce

bakalářský studijní obor
Teleinformatika

Student:	Jakub Lněnička	***ID:***	106597
Ročník:	3	***Akademický rok:***	2009/2010

NÁZEV TÉMATU:

Softwarová podpora analýzy emocionálních stavů

POKYNY PRO VYPRACOVÁNÍ:

Vytvořte v prostředí Matlab knihovnu funkcí pro podporu analýzy emocionálních stavů. Zaměřte se hlavně na automatické, poloautomatické a manuální metody značkování multimediálních záznamů. Cílem bakalářské práce je SW nástroj s grafickým rozhraním, který je využitelný pro vytváření multimodálních emocionálních databází. Dále by tento program měl být schopen komunikaci se vzdáleným serverem pro archivaci výsledků.

DOPORUČENÁ LITERATURA:

[1] PSUTKA, J., MULLER, L., MATOUŠEK, J., RADOVÁ, V.: Mluvíme s počítačem česky. ACADEMIA, Praha 2006. ISBN 80-2100-1309-1
[2] KRČMOVÁ, M.: Fonetika. Elektronické texty. MU Brno 2003.
http://is.muni.cz/do/1499/el/estud/ff/js07/fonetika/materialy/index.html

Termín zadání:	29.1.2010	***Termín odevzdání:***	2.6.2010

Vedoucí práce: prof. Ing. Zdeněk Smékal, CSc.

prof. Ing. Kamil Vrba, CSc.
Předseda oborové rady

UPOZORNĚNÍ:

Abstrakt

Cílem bakalářské práce je vytvořit SW nástroj s grafickým rozhraním, který je využitelný pro vytváření multimodálních emocionálních databází. V úvodu se práce zabývá popisem částí lidského těla vytvářejících hlas a jejich fungováním. Dále se text věnuje problematice zpracování lidského hlasu do digitální formy, velká pozornost je věnována parametrům řečového signálu s důrazem na popis příznaků sloužících k určení vybraných emocí.

Práce se dále zabývá kategorizací emocí a popisem některých z nich. V závěru práce je popsán klasifikátor K-NN, který slouží k určení jednotlivých pocitů a vytvořený program.

Klíčová slova: emoce, určení emocí, Matlab, klasifikátor KNN, základní tón řeči

The goal of my bachelor work is the description of SW tool with the graphic interface which can be made use of for the purpose of developing the multimodal emotional databases.

In its beginning my work deals with the description of the parts of the human body that produce voice (vocal cords) and their functioning. The text is a description of the procedure of transferring the human voice into the digital form where a special attention is paid to the parameters of the speech signal with the emphasis on the description of the symptoms that serve to the purpose of defining the chosen emotions.

This work deals with the categorization of emotions and the description of some of them. In the closing part the K-NN classificator is described that serves to the recognition of the individual feelings by means of a produced software.

Key words: emotions, determine of emotion, Matlab, classifer KNN, the basic of speech.

LNĚNIČKA, J. Softwarová podpora analýzy emocionálních stavů. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2010. 38 s. Vedoucí bakalářské práce prof. Ing. Zdeněk Smékal,CSc.

Prohlašuji, že svoji bakalářskou práci na téma: Softwarová podpora analýzy emocionálních stavů jsem vypracoval samostatně pod vedením vedoucího bakalářské práce s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny uvedeny v seznamu literatury na konci práce.

Jako autor uvedené bakalářské práce dále prohlašuji, v souvislosti s vytvořením této bakalářské práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení §11 a následujících autorského zákona č.121/2000Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení §152 trestního zákona č.140/1961Sb.

V Brně dne..... Podpis autora.....

Obsah

Seznam obrázků

Úvod

Komunikace prostřednictvím mluvené řeči je základním a nejpoužívanějším prostředkem přenosu informace mezi lidmi. Díky řeči je člověk schopen vyjádřit různé myšlenky, nápady a pocity neboli emoce. Umění mluvit a rozumět se učíme už jako malé děti a od té doby ji považujeme za samozřejmou činnost. Jedná se však za jednu z nejsložitějších posloupností akcí, jejichž úspěšné zvládnuti má vliv na průběh celého procesu komunikace. Přenos informace obvykle začíná vytvořením myšlenky (zprávy), kterou chceme sdělit, pokračuje přenosem této myšlenky (tj. vlastní realizací promluvy akustickým řečovým signálem) a končí rozpoznáním akustického signálu posluchačem, včetně porozuměním významu přenášené zprávy, které samozřejmě obsahuje porozumění jejího významu .

V dnešní době se klade stále větší důraz na výzkum v oblasti řečových signálů. Nové objevy a samotný rozvoj techniky tak v různých oblastech otevřely cestu využití počítačových analýz řečového signálu (např. ve zdravotnictví, psychologii, rozpoznání bezpečnostních prvků , aplikace potřebné k úpravě hlasu do přirozenější formy).

Nepřímá komunikace pomocí moderních technologií, např. mobilní telefon, internet apod., je již v dnešní době neodmyslitelnou součástí našeho života. Při dnešním rozvoji digitální techniky je řečový signál pomocí *A/D* převodníků převáděn do číslicového tvaru a nese s sebou mnoho druhů informací. Skutečná informace je hlavní nositelkou myšlenky a podstatnou součástí komunikace. Další, na první pohled podružnou složkou řeči, je informace emoční. Ta nám vyjadřuje buď okamžitý duševní stav mluvčího, nebo jeho postoj, který zaujímá k právě probíranému tématu. Emoční postoj je emoce, kterou v tomto v tomto případě mluvčí vědomě i nevědomě předává dále. Emoce jsou vnímány z prozodie řeči

Rozpoznání řeči stále více proniká do zařízení běžných potřeb ve všech možných odvětvích. Stále více uplatnění nalézají také poznatky o vlastnostech řeči v nepřirozeném emočním stavu (např. vlivem stresu, únavy, apod.). V oblasti rozpoznání pocitů se potýkáme s problematikou nekvalitního materiálu, který je však základem úspěchu výzkumu. Náš výzkum se zaměřuje na získání příznaků, které by co nejlépe dokázaly charakterizovat emoční stav člověka. Samotná práce pak rozebírá vlastnosti u řečového signálu a na řeč se dívá z technického i prozodického pohledu.

1. Prozodie

1.1 Zvuková stránka souvislé řeči

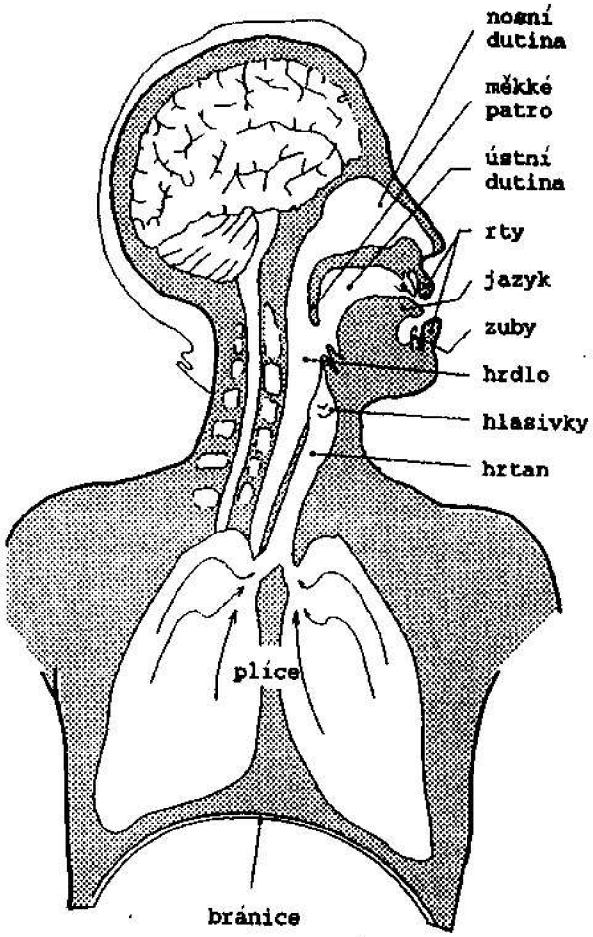
Souvislá řeč není monotónní. Je modulována pomocí síly a výšky hlasu a získává určitý rytmus díky proměnlivému časovému průběhu jednotlivých segmentů a jejich kombinací. Dalšími prostředky modulace je řečové tempo a existence různých typů pauz. Tyto zvukové prostředky se uplatňují na promluvě jako celku. Prostředky modulace jsou jednak přirozenou složkou zvukového signálu řeči (síla hlasu nejen přirozeně existuje, ale mění se i fyziologicky v průběhu řeči), jednak mají i komunikační funkce – zejména při vyjadřování pragmatických složek komunikace. Jsou velmi důležité, protože pouhá napodobení hlásek či slabik bez náležité modulace je velmi těžko srozumitelná.

Na rozdíl od hlásek, které lze na základě zobecnění vyčlenit z proudu řeči jako jednotky segmentální, je modulace řeči i její základní modely, o nichž bude řeč dále, založena vztahem mezi segmenty. Delší slabiku poznáme jen v relaci k jiné stejně strukturované, méně dlouhé slabice, silnější řeč poznáme jen v relaci k méně silné, pauzu na základě toho, jak vypadá úsek bez pauzy atd.. Kromě toho jsou vždy v daném okamžiku přítomny všechny tyto složky, i když v odlišné míře. Není tedy možné je popsat jen základě pár vybraných příznaků a ostatní zanedbat. Snad i to způsobuje, že se jim ve fonetickém zkoumání dostává pozornosti daleko později než hláskám [2].

1.2 Dechové ústrojí

Dechové ústrojí tvoří základní zdroj energie. Je umístěno v hrudním koši a tvořeno přívodní dýchací cestou, plícemi a s nimi funkčně spjatými dýchacími svaly (bránicí). Při nádechu dochází k pohybu vzduchu, který tak poskytuje zdroj energie pro řeč. Při výdechu potom v plicích vzniká výdechový proud vzduchu, který je zásadě základním materiálem pro tvorbu řeči Výdechový proud vzduchu je z plic odváděn z průdušnic (tracheou), a pak prochází hrtanem a nadhrtanovým dutinami, kde se modifikuje, a jako řečový signál je vyzařován rty do okolního prostoru.

Trvání výdechu má vliv na to, jak dlouhý úsek řeči lze vytvořit bez přerušeni. Síla výdechového proudu ovlivňuje způsob fungování hlasového ústrojí, a tím má vliv na sílu hlasu a částečně i na jeho výšku. K vytvoření „slyšitelné“ řeči je zapotřebí z plic vytlačit v rozmezí několik sekund více než 0,5 litru vzduchu. Během běžné řeči se spotřebuje až polovina vitální kapacity plic, tj. maximální rozdíl mezi úplně naplněnými a vypuštěnými plícemi, zatímco při velmi hlasité řeči až 80 %. Další nádech pak opět doplňuje vzduch do plic, čímž mimo jiné dodává nový materiál pro tvorbu řeči. Navenek se nádech projeví jako pauza jinak souvislé řeči [2].



Obr. 1.1 :Hlasový trakt člověka

:

1.3 Hlasové ústrojí

Pod tímto pojmem je možné chápat celý systém pro vytváření hlasu. V našem případě je ale myšlena jenom jako část, kde dochází k samotnému vzniku hlasu. Hlasové ústrojí je uloženo v hrtanu (larynx), který je s plícemi spojen průdušnicí. Z hlediska tvorby řeči nedůležitější část hlasového ústrojí tvoří hlasivky (plicae vocales), které se nacházejí v hrtanové dutině přímo za „ohryzkem“, „Adamovo jablko“. Jsou to dvě ostré slizniční řasy, které vedou napříč hrtanem v místě jeho nejužšího průchodu. Jejich typická délka je asi 15 mm pro muže a 13 mm pro ženu. Z jedné strany jsou napojeny na chrupavky hlasivkového a z druhé strany na chrupavku štítnou. Jsou pokryty sliznicí a jejich základ tvoří hlasový sval. Prostor mezi hlasivkami tvoří hlasivková štěrbina trojúhelníkového tvaru (glottis). Jestliže člověk mlčí, pak hlasivky drží hlasivkovou štěrbinu odkrytou, takže ji může bez odporu procházet vzduch k dýchání. Hlasové ústrojí tak využívá klidového postavení hlasivek [2].

Při vytváření hlasu (fonační) plní hlasivky jinou funkci – nacházejí se v tzv. hlasovém (fonačním) postavení. Výdechový proud vzduchu postupuje bez odporu z plic průdušnicí až k hrtanu. Zde se mu do cesty postaví překážka vytvořená hmotou hlasivek, které cestu vzduchu úplně uzavřou. Stažené hlasivky se pod tlakem vzduchu

stávají pružnými a začínají kmitat. Hlasivky se přitom střídavě postupně otevírají a prudce uzavírají. V důsledku kmitání hlasivek se vzduchový proud (do té doby homogenní) „rozdrobí“ tak, že se víceméně pravidelně střídá vždy kvantum hustšího a řidšího vzduchu – vzniká tzv. vzduchová vlna, kterou vnímáme jako zvuk. Tento periodický proud vzduchových pulzů tvoří základ lidského hlasu. Bývá označován termínem základní (hlasivkový) tón, který představuje nosný zvuk řeči. Frekvence kmitání hlasivkového se označuje f_0 a nazývá se fundamentální frekvence nebo frekvence základního hlasivkového tónu. Tato frekvence je fyzikální charakteristikou řečového signálu a odpovídá výšce hlasu tak, jak ji vnímá posluchač. Udává se, že základní hlasivkový tón má rozsah asi 60-400 Hz. Frekvence kmitání je různá u dětí a dospělých, ale i u žen a mužů. Při normální řeči se pohybuje zhruba kolem jedné oktávy (u mužů asi mezi 80-160 Hz s průměrnou hodnotou 132 Hz, u žen pak mezi 150-300 Hz s průměrnou hodnotou 223 Hz a u dětí dokonce v rozmezí 200 až 600 Hz). Pro hluboké mužské hlasy však může f_0 klesnout až k 50 Hz a naopak u vysokých ženských hlasů muže vystoupat až přes 400 Hz. Při zpěvu se hlasový rozsah zvětšuje asi na dvě oktávy (u školených jedinců až na tři oktávy) a pro sopranistky není výjimkou základní frekvence podstatně převyšující 1000 Hz.

Klidové postavení hlasivek a postavení fonační představují dvě základní protikladné varianty činnosti hlasového ústrojí. Obou těchto poloh (v drobných modifikacích) je využíváno při tvorbě řeči. Fonační postavení má za následek vznik hlasivkového tónu a používá se proto při vytváření znělých zvuků řeči (tj. samohlásek a znělých souhlásek). Oba typy hlásek se především v míře svalového napětí a těsnosti přiblížení hlasivek. Při tvorbě samohlásek je hlasivková štěrbina téměř úplně uzavřená po celé délce a hlasové vazy jsou napjaté, kmitání hlasivek je tak pravidelné a vznikající zvuk má „čistě“tónovou strukturu. Na druhou stranu u znělých souhlásek může být napětí hlasivek menší, kmitání je pak méně pravidelné a charakteristika výsledného zvuku již není čistě tónová. Neznělé zvuky jsou naopak tvořeny při klidovém postavení hlasivek, neobsahují tedy základní hlasivkový tón a vznikají až modifikací výdechového proudu vzduchu nadhrtanových dutinách. Na rozdíl od prostého dýchaní (když člověk mlčí) však při vytváření neznělé řeči nebývají hlasivky úplně povolený a ani hlasivková štěrbina tak nebývá úplně rozevřená [2].

1.4 Artikulační ústrojí

Artikulační ústrojí je posledním ústrojím, které se podílí na tvorbě řeči. Jeho význam spočívá v tom, že umožňuje vytvářet velké množství různých zvuků, které charakterizují mluvený jazyk. Skládá se jednak z nadhrtanových dutin, a jednak z artikulačních orgánů, které jsou v těchto dutinách uloženy nebo je obklopují. Mezi nadhrtanové dutiny řadíme dutiny hrdelní, ústní a nosní. Hranici mezi těmito dutinami tvoří čípek (uvula), špička měkkého patra (velum), které zamezuje nebo umožňuje přístup vzduchu z dutiny hrdelní do dutiny nosní.

Zatímco se nadhrtanové dutiny účastní procesu tvorby řeči pasivně (nepohybují se), artikulační orgány (artikulátory) se účastní tvorby řeči většinou aktivně – tvoří pohyblivé součásti artikulačního ústrojí a svým pohybem mění velikosti nadhrtanových dutin. Z hlediska vytváření řeči mezi nejvýznamnější artikulátory patří jazyk, rty a měkké patro, neboť se podílejí na vytváření největšího počtu různých zvuků. Dalšími artikulátory potom jsou zuby, tvrdé patro nebo čelisti. Artikulátorem je také hrtan, který kromě toho, že se zapojuje do tvorby znělosti, se také může pohybovat nahoru a dolů a měnit tak délku celého hlasového traktu.

Vůbec nejdůležitějším a nejsložitějším artikulátorem je jazyk. Skládá se ze tří částí: hrotu (špičky), hřbetu a kořene, které jsou součástí stejné svalové struktury jazyka, ale každá může do jisté míry fungovat nezávisle. Jazyk je tak velice pružný a přizpůsobivý, schopný tvořit mnoho tvarů a rychle přecházet z jedné pozice do druhé. Právě variabilita umisťování jazyka vytváří nesčetný tvary ústní i hrdelní dutiny a vede tak k vytváření různých zvuků řeči

Výrazných změn ve svém složení doznává zvuk v nadhrtanových dutinách, kam jako výdechový proud vzduchu (v podobě periodického proudu vzduchových pulsů nebo jako prostý proud vzduchu) postupuje z hlasového ústrojí. Tyto změny jsou zásadně dvojího druhu:

- Vytváření tónové struktury. Při průchodu základního hlasivkového tónu nadhrtanovými dutinami dochází vlivem rezonance uvnitř hrdelní, ústní a popř. i nosní dutiny ke změně rozložení akustické energie ve vznikajícím řečovém signálu. Akustická energie se soustředí kolem určitých frekvencí, kterým říkáme formantové frekvence. Oblasti koncentrace (zesílení) akustické energie se nazývají formanty a označují se čísly, počínaje formantem s nejnižší frekvencí. f_1, f_2, …, f_n. Pokud se do procesu vytváření řeči zapojí i nosní dutina, dochází navíc vlivem jejich antirezonančních vlastností k potlačení některých frekvenčních oblastí. Tzv. antiformantů (značí se . A_1, A_2, …, A_m). Vzniká složený zvuk tónového charakteru, kterému říkáme hlas. Tvoří podstatu znělých částí řeči, především samohlásek. Různé zvuky (tj. zvuky s různou spektrální strukturou) přitom vznikají tak, že pohyblivé artikulátory mění tvar nadhrtanových dutin, a tím i frekvenční vlastnosti vy tvářených zvuků.
- Vytváření šumové složky. Artikulátory ovlivňují průchod vzduchu nadhrtanovými dutinami. Svým pohybem mohou průchod na různých místech zúžit, úplně uzavřít nebo naopak uvolnit. Výdechový proud vzduchu se pak prodírá přes vytvořené překážky a podle charakteru překážky vzniká šum různého druhu. Šum tvoří základ těch částí řeči, kterým při popisu jazyka říkáme souhlásky. Různé zvuky se tomto případě vytvářejí pomocí různých typů překážek umístěných na různých místech hlasového traktu.

Během řeči dochází k plynulému pohybu všech artikulátorů. Lze si přitom představit, že po jistou velmi krátkou dobu každý artikulátor zůstává přibližně v určité jedné poloze a soubor všech artikulátorů tak tvoří jistou konfiguraci hlasového traktu. Pohyby artikulátorů mění jednotlivé konfigurace podle předem daného programu. Řečník si totiž nejprve přeje vytvořit posloupnost zvuků, které odpovídají sdělení, jež chce vyjádřit, a potom mozek vysílá signály artikulátorům, jak a jakým způsobem mají fungovat. Výsledná posloupnost konfigurací hlasového traktu je tedy dána posloupností zvuků, které mluvčí vytváří. Je důležité si ale uvědomit, že ke změně konfigurací nedochází „najednou“ – hlasový trakt není diskrétní systém. Přechod artikulátorů z jedné polohy do druhé tedy není okamžitý – dosažení cílového postavení řečových orgánů při vytváření řeči má také za následek, že konfigurace hlasového traktu pro určitý zvuk nezávisí pouze na povaze tohoto zvuku, ale také na okolních zvucích v dané promluvě. Tento jev je typickým pro celý proces artikulace a nazývá se koartikulace. Projevuje se zejména při přechodu z jedné konfigurace hlasového traktu, dané právě vytvářeným zvukem, do konfigurace jiné, dané následujícím zvukem, a to i při velmi

pečlivé výslovnosti. Navenek se koartikulace projeví tím, že se akustické realizace stejné hlásky mění v závislosti na kontextu okolních hlásek [2].

2. Řečový signál

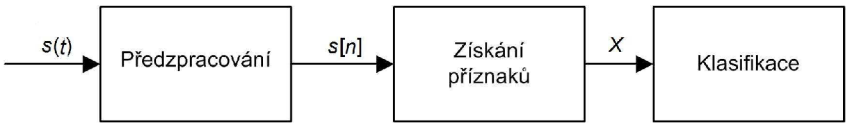
Řečový signál se ve všech oblastech zpracování řeči hodnotí výhradně v číslicové podobě. Využívají se k tomu s výhodou výkonné algoritmy a toolboxy, které jsou zejména od osmdesátých let k dispozici.

Celý proces automatického rozpoznání signálu můžeme rozdělit na tři základní segmentovat stupně .

Zpracováni řeči začíná blokem předzpracování (pre-processing), který obsahuje digitalizaci a několik standardních operací na úpravu signálu. Přímá klasifikace podle časového průběhu signálu $s(t)$ popř. $s[n]$ není možná vzhledem ke značné variabilitě a velkému počtu vzorků. Proto je nutné z řečových signálů získat jen několik málo důležitých příznaků x, aniž by se přitom ztratily důležité části informace obsaženy v signálu. Důležitost informace je dána konečným cílem zpracování, např. při kódování řeči je důležité zachování tvaru signálu, při rozpoznání mluvčího jsou naopak příznaky charakteristické pro jednotlivé mluvčí. Pod pojmem příznaky (features) rozumíme vlastnosti obrazce vyjádřené kvantitativně. Signál v této formě převeden ze „signálového prostoru" do „příznakového prostoru" je již připraven k vlastnímu zpracování informace a může být provedena klasifikace podle vektoru příznaků x.

Předzpracování

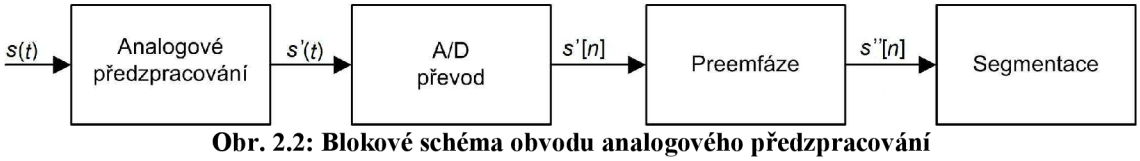
Lidská řeč a tím i řečový signál jsou značně variabilní. Nikdo není schopen vyslovit slovo dvakrát naprosto stejně, tzn. dodržet stejný přízvuk, výšku tónu, hlasitost a rychlost promluvy. Ještě větší rozdíly jsou mezi různými mluvčími. Podstatný vliv na charakter řečového signálu mají také rušení, okolní zvuky a rovněž zkreslení při přenosu signálu, tzn. kmitočtové charakteristiky mikrofonů, filtrů a zesilovačů a vlastnosti přenosových cest při dálkovém přenosu řeči. Jmenované vlivy snižují celkovou úspěšnost rozpoznání procesu. Proto je účelné některé z uvedených vlivů potlačit hned na počátku procesu vhodným předzpracováním.



Obr. 2.1: Blokové schéma obvodu předzpracování

2.1 Analogové předzpracování

Analogové předzpracování, tj. manipulace s řečovým signálem do té doby, než bude prezentován sledem „vzorků", začíná převodem změn akustického tlaku na elektrický signál. Při „živém" snímáním řeči může vhodný mikrofon zaručit velmi dobrý poměr signál/šum (řeč/zvuk pozadí). Pokud je odstup mikrofonu od zdroje řeči konstantní a velmi malý, pak lze zanedbat vliv akustiky místnosti (okolí).



Obr. 2.2: Blokové schéma obvodu analogového předzpracování

Takto získaný signál je v rozsahu několika milivoltů a musí být zesílen pokud možno bez šumu a s lineární kmitočtovou závislostí v pracovním pásmu následujících stupňů zpracování. Přitom je nutné dbát na krátká spojovací vedení (požití koaxiálních kabelů). Je rozumné realizovat zesilovací jednotku dvěma stupni: předzesilovačem v přímé blízkosti mikrofonu a dalším stupněm před *A/D* převodníkem.

Analogový řečový signál $s(t)$ je nutné omezit dolní propusti s ohledem na následné vzorkování. Podle známého vzorkovacího teorému musí být vzorkovací kmitočet f_{vz} více než dvojnásobně vyšší než je kmitočet dolní propusti f_{dp} ($2 \cdot f_{vz} > f_{dp}$).

Střední úroveň řečového signálu se při normální řeči mění obvykle o několik decibelů v časovém rozmezí několika sekund. Změnou poloh mikronu a úst mluvčího lze způsobit opět rozdíl několika decibelů. Mnoho parametrů signálu je závislých na kolísaní hlasitosti. Protože tyto efekty nemají fonetický význam, je žádoucí vyrovnávat celkovou intenzitu řečového signálu hned na počátku zpracování ještě v analogové podobě.

2.2 Analogově číslicový převod

Analogově předzpracovaný řečový signál je nyní digitalizován obvykle vzorkováním na kmitočtu 8-22 kHz a kvantováním s rozlišením 8-16 bitů. Parametry digitalizace jsou buď dány zdrojem resp. přenosovou cestou signálu, nebo si je při snímání řeči volíme sami s ohledem na účel zpracování řečového signálu. Obecně lze říci, že pro rozpoznání obsahu řeči nám postačí nižší kvalita digitalizace (8-12 kHz, 8-10 bitů) , zatímco při rozpoznání mluvčího požadujeme kvalitnější číslicový signál. Nejvyšší kmitočtové a amplitudové rozlišení je vhodné pro rozpoznání emočních stavů mluvčích a diagnostická vyšetření hlasu.

Zatímco charakter vzorkovaného signálu nám předurčuje minimální vzorkovací kmitočet, při volbě maximálního vzorkovacího kmitočtu nejsme nijak omezeni. Naproti tomu při volbě kvantování jsme charakterem vzorkovaného signálu omezeni při stanovení maximálního smysluplného počtu kvantovacích úrovní.

Počet rozlišení úrovní ideálního signálu, který neobsahuje poruchy, je teoreticky neomezený. V praxi se však vždy v určité míře vyskytují v signálu poruchy různého druhu. Přítomnost poruch má za následek omezení počtu rozlišitelných úrovní signálu. Reálný počet rozlišitelných amplitudových úrovní signálu a tím současně maximální počet kvantizačních úrovní m je podle Shannona dán vztahem:

$$m = k\sqrt{1+\frac{P_s}{P_p}}, \tag{2.1}$$

kde P_s je maximální výkon signálu, P_p je střední výkon poruch, k je konstanta typu šumu (pro bílý šum $k = 1$)

Při kvantizačním procesu dochází k určité ztrátě informace vlivem zaokrouhlování okamžitých velikostí signálu. Tato ztráta se nazývá kvantizační zkreslení nebo kvantizační šum. Odstup signálu od kvantizačního šumu je v decibelech pro *B*-bitový převod dán vztahem:

$$SNR = 6B - 7{,}24 . \tag{2.2}$$

Protože dynamický rozsah řečového signálu je asi 60 dB, je pro jeho kvalitní převod (např. účely záznamu) zapotřebí B = 11-12 bitů. Jsou-li hodnoty signálu *s(t)* rozloženy přibližně rovnoměrně v celém intervalu:

$$|s(t)| \le S_{\max} ,$$

nabízí se provést rovnoměrné (uniformní) kvantování celého rozsahu signálu do počtu pásem 2^B, při šířce pásma (kvantizační krok) :

$$\Delta = \frac{2S_{\max}}{2^B} .$$

Rozložení okamžitých hodnot řečových signálů však spíše připomíná exponenciální průběh kolem střední hodnoty[2].

2.3 Preemfáze

Statisticky zjištěné dlouhodobé spektrum řečového signálu ukazuje, že střední část spektra klesá se sklonem 6 dB/oktávu. Podstatná část celkové energie řečového signálu (u některých mluvčích více než polovina) leží v kmitočtovém pod hranicí nad 300 Hz, ačkoli užitečné informace jsou téměř kompletně obsaženy v pásmu nad 300 Hz. Vezmeme-li navíc úvahu, že kvantizační šum vykazuje rovnoměrné spektrum, je jeho negativní vliv podstatné větší na energeticky slabší, ale důležitější vyšší složky spektra řečového signálu. U znělých zvuků navíc obvykle první formant energeticky silně převyšuje ostatní formanty.

Uvedené efekty lze částečně zmírnit filtrací řečového signálu číslicovým filtrem s charakteristickou horní propustí:

$$H(z) = 1 - \lambda z^{-1} . \tag{2.3}$$

Filtrace prováděna před vážením segmentu na zdůraznění vyšších kmitočtů se nazývá preemfáze (preemphasis) a časové oblasti naplňuje vztah:

$$s''(n) = s'[n] - \lambda s'[n-1] . \tag{2.4}$$

Koeficient preemfáze leží obvykle v intervalu λ (0,9-1,0). Někdy je vhodné použít adaptivní preemfázi, při které se λ mění s časem podle podílu prvních dvou autokorelačních koeficientů [4]:

$$\lambda = \frac{R[1]}{R[0]}. \tag{2.5}$$

2.4 Segmentace pomocí oken

Protože je rychlost fyziologické funkce orgánu člověka omezena, a proto lze nalézt v řečovém signálu, krátké úseky, v nichž se vlastnosti řeči mění dostatečně pomalu, je téměř výhradně zpracováván metodami tzv. krátkodobé Tyto metody vycházejí z kvazistacionární podstaty řečového signálu, tj. možnosti přijetí předpokladu, že vlastnosti signálu se v čase mění „pomalu“ Signál je za tím účelem rozdělen na ekvidistantní časové úseky – segmenty (frame) o délce N vzorků a každý segment je potom popsán vektorem příznaků.

Délka segmentu musí být na jedné straně dostatečně malá, aby bylo možné naměřené parametry uvnitř segmentu aproximovat konstantními hodnotami a na druhé straně dostatečně velká , aby bylo zaručeno, že požadované parametry budou bezchybně změřeny.

Oba protichůdné požadavky jsou vcelku splněny pro úseky řeči dlouhé 10 ms až 25 ms, což souvisí se změnami nastavení lidského hlasivkového ústrojí, které probíhají v nejkratším intervalu 10 ms až 25 ms. U takových segmentů platí přibližně Gaussovo rozložení hustoty pravděpodobnosti okamžité velikosti řečového signálu. Celé slovo je rozděleno celkem na J segmentů, přičemž všechny segmenty mají stejnou délku odpovídající N vzorkům. Přitom se dva sousední segmenty mohou překrývat. Částečným překrýváním segmentů se dosáhne většího vyhlazení časových průběhů parametrů signálu, ale zpomalí se časový posun a částečně se zvýší výpočetní nároky.

Řečový segment $s[n]$ o N vzorcích může být vytvořen z řečového signálu a přidělit jim určitou váhu. Váhová funkce w(n) určuje typ okna. Nejčastěji používanými typy oken při zpracování řečového signálu jsou:

pravoúhlé	$w[n] = 1$	pro $n = 1,2,N$
	$w[n] = 0$	pro ostatní n
Hammingovo	$w[n] = 0,54 - 0,46\cos(2\pi n/N)$	pro $n = 1,2,N$
	$w[n] = 1$	pro ostatní n

V obou případech je N délka okna a tím současně také délka vybraného segmentu řeči vyjádřena v počtu vzorků. Časový průběh definovaných typů oken a jejich aplikace na řečový signál je znázorněn na obr. 3.4. Přestože pravoúhlé okno je jednodušší, často se upřednostňuje použití Hammingova okna, vzhledem k tomu, že potlačuje vzorky na okrajích segmentů, čímž se zvyšuje stabilita některých výpočtů. Zvolené okno se pohybuje po časové ose krokem N vzorků v případě, že segmenty na sebe navazují nebo s krokem menším než N vzorků, pokud se segmenty překrývají.

Obě okna v podstatně představují filtr typu dolní propusti. Spektrum vybraného segmentu získané Fourierovou transformací reprezentuje výsledek konvoluce skutečného spektra daného úseku řečového signálu se spektrem použité okénkové funkce. V takovém případě je důležité znát, jak vypadá spektrum okénkové funkce a jaké závěry o skutečném spektru řeči můžeme vyvodit z výsledné konvoluce.

Násobení řečového signálu pravoúhlým oknem vede ke dvěma nežádoucím efektům – rozmazání a rozptylu spektra. Oba efekty souvisejí s tím, že spektrum pravoúhlého okna tvořeno jedním hlavním lalokem a větším množstvím vedlejších laloků. Konvolucí spektra okna se spektrem signálu se jediná spektrální čára ve spektru signálu rozšíří (rozmaže) na tvar hlavního laloku. Šířka hlavního laloku tak určuje kmitočtové rozlišení *DFT* a pro délku okna $N \cdot T_{vz}$ je dána vztahem:

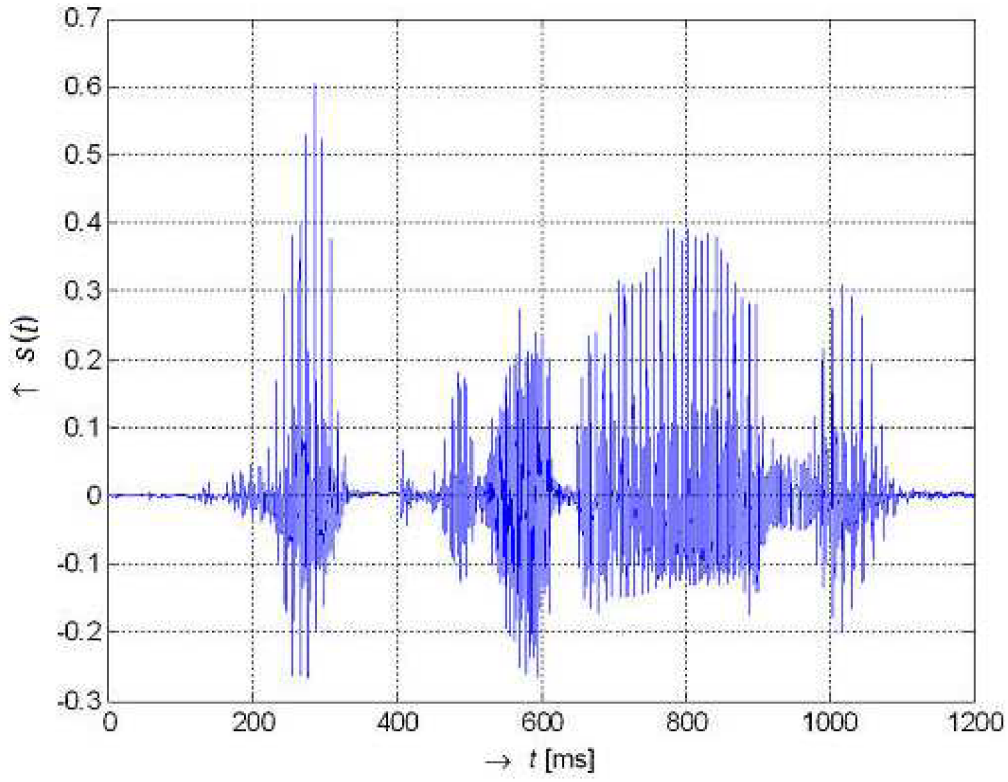
$$\mathrm{DFT} = \frac{2}{N \cdot T_{vz}}, \tag{2.5}$$

kde T_{vz} je vzorkovací perioda. Znamená to, že chceme-li dosáhnout velkého spektrálního rozlišení (při stejném vzorkování), musíme volit N co největší. Avšak při dlouhém analyzovaném úseku budou rychlé spektrální změny průměrovány a nemohou být detekovány. Druhým nežádoucím efektem (rozptyl spektra) , který je způsoben vedlejšími laloky ve spektru okna, je skutečnost, že ve spektru nevzorkovaného řečového signálu se objeví nové spektrální čáry vně hlavního laloku. Tento efekt nelze potlačit změnou délky okna, můžeme ho ovlivnit pouze tvarem okna. U pravoúhlého okna je výška prvního vedlejšího laloku 13 dB maximem hlavního laloku.

U řečového signálu (zejména v jeho znělých úsecích) se vyskytují rozdíly mezi nejsilnějšími a nejslabšími kmitočtovými komponenty více než 40 dB. Použitím pravoúhlého okna nemohou být slabé komponenty ve spektru signálu vůbec postihnuty. Řešením tohoto problému je použití jiného vhodnějšího typu okna, obvykle Hammingova. Toto okno má sice ve spektru zhruba dvojnásobně široký hlavní lalok, ovšem útlum vedlejších laloků 43 dB je podstatně lepší [4].

3. Parametry řečového signálu

Vezměme například řečový signál, který není zabarven žádnou emocí a je tedy neutrálního charakteru. Tato příkladová promluva bude rozdělena na 71 segmentů o délce 40 ms a každý segment vážen Hammingovým oknem. Následně jsou pro každý segment počítány všechny vypsané parametry, které také nazýváme příznaky [2].



Obr. 3.1: Časový průběh mužské řečové promluvy

3.1 Střední počet průchodů signálu nulovou rovinou (ZCR)

Diky tomuto parametru můžeme rozlišit, zda je úsek promluvy řečí nebo šumem. Tento parametr je počítán pro každý rámec řeči. Předpokládá se, že před rozdělením do rámců je signál předzpracován metodami uvedenými v kapitole 3. Rámec řeči je do označen x(n), kde $0 \leq n \leq I_{seg} - 1$:

$$ZCR = \sum_{n=0}^{I_{seg}} \left| sign(x[n] - sign(x[n-1]) \right| \tag{3.1}$$

Kde znaménková funkce “sign” je definována [2]:

$$sign(x[n]) \quad \begin{cases} +1 \ \ pro \ x[n] \geq 0 \\ 0 \ \ pro \ x[n] = 0 \\ -1 \ \ pro \ x[n] \leq 0 \end{cases}.$$

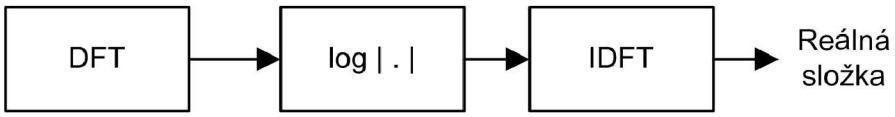
3.2 Krátkodobá energie

Pomocí tohoto parametru lze rozlišit úseky ticha (nízká energie) od úseků řeči (vysoká energie). Při měření krátkodobé energie je vhodnější volit kratší délku mikrosegmentů 20 ms až 40 ms. Hodnoty funkce poskytují pro každý segment informaci o průměrné hodnotě energie v mikrosegmentu. Díky velkému dynamickému rozsahu energie (několik řádů) používáme její logaritmus [2]:

$$E = \frac{1}{I_{seg}} \sum_{n=0}^{I_{seg}-1} |x(n)|^2 \ . \tag{3.2}$$

3.3 Kesptrum

Řeč je dána kovolucí buzení impulzní charakteristiky filtru. Pokud jsou dva signály konvoluovány, je obtížné je získat zpět (provést dekonvoluci). Můžeme se o to pokusit tak, že na určitém místě zavedeme do transformací nelinearitu, která dokáže převést součin na součet. Jednotlivé komponenty součtu pak již lze oddělit.



Obr. 3.2: Blokové scháma kepspestrální analýzy

Komplexní kepstrum buzení se sestává z pulsů, které se objevují v intervalech odpovídajících periodě základního hlasivkového tónu. Protože kompletní kepstrum impulsní odezvy hlasového traktu je soustředěno kolem $n = 0$, kompletním kepstrem buzení jsou pulzy v intervalech úměrných periodě základního hlasivkového tónu f_0. Kesptrální hodnoty f_0 reprezentující hlasivkový trakt lze extrahovat z úplného kepstra pomocí lineárního systému, ve kterém jsou složky kepstra pro malé hodnoty $|n|$ násobeny hodnotou jedna a ostatní nulou. Postup kepspetrální anlýzy je na obrázku 4.2. Předzpracovaný signál je přiveden na vstup bloku *DFT*, jeho výstup přichází na blok $\log|.|$.Tento výsledek podrobíme *IDFT*. Výsledné kepstrum je reálná složka tohoto bloku.

První koeficient kepstra představuje energii signálu. Koeficienty s nízkým pořadím (dolní kvefrence) popisují pomalé změny ve spektru signálu, tzn. formantovou strukturu a tím i charakteristiku hlasového traktu. U znělých úseků řeči se v kepstru vyskytuje výrazná špička, která svou polohu určuje základní tón řeči.

Odhad spektrální hustoty výkonu pomocí přímé *DFT*:

$$c[n] = DFT^{-1}\{\ln\left|DFT\left[s[n]\right]^2\right|\}. \tag{3.3}$$

Diskrétní Fourierova transformace bývá samozřejmě implementována pomocí rychlého algoritmu *FFT* (Fast Fourier Transform). Tedy jestli má nezávislá proměnná n kvefrence. Kepstrum vzniklo přesmyčkou ze slova spektrum.

Pro získání základního tónu řeči (*ZTŘ*) z kepstra použijeme znělé segmenty signálu. Na kepstrum je aplikován filtr, který odstraní z kepstra charakteristiku hlasového traktu, tzn. 40 vzorků je nulováno. Základní tón je odpovídající kvefrence maximální hodnoty kepstra tohoto segmentu [2].

3.4 Základní tón řeči pomocí autokorelační funkce

Autokorelační funkce je definována pro rámec x(n) jako:

$$R[m] = \sum_{n=0}^{N-1-m}\left|x(n) - x[n+m]\right|. \tag{3.4}$$

Maximum této funkce hledáme pro $i \in [L_{min}, L_{max}]$, kde L_{min} je minimální povolená hodnota periody základního tónu ve vzorcích a L_{max} je maximum. Index maximální hodnoty označíme i_{max} a hodnotu R_{max}. Pokud je:

$$\frac{R_{\max}}{R[0]} > \alpha, \tag{3.5}$$

kde α se volí přibližně 0,3, prohlásíme rámec za znělý a udává periodu základního tónu řeči (*ZTŘ*). V opačném případě prohlásíme rámec za neznělý [2].

3.5 Jitter

Frekvence základního tónu ve skutečnosti není konstantní (hlasivkové pulsy nejsou ryze periodické). V delších řečových úsecích s totiž projeví vliv intonace promluvy. Délka periody i amplituda jednotlivých pulzů základního hlasivkového tónu se mírně liší i v rámci krátkého signálu (obvykle již se z periody na periodu). Takové mírně kolísání délky základní periody se nazývá Jiter a je nezávislé na duševním stavu mluvčího.

Jiter je definován jako střední rozdíl délek sousedních period, dělených střední délkou periody. V tomto případě je ZTŘ počítán z kepstra [2]:

$$Jitter = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_0^{(i)} - T_0^{(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N-1}T_0^{(i)}}. \tag{3.6}$$

$T_0^{(i)}$, i = 1,2,…N je základní perioda a N je rovno počtu period

3.6 Shimmer

Kolísání velikosti impulzů (shimmer). Je definován [2]:

$$Shimmer = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A^{(i)} - A^{(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N-1}A^{(i)}}. \tag{3.7}$$

$A^{(i)}$, i = 1,2,…N je rozkmit a N je rovno počtu impulsů.

3.7 Spektrum

Periodické signály, jejichž analýza a syntéza je běžná (akustické aplikace), lze analyzovat pomocí Fourierových řad:

$$s(t) = \sum_{k=-\infty}^{\infty} c_k e^{jk\Omega t},\ \Omega = \frac{2\pi}{T_s}.$$

Fourierova transformace má za určitých předpokladů k Fourierově řadě bezprostřední vztah. Mějme vzorkování zvoleno tak, aby perioda signálu byla celistvým násobkem vzorkovací periody T, tedy T_s=N·T (fázovým závěsem, dodatečným vzorkováním). Pokud nebude podmínka vzorkování splněna, nebudou koeficienty odpovídat skutečnosti a výsledné poskládání nebude tvořit přesnou funkci. Funkce s(t) je frekvenčně omezená s horní mezní frekvencí a vzorkování splňuje vzorkovací teorén $\omega_{\max} < \frac{\pi}{T} = \frac{N\Omega}{2}$, pak $c_k = 0$ pro $|k| > \frac{N}{2}$

Spektrum počítáme z konečného počtu (N) vzorků. Původně nekonečný signál je vynásoben oknem o N vzorcích:

$$s(\omega) = STFT\{s(nT)\} = \sum_{n=0}^{N-1} s[nT]e^{-jknT}, \tag{3.8}$$

Pomocí DFT získáme diskrétní verzi signálu :

$$s(nT) = \frac{1}{N}k\sum_{n=0}^{N-1} S(nT)e^{jk\Omega nT} = \frac{1}{N}\sum_{k=0}^{N-1} S(nT)e^{jk\Omega\frac{2\pi}{N}}. \tag{3.9}$$

Pokud zaměníme spojitý signál $s(t)$ za diskrétní $s[nT]$, dostaneme koeficient Fourierovy transformace (jednotlivé frekvence) [2]:

$$c_k = \frac{1}{N} k \sum_{n=0}^{N-1} S(nT) e^{-jk\Omega nT} . \tag{3.10}$$

3.8 NHR

NHR neboli Noise to Harmonic ration poměr hodnot neharmonické spektrální energie ve frekvenčním pásmu 1500-4500 Hz, vzhledem k neharmonické spektrální energii ve frekvenčním pásmu 70-4500 Hz [2].

3.9 Popis tvaru spektra

3.9.1 Spektrální centroid

Spektrální centroid je těžiště nebo také centrum spektra. Používáme je v oblasti digitálního zpracování signálu, aby charakterizoval audio spektrum. Naznačuje, kde leží největší část spektra. Spektrální centroid je počítán jako vážený průměr z frekvencí, přítomných v signálu, které jsou vážené odpovídající amplitudou [2]:

$$\mu = \frac{\sum_{n=0}^{N-1} f[n] \cdot x[n]}{\sum_{n=0}^{N-1} x[n]} . \tag{3.11}$$

3.9.2 Spektrální rozptyl

Spektrální rozptyl je rozptyl spektra okolo její střední hodnoty [2]:

$$\sigma = \frac{\sum_{n=0}^{N-1} (f[n] - \mu)^2 \cdot x[n]}{\sum_{n=0}^{N-1} x[n]} . \tag{3.12}$$

3.9.3 Spektrální šikmost

Šikmost udává hodnotu nesymetrie rozdělení okolo její střední hodnoty. Je počítána z momentu třetího řádu [2]:

$$m_3 = \frac{\sum_{n=0}^{N-1} (f[n] - \mu)^3 \cdot x[n]}{\sum_{n=0}^{N-1} x[n]}. \tag{3.13}$$

z toho vyplývá, že šikmost je

$$\gamma_{\check{s}} = \frac{m_3}{\sigma^3} \tag{3.14}$$

Míra asymetrie rozložení tedy je:

$\gamma_{\check{s}} = 0$, odpovídá symetrickému rozložení,
$\gamma_{\check{s}} < 0$, odpovídá rozložení více energie na pravé straně,
$\gamma_{\check{s}} > 0$, odpovídá rozložení více energie na levé straně .

3.9.4 Spektrální špičatost

Udává hodnotu špičatosti rozložení okolo jeho střední hodnoty. Je počítána z momentu 4. řádu [2]:

$$m_4 = \frac{\sum_{n=0}^{N-1} (f[n] - \mu)^4 \cdot x[n]}{\sum_{n=0}^{N-1} x[n]}. \tag{3.15}$$

Špičatost tedy je:

$$\gamma_1 = \frac{m_4}{\sigma^4} - 3 . \tag{3.16}$$

Míra špičatosti rozložení tedy je:

$\gamma_1 = 0$, odpovídá normálnímu rozložení,
$\gamma_1 < 0$, odpovídá ploššímu rozložení,
$\gamma_1 > 0$, odpovídá špičatějšímu rozložení.

3.9.5 Spektrální sklon

Udává hodnotu snižující se spektrální amplitudy. Je počítán jako lineární regrese spektrální amplitudy, což představuje aproximaci daných hodnot polynomem prvního řádu (přímkou) metodou nejmenší čtverců [2]:

$$sklon = \frac{1}{\sum_{n=0}^{N-1} x[n]} \cdot \frac{\sum_{n=0}^{N-1} f[n] \cdot x[n] - \sum_{n=0}^{N-1} f[n] \cdot \sum_{n=0}^{N-1} x[n]}{\sum_{n=0}^{N-1} f^2[n] - \left(\sum_{n=0}^{N-1} f^2[n] \right)^2}. \tag{3.17}$$

3.9.6 Spektrální plochost

Spektrální plochost neboli Flatness se používá v oblasti digitálního zpracování signálu za účelem charakterizace audio spektra. Vysoká hodnota spektrální plochosti naznačuje, že spektrum má podobné množství energie ve všech spektrálních pásmech – to znamená bílý šum, a graf spektra, by vypadal relativně ploše a hladce. Nízká hodnota spektrální plochosti naznačuje, že spektrum je soustředěno v relativně malém počtu pásem – to znamená směs sinusových vln a že spektrum vypadá "špičatě". Spektrální plochost se vypočítá vydělením geometrického průměru výkonového spektra střední hodnotou výkonného spektra [2]:

$$Flatness = \frac{\sqrt{\prod_{n=0}^{N-1} x[n]}}{\frac{1}{N} \sum_{n=0}^{N-1} x[n]}. \tag{3.18}$$

3.10 Mel-frekvenční kepstrální koeficienty - MFCC

Výpočet kepstra v kapitole 3.3 příliš neodpovídá lidskému slyšení:

- *DFT* má všude stejné frekvenční rozlišení.
- Lidské ucho má na nízkých frekvencích větší rozlišení než na vysokých.
- Pro rozpoznávače řeči chceme přiblížit kepstrum slyšení.

U *MFCC* postupujeme tak, že na frekvenční osu rozmístíme nelineárně filtry.Musíme energii na jejich výstupu, a tuto použijeme místo DFT při výpočtu kepstra.

Při konstrukci filtru musíme nelineárně upravit frekvenční osu a na upravené ose pak filtry rozmístit rovnoměrně. Používaná nelineární úprava využívá převodu Hertzů na Mely:

$$F_{Mel} = 2959 \cdot \log_{10}\left(1 + \frac{F_{Hz}}{700}\right). \tag{3.19}$$

Lineární rozmístění filtrů na Mel-ové ose má za následek nelineární rozmístění na standardní kmitočtové ose v Hz (Obr. 5.19).
Při výpočet energií z jednotlivých filtrů (frekvenčních pásem) bychom mohli skutečně zkonstruovat banku filtrů, vstupní signál filtrovat v časové oblasti a počítat energii pomocí $\sum_n s_i^2(n)$. Toto by však bylo příliš. složité, proto využijeme DFT, umocníme, vynásobíme trojúhelníkovým oknem a sečteme. Tento postup je mj. použit ve standardním toolkitu pro rozpoznávání řeči HTK Hidden Markov Model ToolKit z University of Cambridge.

Zbývá provést zpětnou FT logaritmu těchto energií. Zpětnou FT můžeme realizovat pomocí diskrétní cosinové transformace (DCT), která nahrazuje inverzní FT (bez odvození: využíváme symetrie spektra a toho, že výsledek musí vyjít reálný):

$$c_{mf}(n) = \sum_{i=1}^{K} \log m_k \cos\left[n(k-0{,}5)\frac{\pi}{K} \right]. \tag{3.20}$$

Výsledkem jsou mel-frekvenční koeficienty (MFCC)

4. Emoce

Jen asi 10 % všech informací z řečového signálu nám dává informaci o emočním stavu mluvčího.

Při rozpoznání řeči se lze soustředit na různé aspekty. Jedním z cílů bylo vytvořit syntetizovanou řeč co nejpřirozenější, dalším cílem bylo rozpoznat citový obsah z řeči. Řečový signál zbarvený emocemi nám dává komplexnější pohled na řečníka. Na emotivní stav mluvčího reagují posluchači a přizpůsobují své chování podle druhu emoce, kterou mluvčí vyjadřuje. Například smutným lidem ukazujeme empatie, rozhněvaných se bojíme. Pro určení emočního stavu mluvčího na základě prozodických vlastností a kvality hlasu musíme roztřídit zvukové rysy v řeči a přiřadit je k náležejícím emocím.

Nalezení vhodných akustických vlastností k jednotlivým emocím není příliš jednoduché, také proto si výsledky občas odporují. Je těžké definovat, které příznaky se vztahují k emotivní řeči.

Nejjednodušší přístup k popisu emocí je použití kategorie používané v běžném hovorovém jazyce. Toto rozdělení umožňuje různé způsoby dělení kategorií, které mohou být použity pro popis emočního stavu a emocí. Podle emočního výzkumu se emoce rozdělují do dvou kategorii: primární a sekundární.

4.1 Primární emoce

Kategorie obsahuje takové emoce, které jsou „čisté“ a „jednoduché“. Tyto emoce mají jen několik forem, které jsou od sebe kvalitativně odlišné. Každá forma má příznaky, kterými se od sebe od ostatních odlišuje. Seznam základních emocí je jenom taková formální dohoda – strach, vztek, štěstí / radost, smutek a nuda/. Občas sem lze zařadit překvapení, hněv i pohrdání.

4.2 Sekundární emoce

Tato kategorie obsahuje emoční stavy, které jsou odvozeny z emocí primárních jejich smícháním. Tyto odvozené emoce pokrývají velký rozsah emočních stavů, avšak málo z nich mohlo být považováno za emoce základní. Mluvíme zde například o pocitech: žal, zalíbení / něžnost, sarkasmus / ironie, překvapení / údiv, nenávist / odpor.

5. Klasifikátor

Pro klasifikaci jsme zvolili klasifikátor K-NN

5.1 K-NN klasifikátor.

Klasifikace KNN neboli nejbližších sousedů spadá mezi neparametrické metody klasifikace.K-NN algoritmy pracují na principu, že přiděluje studovaný objekt do skupiny podle jeho pozice v prostoru, ke které má nejmenší vzdálenost. Souřadnice objektu reprezentují hodnoty parametrů, které popisují daný objekt např. v našem případě základní tón apod.. Pozice skupin jsou určeny pomocí tréninkové množiny.

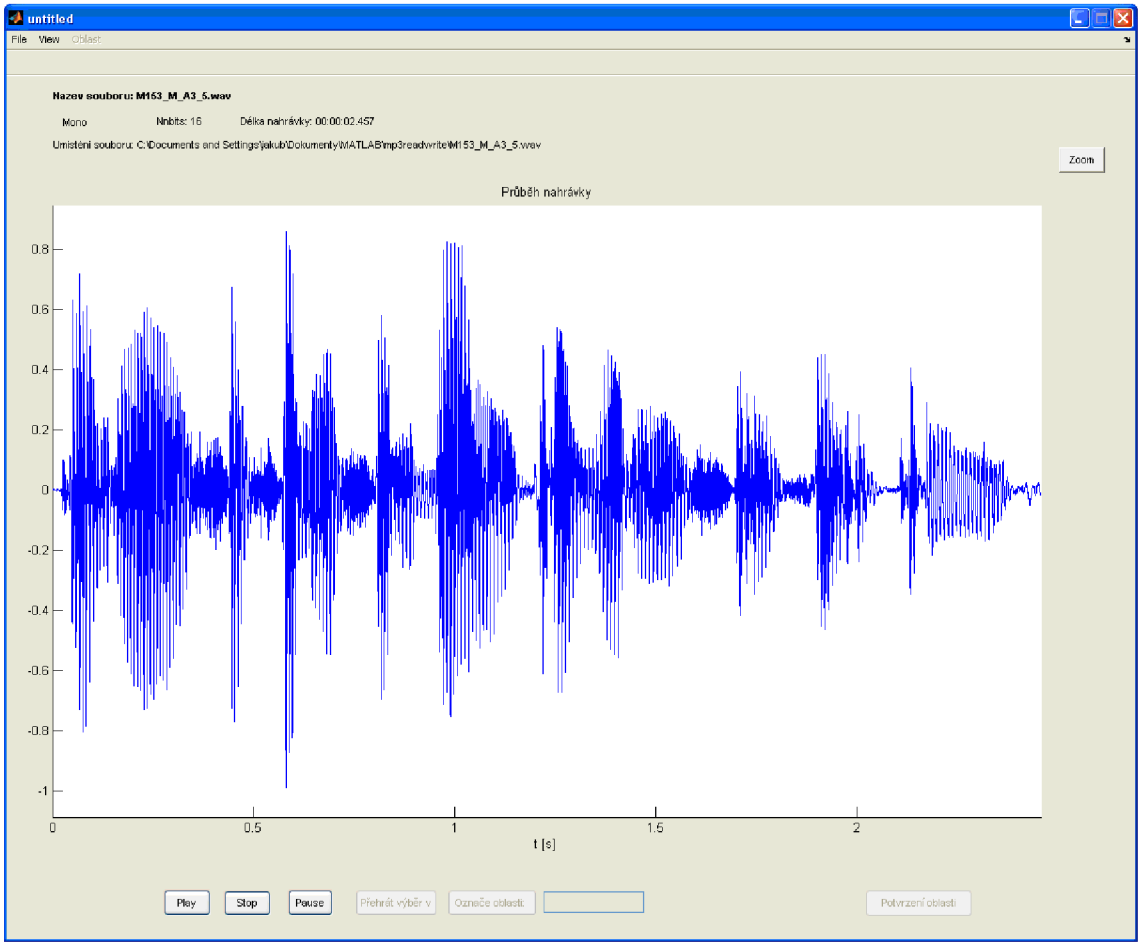
Vzdálenost je určována euklidovsky, tedy vzdálenost vektorů *u* a *v* je:

$$|u - \mathrm{v}| = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2} \ . \tag{5.1}$$

6. Navržený program

Hlavním účelem programu SW podpory analýzy emocionálních stavů je vyznačit oblasti, na které budou později aplikovány výpočty, sloužící pro určení emocí. Některé výpočty, které jsou použity při sestavení programu již byly popsány v kapitole 3.

6.1 Načítání dat



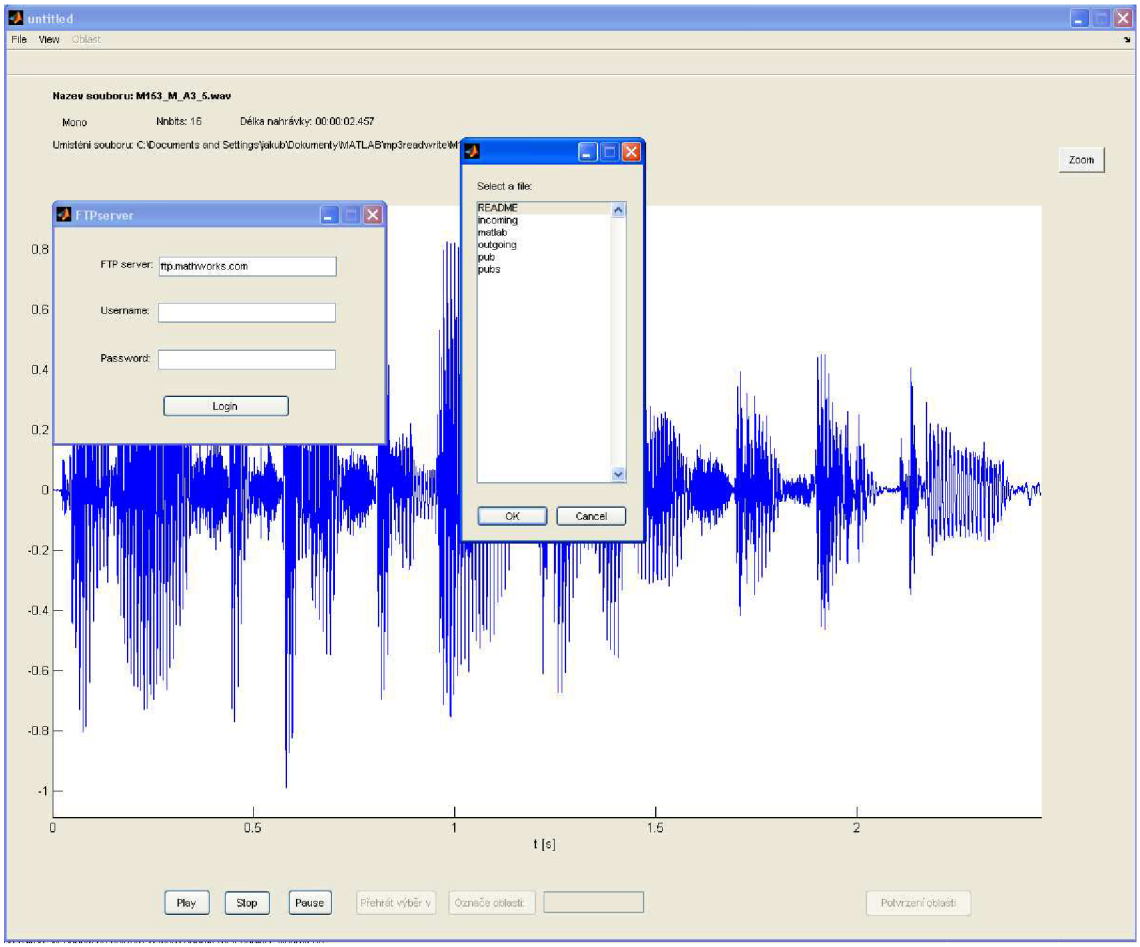
Obr. 6.1: Program po načtení nahrávky

Uživatel si po spuštění programu může vybrat záložce „File“ možnosti „open“. Tím si vybere nahrávku, kterou chce načíst, a to buď ve formátu wav nebo mp3, případně pomocí možnosti „import to FTP server“ si může stáhnout soubor z FTP serveru do pracovního adresáře.

Pro načtení se v případě wav formátu nahrávky používá funkce v Matlabu wavread, jejichž výstupními parametry jsou proměné *y*, *fs*, *nbits*. Proměná *y* reprezentuje vzorky daného signálu. Proměnná *Fs* je vzorkovací frekvence v Hz, která byla použita při zjišťování hodnot vzorků v průběhu signálu, a proměnná *nBits* udává počet bitů použitých při kódování signálu.

Pro načtení mp3 formátu se používá funkce *mp3read*, která není součástí základního balíčku Matlabu, ale byla dodatečně vytvořena pro možnost práce s tímto formátem. Výstupní proměnné jsou přitom stejné jako v případě funkce *wavread*.

Při stahování z FTP serveru, se zobrazí okno, kde uživatel napíše server, ze kterého chce data stahovat. Pro přihlášení může zadat své uživatelské jméno a heslo, pokud je nezadá, bude program považovat přihlášení jako anonymní. Po potvrzení serveru kliknutím na tlačítko „login“, se počítač přihlásí ke zvolenému severu a zobrazí se okno, kde je uveden obsah serveru viz. obr. 6.2. Po zvolení souboru, který chceme stáhnout, se námi označený soubor zkopíruje do počítače. Pro stahovaní dat se využívá matlabovská funkce *mget*.



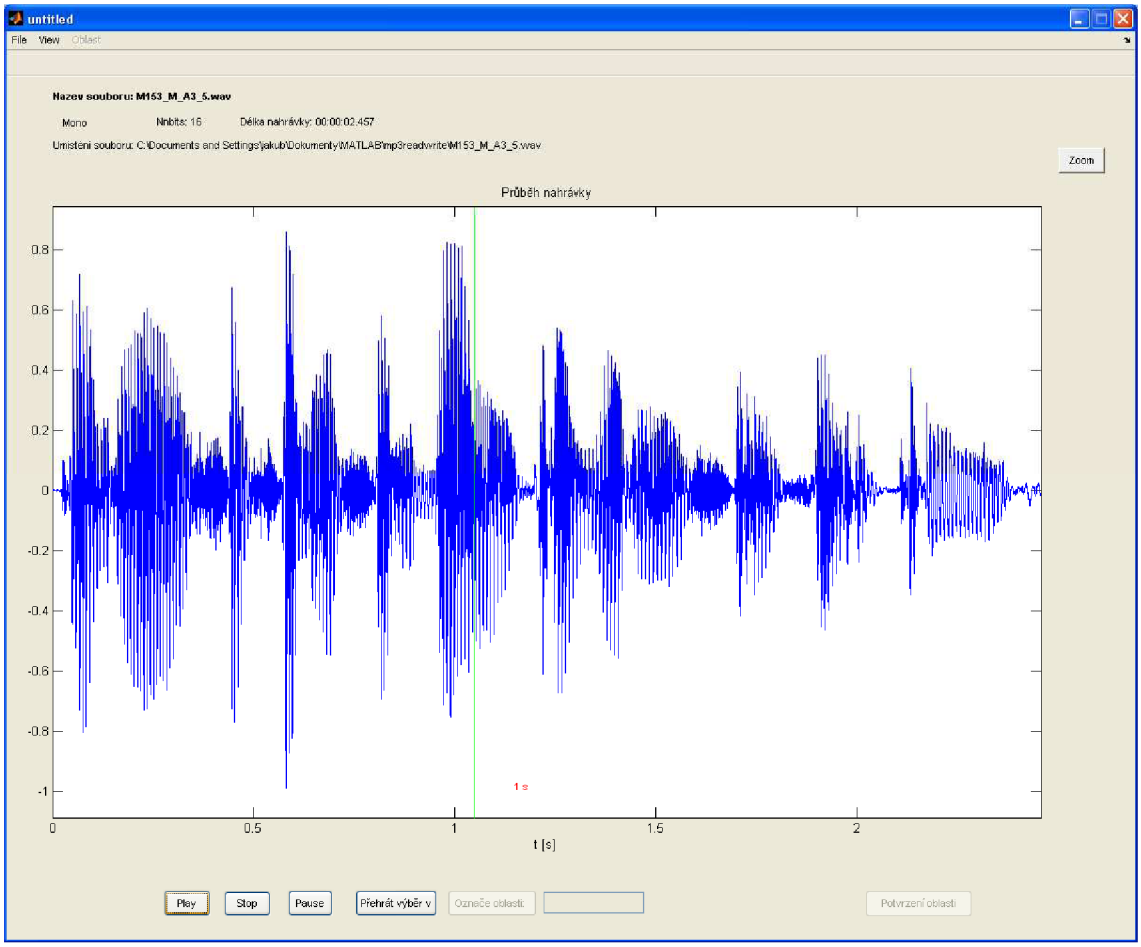
Obr. 6.2: Načítání dat z FTP serveru

Při otevírání nahrávky, se nejdřív otevře okno, ve kterém se vybere soubor. Pro tento krok se využívá funkcí *uiiget*, jejichž výstupem je název souboru a umístění souborů, které se využívají jako specifikace pro to jaká funkce se použije pro načtení a také určuje cestu k souboru. Po zvolení souboru se zobrazí průběh jejího signálu, název souboru, jeho umístění v PC, formát nahrávky počet bitů a délka nahrávky jak je vidět na obr. 6.1. Časová osa je v s a její hodnoty lze vypočítat pomocí vztahu:

$$t_n = \frac{n}{fs}. \tag{6.1}$$

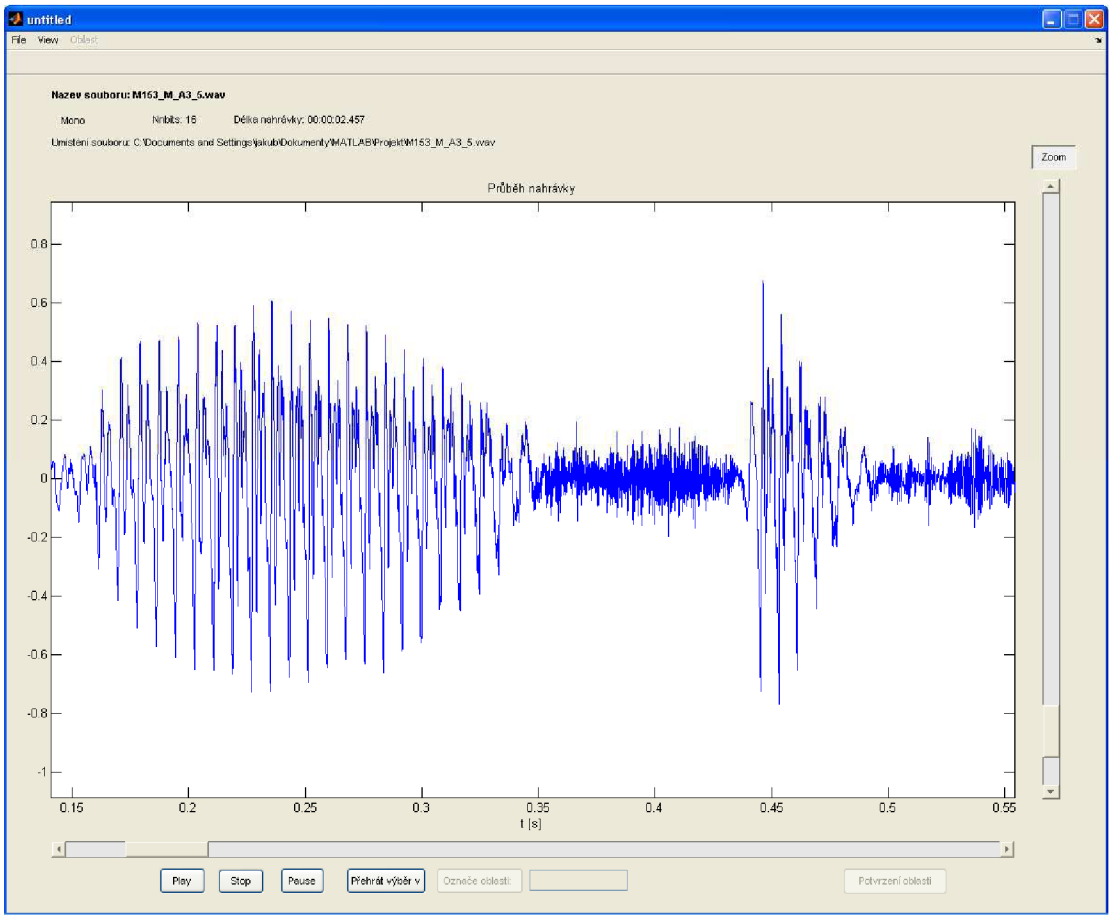
6.2 Práce s nahrávkou

Uživatel si následovně může nahrávku přehrát pomocí tlačítka „Play“, zastavit přehrávání pomocí tlačítka „Stop“, nebo jenom pozastavit tlačítko „Pause“. Pro lepší orientaci, kde se pří přehrávaní nacházíme, se do průběhu zobrazuje ukazatel s danou pozicí obr. 6.3. V případě stereo formátu je možná volba kanálu, který má být přehrán popmenu, které se zobrazí v levém dolní rohu ale jenom při stereo formátu.



Obr. 6.3 Zobrazení aktuální pozice v nahrávce

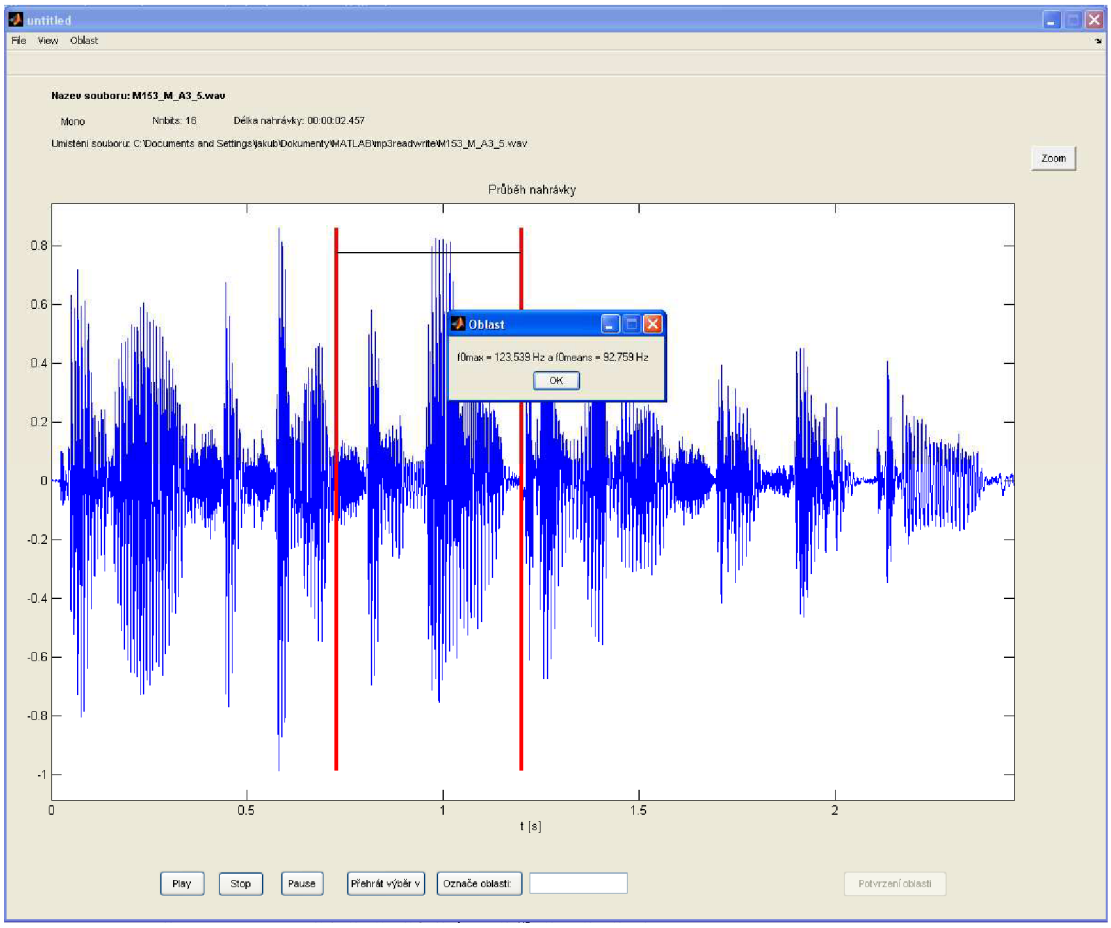
Pro přehrávaní se využívá funkce audiopplayer, která vytvoří objekt sloužící pro tento účel. Dále tím, že si uživatel klikne na tlačítko „Zoom“, si povolí možnost přibližovat nahrávku v časové oblasti, pro zlepšení orientace, a hodnota přiblížení se pak nastavuje pomocí slideru, který se po stisknutí tlačítka objeví od zobrazeného průběhu. Pokud se použije přiblížení, tak se objeví další slider dole pod zobrazeným průběhem, který slouží pro posouvání zobrazené oblasti nahrávky obr 6.4.



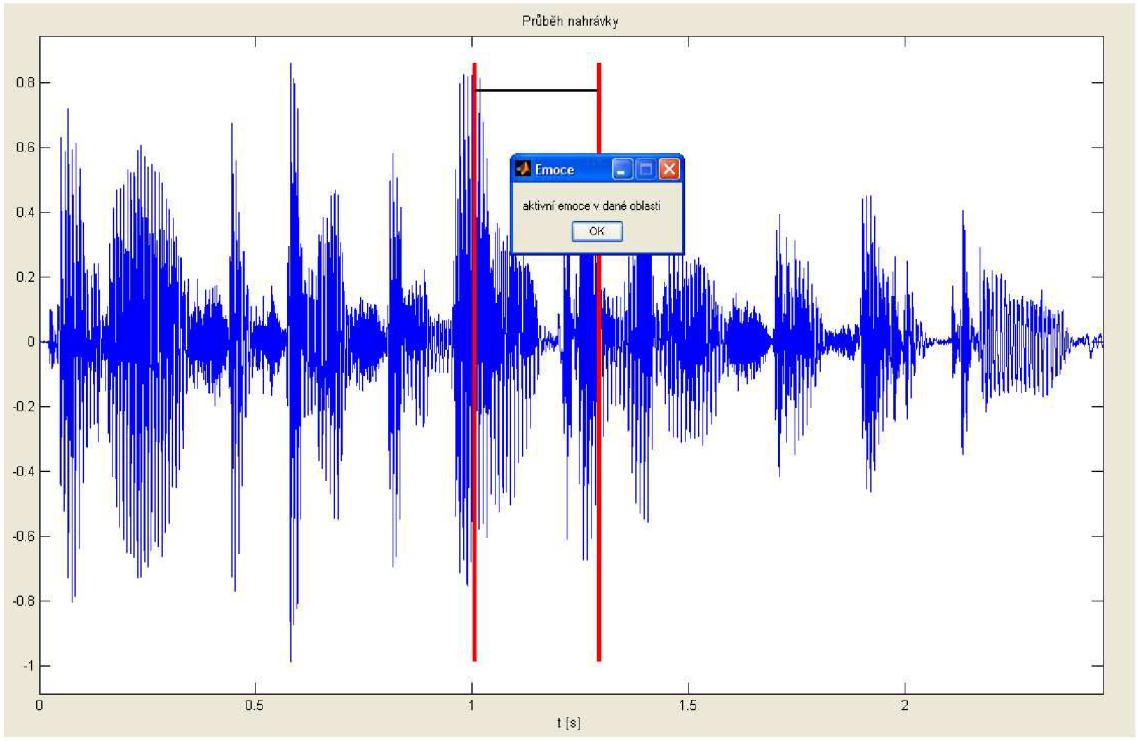
Obr. 6.4 Přiblížení průběhu signálu

Vyznačení oblasti k přehrání si může uživatel vyznačit pomocí myši. Pro tento proces se zase využívá funkce *audioplayer*, u kterého se ale navíc specifikuje počátek a konec přehrávání, které je určeno příslušným číslem vzorků, které určují počátek a konec. Po vyznačení oblasti se objeví vedle tlačítka „Pause“ tlačítko pro přehrávání dané oblasti. Pokud bude mít uživatel zájem může po přehrání tuto oblast pojmenovat. Dále se při označení oblasti povolí nahoře na liště povolí menu „Oblast“, ve které jsou možnosti „hodnoty základního tónu“, která vypočítat zobrazí maximální a střední hodnotu základního tónu obr. 6.5, „tvar spektra“, která nám ukáže hodnoty tohoto parametrů popisující tvar spektra obr. 6.7, nebo „Emoce“ pomocí které se dá zjistit jaká emoce se v daném úseku projevuje.

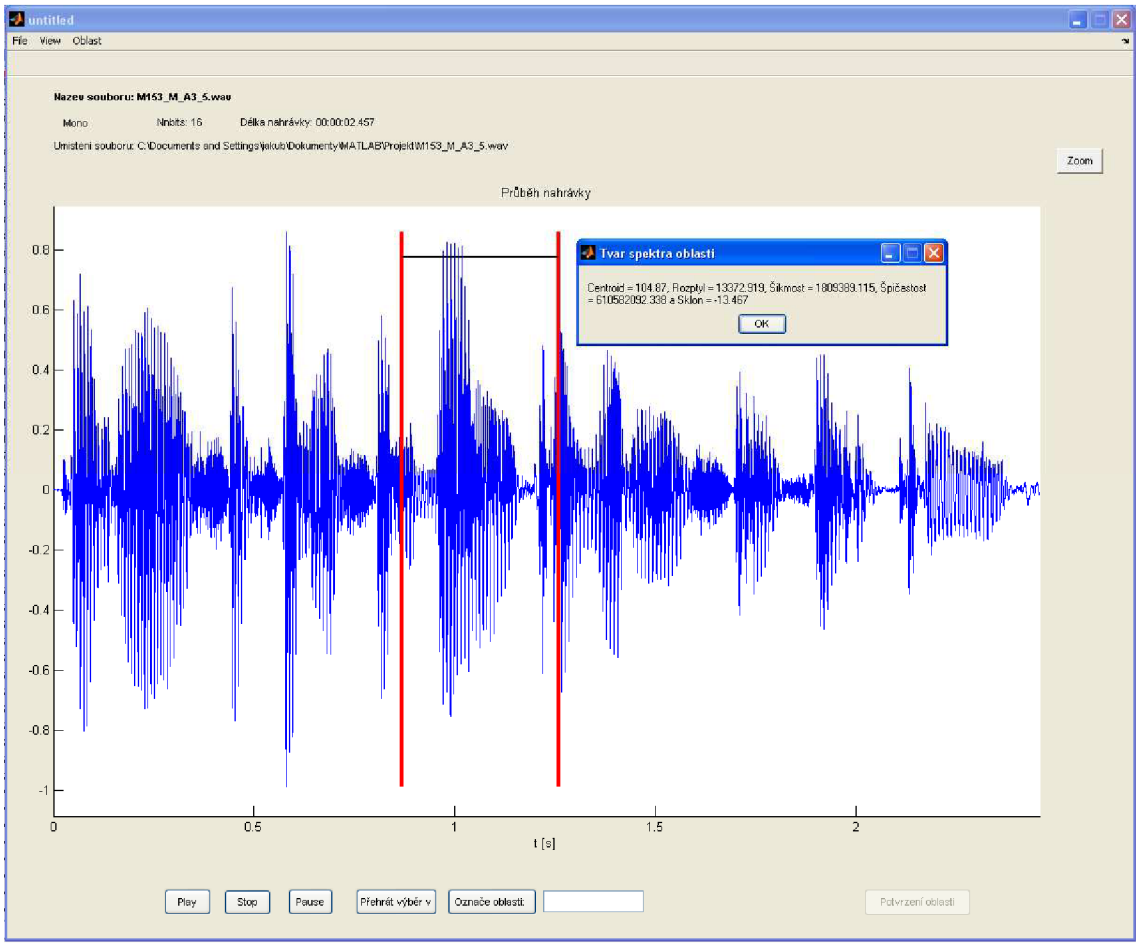
Proces určení emoce v dané oblasti probíhá, tak že se vypočítá její MFCC koeficienty, které se využijí jako hodnoty vektoru pro klasifikátor KNN, který určí typ emoci, pomocí toho že srovná hodnoty vektoru oblasti s hodnotami vektorů emocí, které jsou určeny v tréninkové sekvenci. Výstupem klasifikátoru jestli se v oblasti projevuje pasivní nebo aktivní emoce. Výsledek tohoto procesu se v programu zobrazí v dialogovém ohně obr. 6.6.



Obr. 6.5 Výpočet základního tónu pro danou oblast

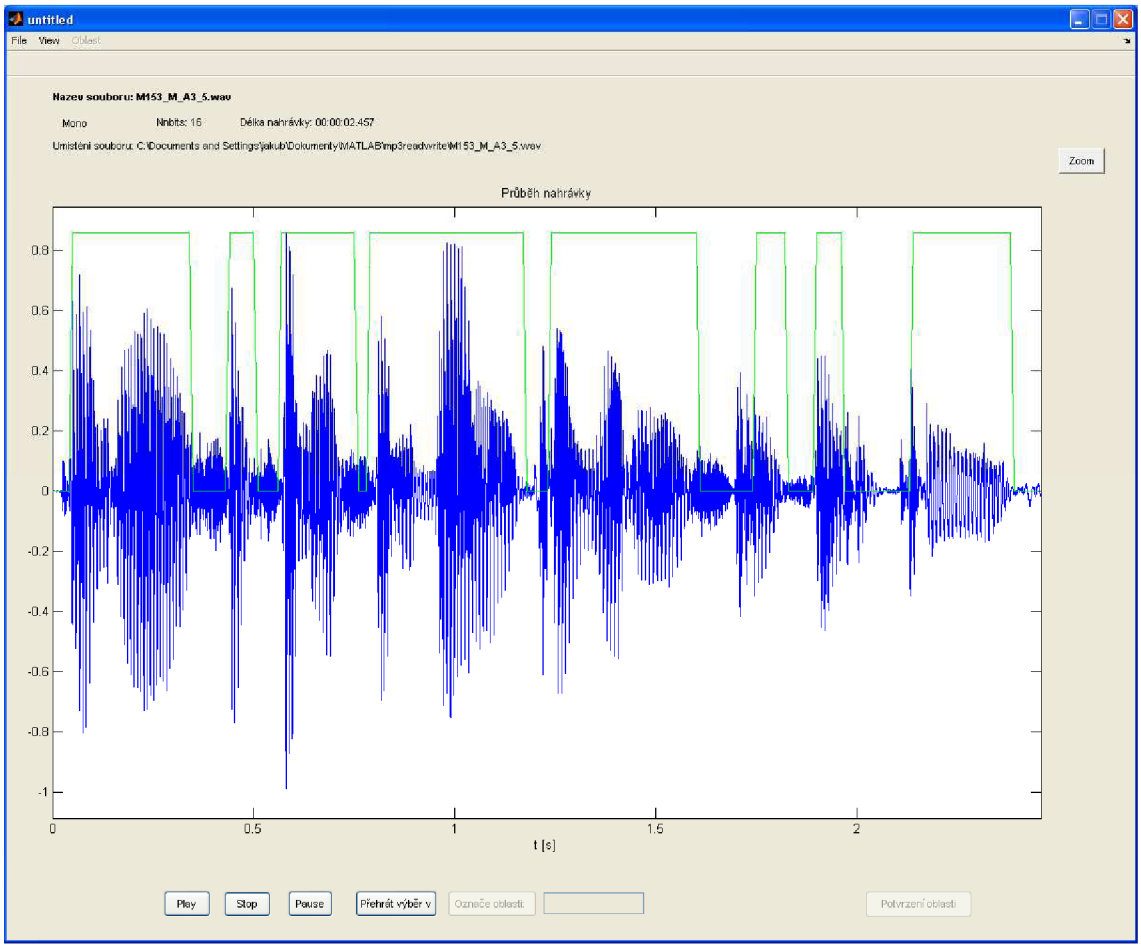


Obr. 6.6 Určení emoce



Obr. 6.7 Zobrazení parametrů popisujících tvar spektra v dané oblasti

Mezi další možnosti, které mohou umožnit lepší orientaci v nahrávce, v níž se může vykytovat emoce, je možnost zobrazení některých parametrů řeči, to znamená energie, základního tónu a znělosti řeči viz obr. 6.8. Výpočet základního tónu se dá upravit, tak že se dají nastavit parametry pro jeho výpočet , aby bylo možné najít nepřesnější výsledek

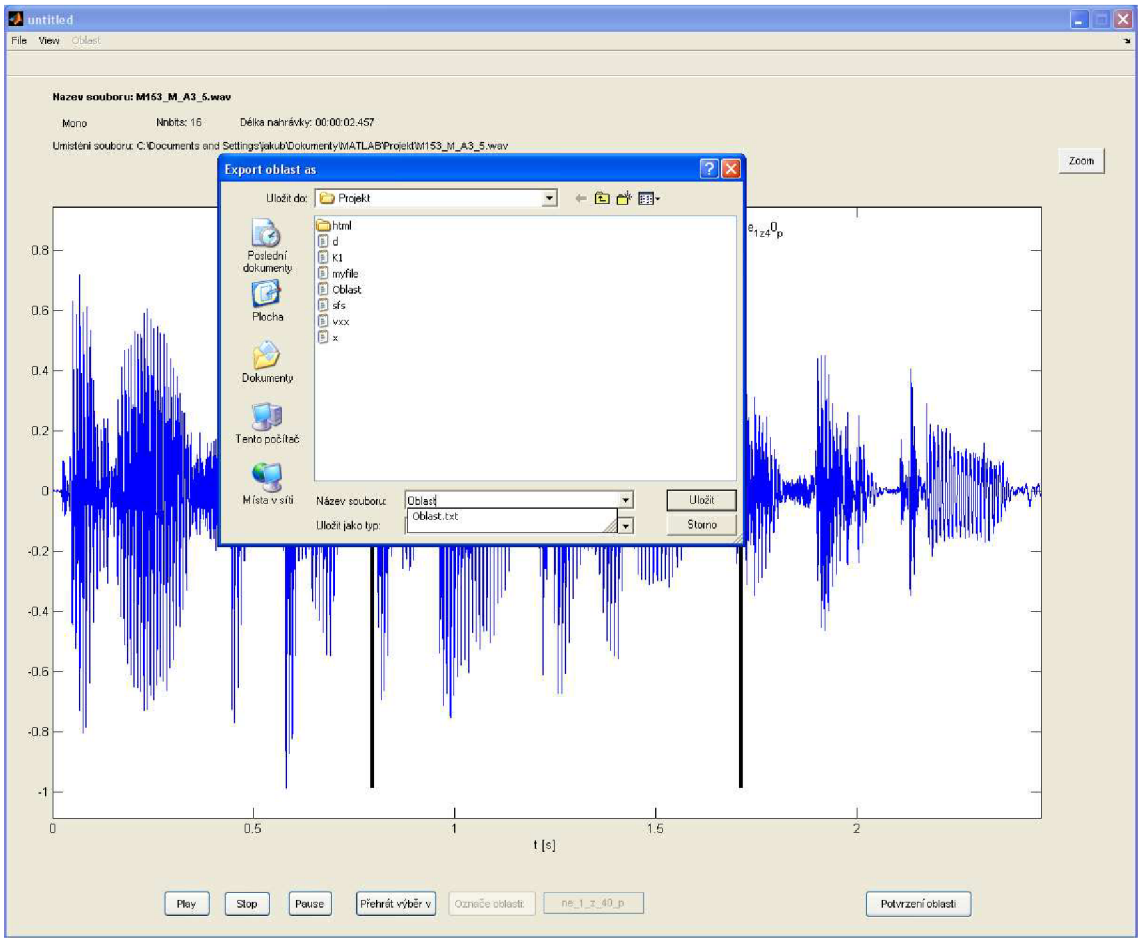


Obr. 6.8: Zobrazení znělých segmentů

.

6.3 Ukládání výsledků

Nakonec po vyznačení všech sektorů, si může uživatel záznam označených oblastí exportovat do počítače v textovém formátu, pro kterou se používá nejdříve funkce *uiiput*, která zobrazí okno pro určení místa, kam chce uživatel uložit soubor a název souboru obr. 6.8, následovně potom se použije funkce *dlmwrite*, která vytvoří textový dokument, obsahující daný záznam, nebo následovně poté si může uložit výsledky na FTP server. Při prvním ukládání na FTP server je postup podobný jako pří stahování z FTP serveru. Nejprve musí uživatel přihlásit na server, kam chce ukládat. Zase pokud uživatel nevyplní při přihlašování uživatelské jméno a heslo, tak program považuje přihlašování jako anonymní. Při dalším ukládání výsledků se už nemusí přihlašovat a výsledky se automaticky pošlou na předtím určený server. Pro ukládání se používá funkce *mput*.



Obr. 6.9 Ukládání výsledku do počítače

7. Závěr

Cílem projektu bylo vytvořit SW nástroj s grafickým rozhraním, který je využitelný pro vytváření multimodálních emocionálních databází.

Potom co jsem provedl rozbor různých parametrů řečových signálů, tak jsem z jejich výsledků zvolil pro rozpoznání příznaků ze signálu, z spektra i jeho z kepstra, sloužící vyhledání emocí v nahrávce. Jako klasifikátor jsem zvolil K-NN.

Výsledný program vyhovuje zadání bakalářské práce a ukazuje možnosti svého využití pro splnění stanoveného cíle, což je vytvoření multimodálních emocionálních databází. Dále je možné pomocí něho vypočítávat parametry řeči, které jsou např.základní tón, energie signálu a další.

Pokud bych se na tomto programu dále pracoval dal by se z něho vytvořít univerzální program sloužící pro vypočet různých parametrů signálu, které by mohli popisovat kvalitu řeči, určení mluvčího, nebo např. určení jestli je v dané nahrávce hlas nebo hudba.

Seznam použité literatury

[1] KRČMOVÁ, M.: Fonetika. Elektronické texty. MU Brno 2003. http://is.muni.cz/do/1499/el/estud/ff/js07/fonetika/materialy/index.html

[2] PSUTKA, J., MULLER, L., MATOUŠEK, J., RADOVÁ, V.: Mluvíme s počítačem česky. ACADEMIA, Praha 2006. ISBN 80-2100-1309-1

[3] SMÉKAL, Z.: Číslicové zpracování signálů. Skripta VUT, Brno 2009 (elektronické texty)

[4] SMÉKAL, Z.: Číslicové zpracování řeči. Skripta VUT, Brno 2009 (elektronické texty)

[5] ZAPLATÍLEK, K., DOŇAR, B.: MATLAB – začínáme se signály. BEN, Praha 2010. ISBN 80-7300-200-0

[6] ZAPLATÍLEK, K., DOŇAR, B.: MATLAB – tvorba uživatelských aplikací. BEN, Praha 2005. ISBN 80-7300-133-0

[7] ZAPLATÍLEK, K., DOŇAR, B.: MATLAB – pro začátečníky. BEN, Praha 2005. ISBN 80-7300-175-6

Seznam použitých veličin, symbolů a zkratek

A/D	Převodník analogového signálu na digitální
f_0	Frekvence základního hlasivkového tónu [Hz]
T_0	Perioda základního hlasivkového tónu [s]
$f_1, f_2, \ldots, f_n$	Formantové frekvence [Hz]
$A_1, A_2, \ldots, A_m$	Antiformátové frekvence [Hz]
$s(t)$	Časově spojitý signál
$s(n)$	Diskrétní signál
$s''(n)$	Digitální signál
$s''(n)$	Digitální signál upravený preemfází
f_{vz}	Vzorkovací frekvence [Hz]
T_{vz}	Vzorkovací perioda [s]
f_{dp}	Mezní frekvence dolní propusti [Hz]
M	Maximální počet kvantizačních úrovní při A/D převodu [-]
P_s	Maximální výkon signálu [W]
P_p	Střední výkon poruch [W]
SNR	Odstup signálu od kvantizačního šumu [-]
Δ	Kvantizační krok [-]
λ	Činitel preemfáze [-]
DFT	Diskrétní Fourierova transformace [Hz]
ZCR	Střední počet průchodů signálu nulovou rovinou [-]
E	Krátkodobá energie [-]
I_{seg}	Celkový počet rámců řeči [-]
$x[n]$	Rámec řeči [-]
$f(n)$	Daná frekvence obsažená ve spektru signálu [-]
$x(n)$	Amplituda dané frekvence [-]
$c[n]$	Spektrální hustota [s]
$w[n]$	Pravoúhlé okénko [-]
FFT	Rychlá Fourireova transformace [Hz]
$R[m]$	Autokorelační funkce [-]
R_{max}	Maximální hodnota Autokorelační funkce [-]
L_{min}	Minimální povolená hodnota periody základního tónu ve vzorcích [s]
L_{max}	Maximální povolená hodnota periody základního tónu ve vzorcích [s]
$Jitter$	Střední rozdíl délek sousedních period [s]
$Shimmer$	Kolísání velikosti impulsů [-]
NHR	Noise to harmonic [-]
A	Velikost amplitudy řečového signálu [-]
ω_{max}	Maximální vlnová rychlost řečového signálu [-]
$S(w)$	Spektrum signálu [-]
$S(nT)$	Diskrétní verze signálu [-]
c_k	Koeficient Fourierovy řady [-]
M	Spektrální centroid [-]
σ	Spektrální rozptyl [-]
m_n	Moment n-tého řádu [-]
$\gamma_š$	Spektrální šikmost [-]
γ_1	Spektrální špičatost [-]
$sklon$	Spektrální sklon [-]

Flatness	Spektrální plochost [-]
ZTŘ	Základní tón řeči
IDFT	Inverzní diskrétní Fourierova transformace [Hz]
$\| u - v \|$	Vzdálenost dvou vektorů [-]
k	Počet vektorů
u, v	Směrové vektory
MFCC	Mel-frekvenční kepstrální koeficienty [-]