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

ROBUSTNÍ DETEKCE KLÍČOVÝCH SLOV V ŘEČOVÉM SIGNÁLU

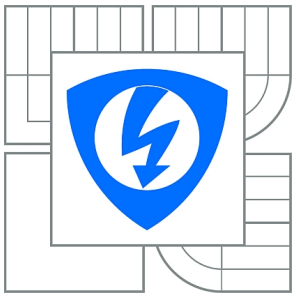
DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE Bc. VÁCLAV VRBA
AUTHOR

BRNO 2014

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

ROBUSTNÍ DETEKCE KLÍČOVÝCH SLOV V ŘEČOVÉM SIGNÁLU

ROBUST DETECTION OF KEYWORDS IN SPEECH SIGNAL

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

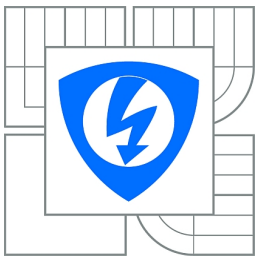
AUTOR PRÁCE — Bc. VÁCLAV VRBA

AUTHOR

VEDOUCÍ PRÁCE — Ing. HICHAM ATASSI

SUPERVISOR

BRNO 2014



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Diplomová práce

magisterský navazující studijní obor
Telekomunikační a informační technika

Student:	Bc. Václav Vrba	***ID:***	159234
Ročník:	2	***Akademický rok:***	2013/2014

NÁZEV TÉMATU:

Robustní detekce klíčových slov v řečovém signálu

POKYNY PRO VYPRACOVÁNÍ:

Proveďte rozbor nejmodernějších metod využívaných pro detekci klíčových slov v řeči. Navrhněte a následně naimplementujte v prostředí Matlab systém pro rozpoznání izolovaných slov využitím libovolné řečové databáze. Systém by měl být schopen pracovat v reálném prostředí, kde se mohou vyskytnout různé šumy a odrazy. Při návrhu je zapotřebí porovnat různé metody pro extrakci příznaků a klasifikaci z hlediska úspěšnosti klasifikace, robustnosti a výpočetní náročnosti.

DOPORUČENÁ LITERATURA:

[1] Psutka J.. Komunikace s počítačem mluvenou řečí. Academia, Praha 1995.
[2] Psutka J., Müller L., Matoušek J., Radová V.. Mluvíme s počítačem česky. Academia, Praha 2006.
[3] Sigmund M.. Analýza řečových signálů. Skripta, VUT, Brno 2000.
[4] R. Duda, P. Hart, D. Stork, Pattern Classification, druhé vydání. Wiley, 2003.

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

Vedoucí práce: Ing. Hicham Atassi
Konzultanti diplomové práce:

doc. Ing. Jiří Mišurec, CSc.
Předseda oborové rady

UPOZORNĚNÍ:

ABSTRAKT

Diplomová práce je rozdělena do dvou částí teoretické a praktické. V teoretické části je zaměřena na metody analýzy a rozpoznání řečových signálů. V praktické části byl vytvořen systém pro rozpoznávání izolovaných slov v prostředí Matlab nezávislý na mluvčím zvlášť pro muže a ženy. Dále byly vytvořeny dvě řečové databáze pro využití v kokpitu a proběhlo testování a evaluace včetně vlivu přidaného šumu.

KLÍČOVÁ SLOVA

KWD, MFCC, LPCC, DTW, KNN, HMM, Matlab, HTK

ABSTRACT

The master thesis is divided into two parts theoretical and practical. The theoretical part is focused on methods of analysis and detection of speech signals. In the practical part the system for isolated word recognition was created in Matlab. The system is speaker independent separately for men and women. Also two speech databases were created for further use in the aircraft cockpit. Tests and evaluations were performed even with added noise.

KEYWORDS

KWD, MFCC, LPCC, DTW, KNN, HMM, Matlab, HTK

VRBA, Václav *Robustní detekce klíčových slov v řečovém signálu*: diplomová práce. BRNO: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2014. 51 s. Vedoucí práce byl Ing. Hicham Atassi

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Robustní detekce klíčových slov v řečovém signálu“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové 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 autorského 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.

BRNO ....

(podpis autora)

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu diplomové práce panu Ing. Hichamu Atassimu, za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci. Dále bych rád poděkoval spolupracovníkům z Honeywellu a SŽDC.

BRNO ....

(podpis autora)

Faculty of Electrical Engineering
and Communication
Brno University of Technology
Purkynova 118, CZ-61200 Brno
Czech Republic

http://www.six.feec.vutbr.cz

PODĚKOVÁNÍ

Výzkum popsaný v této diplomové práci byl realizován v laboratořích podpořených z projektu SIX; registrační číslo CZ.1.05/2.1.00/03.0072, operační program Výzkum a vývoj pro inovace.

BRNO ....

(podpis autora)

OBSAH

Úvod **11**

1 Historie a problematika řečových technologií **12**

2 Analýza řečového signálu **13**
2.1 Kódování tvaru vlny 13
2.1.1 Pulsní kódová modulace (PCM) 13
2.1.2 Další metody kódování tvaru vlny 15
2.2 Lineární prediktivní analýza 15
2.2.1 Lineární prediktivní kódování (LPC) 15
2.2.2 Perceptivní lineární prediktivní kódování (PLP) 16
2.3 Homomorfní zpracování řečového signálu 17
2.3.1 Kepstrální koeficienty lineárního prediktivního kódování (LPCC) 18
2.3.2 Melovské kepstrální koeficienty (MFCC) 18

3 Metody rozpoznání řeči **21**
3.1 Dynamické borcení času (DTW) 21
3.2 K nejbližších sousedů (KNN) 21
3.3 Skryté Markovovy modely (HMM) 22
3.3.1 Akustické modelování pomocí HMM 23

4 Databáze **24**
4.1 Timit 24
4.2 Aurora 5 24

5 Praktická část **25**
5.1 DTW 26
5.1.1 Výběr tréninkové a testovací množiny 26
5.1.2 Úspěšnost rozpoznávání pro různé délky rámce 26
5.1.3 Porovnání výsledků 28
5.1.4 Úspěšnost rozpoznávání pro různé počty kepstrálních koeficientů 30
5.1.5 Porovnání výsledků 33
5.2 HMM 34
5.2.1 Tvorba modelů 34
5.2.2 Tvorba databáze Basic 36
5.2.3 Testování a analýza 37
5.2.4 Testování a analýza uměle zašumělých dat 38
5.2.5 Tvorba databáze Speech4EFB 39

5.2.6 Testování a analýza 43
5.2.7 Testování a analýza uměle zašumělých dat 43

6 Závěr **46**

Literatura **48**

Seznam symbolů, veličin a zkratek **50**

SEZNAM OBRÁZKŮ

2.1 Ilustrace procesu vzorkování [1] 14
2.2 Ilustrace procesu kvantizace [1] 14
2.3 Křivky stejné hlasitosti [1] 16
2.4 Banka trojúhelníkových filtrů a) v melovské škále (a) a v původní netransformované škále (b) [1] 19
3.1 Princip metody KNN 22
5.1 Úspěšnost rozpoznání při proměnných délkách rámce pro MFCC . . . 27
5.2 Úspěšnost rozpoznání při proměnných délkách rámce pro MFCC . . . 28
5.3 Úspěšnost rozpoznání při proměnných délkách rámce pro LPCC . . . 29
5.4 Úspěšnost rozpoznání při proměnných délkách rámce pro LPCC . . . 29
5.5 Srovnání úspěšnosti rozpoznání při délce rámce 30 ms pro MFCC a LPCC 30
5.6 Srovnání úspěšnosti rozpoznání při délce rámce 30 ms pro MFCC a LPCC 31
5.7 Úspěšnost rozpoznání při různém počtu MFCC koeficientů 32
5.8 Úspěšnost rozpoznání při různém počtu MFCC koeficientů 32
5.9 Úspěšnost rozpoznání při různém počtu LPCC koeficientů 33
5.10 Úspěšnost rozpoznání při různém počtu LPCC koeficientů 34
5.11 Srovnání úspěšnosti rozpoznání MFCC a LPCC pro 12 koeficientů . . 35
5.12 Srovnání úspěšnosti rozpoznání MFCC a LPCC pro 12 koeficientů . . 35
5.13 Úspěšnost rozpoznání frází ze zašumělé databáze Basic 39
5.14 Rozmístění nahrávacích zařízení v leteckém simulátoru: 1 - sluchátka, 2 - mikrofon, 3 - tablet 40
5.15 Úspěšnost rozpoznání frází z databáze Speech4EFB 44
5.16 Úspěšnost rozpoznání frází ze zašumělé databáze Speech4EFB pro nahrávky ze sluchátek 45

SEZNAM TABULEK

4.1 Rozdělení mluvčích podle dialektů 24
5.1 Různé úrovně hlasitosti dle FAA [3] 25
5.2 Informace o databázi Aurora a výběru nahrávek pro další práci 26
5.3 Úspěšnost rozpoznání při proměnných délkách rámce pro MFCC . . . 27
5.4 Úspěšnost rozpoznání při proměnných délkách rámce pro LPCC . . . 28
5.5 Srovnání úspěšnosti rozpoznání při proměnných délkách rámce pro MFCC a LPCC 30
5.6 Úspěšnost rozpoznání při různém počtu MFCC koeficientů 31
5.7 Úspěšnost rozpoznání při různém počtu LPCC koeficientů 33
5.8 Srovnání úspěšnosti rozpoznání MFCC a LPCC pro různý počet koeficientů 34
5.9 Rozdělení mluvčích 36
5.10 Údaje o jednotlivých mluvčích 37
5.11 Úspěšnost rozpoznání frází z databáze Basic 38
5.12 Úspěšnost rozpoznání frází ze zašumělé databáze Basic 38
5.13 Rozdělení mluvčích 41
5.14 Údaje o jednotlivých mluvčích 42
5.15 Úspěšnost rozpoznání frází z databáze Speech4EFB 43
5.16 Úspěšnost rozpoznání frází ze zašumělé databáze Speech4EFB pro nahrávky ze sluchátek 44

ÚVOD

Rozpoznání řeči nabývá v poslední době stále více na významu. Vzhledem k prudkému rozvoji v oblasti dotykových zařízení, například tabletů a chytrých telefonů, díky nimž dochází k enormnímu růstu výpočetních výkonů těchto zařízení, které dnes směle konkurují klasickým stolním PC či laptopům, se možnosti zpracování mluvené řeči posouvají směrem ke koncovým uživatelům a pro ně určených aplikací. Trend je v tomto směru neúprosný a směřuje od ovládání pomocí myši a klávesnice přes ovládání dotykem k ovládání pomocí hlasových příkazů.
Jedna z oblastí zpracování řeči se věnuje právě detekci klíčových slov v řečovém signálu. Zajímavou možností je implementace v letectví. V současné době totiž nemá žádná společnost aplikaci s hlasovým ovládáním certifikovanou pro letecký průmysl. Ve své diplomové práci si kladu za cíl vytvořit základní kameny pro budoucí vývoj a testování vytvořením systému v prostředí Matlab pro rozpoznávání izolovaných slov. Struktura práce je rozdělena do dvou částí teoretické a praktické. V teoretické části jsou řešeny metody analýzy a rozpoznávání řečových signálů. V praktické části, která byla realizována ve spolupráci s firmou Honeywell sekcí Aerospace, se zabývám realizací vlivu různých změn parametrů při homomorfním zpracování řečového signálu na úspěšnost rozpoznání. Dále pak vývojem prototypu systému nezávislém na mluvčím zvlášť pro ženy a muže, tvorbě dvou řečových databází zohledňující další využití v kokpitu dopravního letadla a následným testováním a analýzou včetně vlivu šumu.

1 HISTORIE A PROBLEMATIKA ŘEČOVÝCH TECHNOLOGIÍ

Základním způsobem přenosu informací mezi lidmi je mluvená řeč. Později se pro uchování těchto informací vyvinula její psaná forma. Při komunikaci s výpočetní technikou byl vývoj přesně opačný. Nejprve jsme výpočetní techniku naučili číst a psát a teprve se zvyšujícím se výpočetním výkonem a s rostoucí kapacitou paměti učíme stroje mluvit a rozumět mluvené řeči. Konečným cílem těchto snah je vytvoření plnohodnotného partnera pro mluvený dialog.

První pokusy s hlasovým syntetizérem byly popsány již v druhé polovině 18. století, kdy je nezávisle na sobě prováděli von Kempelen a Kratznestein. Poté bylo provedeno mnoho dalších experimentů s analýzou, syntézou a rozpoznáváním řeči. Větší rozvoj řečových technologií nastal až s nástupem číslicových počítačů, který umožňoval digitalizaci a číslicové zpracování řeči. Metody pro porovnávání a vyčíslení podobnosti dvou promluv byly také velice ovlivněny rozvojem výpočetní techniky. V sedmdesátých letech došlo k rozvoji rozpoznávání jednotlivých izolovaně vyslovených slov a slovních spojení, které zahrnovalo i proměnlivost tempa řeči. V průběhu osmdesátých let byla vyvinuta technika klasifikace řeči, založená na statistickém přístupu. Postupně došlo ke změně základních využívaných částí pro modelování řeči od jednotlivých slov k subslovním jednotkám (fonémy, alofony, trifony, apod.), které jsou využívány dodnes. K obdobnému vývoji došlo i v oblasti syntézy řeči.

Problematiku řečových technologií můžeme rozdělit na několik samostatných částí: zpracování řečového signálu, rozpoznávání řeči, porozumění významu a syntéza řeči. V současnosti se spíše využívají jednotlivé části samostatně a pro jednotlivé specifické oblasti (hlasové ovládání strojů, automatické hlasové přepojování hovorů, automatický přepis diktátů, automatické titulkování, vyhledávání v řečových databázích, automatické předčítání, apod.) a pokračují snahy o vytvoření komplexních dialogových systémů, které umožní komunikaci neomezeným přirozeným jazykem nebo jazyky včetně automatického překládání.

2 ANALÝZA ŘEČOVÉHO SIGNÁLU

Hlavními oblastmi zpracování řečových signálů jsou kódování a přenos řeči, rozpoznání řeči, syntéza řeči, identifikace a verifikace osob podle hlasu. Požadavky na způsob zpracování řečového signálu se pro jednotlivé oblasti značně liší. Zatím co pro efektivní kódování a přenos řečového signálu je rozhodující přenosová rychlost, tak pro rozpoznání řeči je hlavní kvalita informačního obsahu a pro identifikaci a verifikaci řečníka jsou nejdůležitější odlišnosti hlasů.
Většina metod analýzy řečového signálu vychází z předpokladu, že se vlastnosti řečového signálu v průběhu času mění pomalu. Pro zpracování řeči se tak využívají metody krátkodobé analýzy, při kterých se signál rozdělí na krátké časové úseky (mikrosegmenty) o typické délce 10 ms. Takto vzniklé oddělené krátké zvuky jsou popsány číslem nebo souborem čísel. Poté jednotlivé mikrosegmenty opět spojíme dohromady a dostaneme výsledné časové číselné posloupnosti, které jsou výsledkem analýzy a popisují promluvený celek. Metody krátkodobé analýzy používají většinou jako vstup data získaná digitalizací signálu (kódováním tvaru vlny).

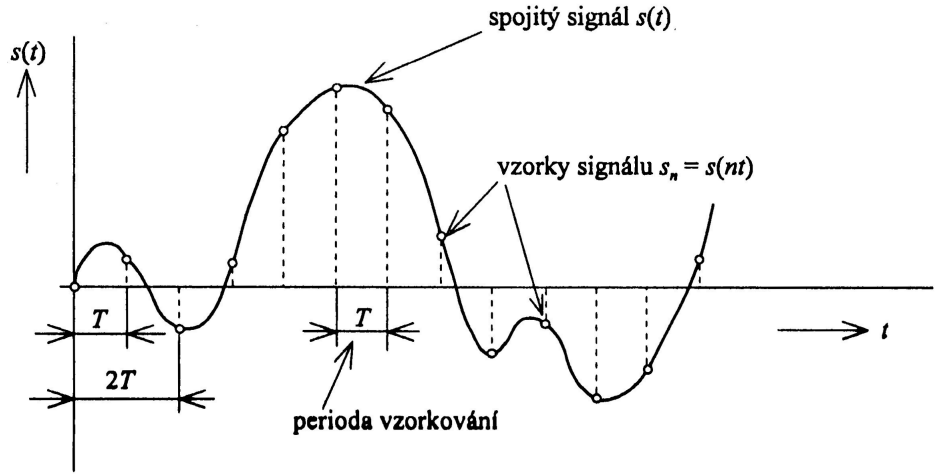
2.1 Kódování tvaru vlny

2.1.1 Pulsní kódová modulace (PCM)

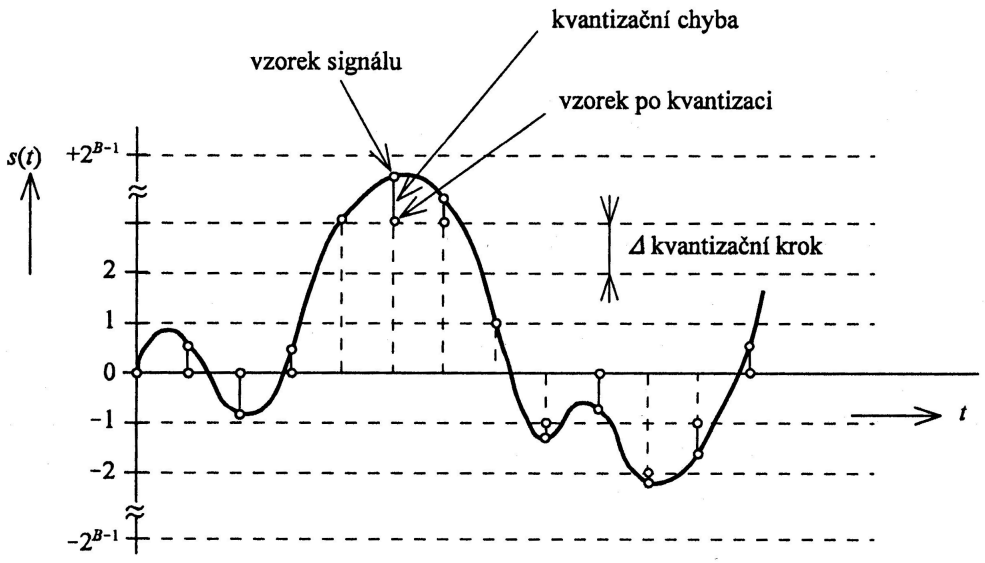
Pro analýzu řeči je nutné nejprve převést analogový signál, který obvykle získáme snímáním řeči pomocí mikrofonu, do číslicového tvaru. Tento proces, který nazýváme pulsní kódová modulace, se skládá ze dvou postupných kroků: vzorkování a kvantizace s kódováním. Vzorkování (obr.2.1) je transformace časově spojitého signálu $s(t)$ na diskrétní posloupnost $s_n = s(nT)$, kde T je perioda vzorkování a $n \in < 0, \infty >$. Frekvence vzorkování $F_v = 1/T$ je v souladu s Nyquistovým vzorkovacím teorémem omezena a musí splňovat $F_v \geq 2F_m$, kde F_m je horní hranice frekvenčního pásma analogového signálu $s(t)$ a dolní hranice frekvenčního pásma je 0 Hz. Pak platí

$$s(t) = \sum_{n=-\infty}^{\infty} s(nT) \left[\frac{\sin \pi(\frac{t}{T} - n)}{\pi(\frac{t}{T} - n)} \right] \tag{2.1}$$

Při porušení vzorkovacího teorému dochází ke zkreslení složek vyšších frekvencí. Kvantizace (obr. 2.2) s následujícím kódováním je aproximací analogové hodnoty vzorku signálu jednou z konečného počtu číselných hodnot.
Kvantizér s rovnoměrně rozloženými kvantizačními úrovněmi je charakterizován počtem úrovní kvantování a kvantizačním krokem Δ. Počet úrovní kvantování se obvykle volí ve tvaru 2^B, kde B je počet bitů v binárním kódu. Pro pokrytí celého



Obr. 2.1: Ilustrace procesu vzorkování [1]



Obr. 2.2: Ilustrace procesu kvantizace [1]

rozsahu signálu volíme charakteristiky kvantizéru podle

$$2S_{max} = \Delta 2^B \tag{2.2}$$

kde S_{max} je maximální úroveň vzorkovaného signálu. V průběhu kvantizace dochází ke kvantizačnímu zkreslení (kvantizační šum) v důsledku přiřazení měřených okamžitých velikostí signálu obvykle k nejbližší nižší hodnotě kvantizační úrovně. Rozložení kvantizačního šumu v rozsahu kvantizačního kroku je rovnoměrné. Pro kvalitní zá-

znam řečového signálu je potřeba jedenácti až dvanácti bitový převod a pro vysoce kvalitní záznam se doporučuje šestnácti bitový převod.

2.1.2 Další metody kódování tvaru vlny

Protože metoda pulsní kódové modulace je značně informačně redundantní a není možné dále snižovat frekvenci vzorkování, je pozornost soustředěna na metody, které snižují počet bitů na vzorek. Možnosti tohoto snížení souvisí s širokým dynamickým rozsahem řečového signálu (kódování μ-law a A-law) a se značnou korelací sousedních vzorků signálu (diferenční pulsní kódová modulace - DPCM).

2.2 Lineární prediktivní analýza

2.2.1 Lineární prediktivní kódování (LPC)

Jednou z nejefektivnějších metod analýzy řečového signálu je lineární prediktivní kódování, které je založeno na snaze odhadnout přímo z řečového signálu parametry modelu vytváření řeči. Výhodou této metody je relativně přijatelná výpočetní zátěž při odhadování těchto parametrů.
Model vytváření řeči je založen na spojení modelů jednotlivých fází vzniku řečového signálu, který vzniká nejprve v hlasivkách a poté ho ovlivňuje průchod hlasivkovým traktem a vlastní vyzařování zvuku. Při modelování vycházíme z předpokladu, že typ buzení a vlastnosti hlasového traktu zůstávají téměř konstantní po krátké časové úseky (10 - 30 ms). Model produkce řeči se skládá z lineárního modelu hlasového traktu s pomalu se měnícími parametry, který je buzen periodickým sledem pulzů (znělá řeč) nebo náhodným šumem (neznělá řeč).
Princip metody LPC vychází z předpokladu, že k-tý vzorek signálu $s(k)$ lze popsat jako lineární kombinaci Q předchozích vzorků, které určují řád modelu, a buzení $u(k)$ s koeficientem zesílení G:

$$s(k) = -\sum_{i=1}^{Q} a_i s(k-i) + Gu(k) \tag{2.3}$$

Přenosovou funkci modelu $H(z)$ můžeme vyjádřit jako

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{A(z)} = \frac{G}{1 + \sum_{i=1}^{Q} a_i z^{-i}} \tag{2.4}$$

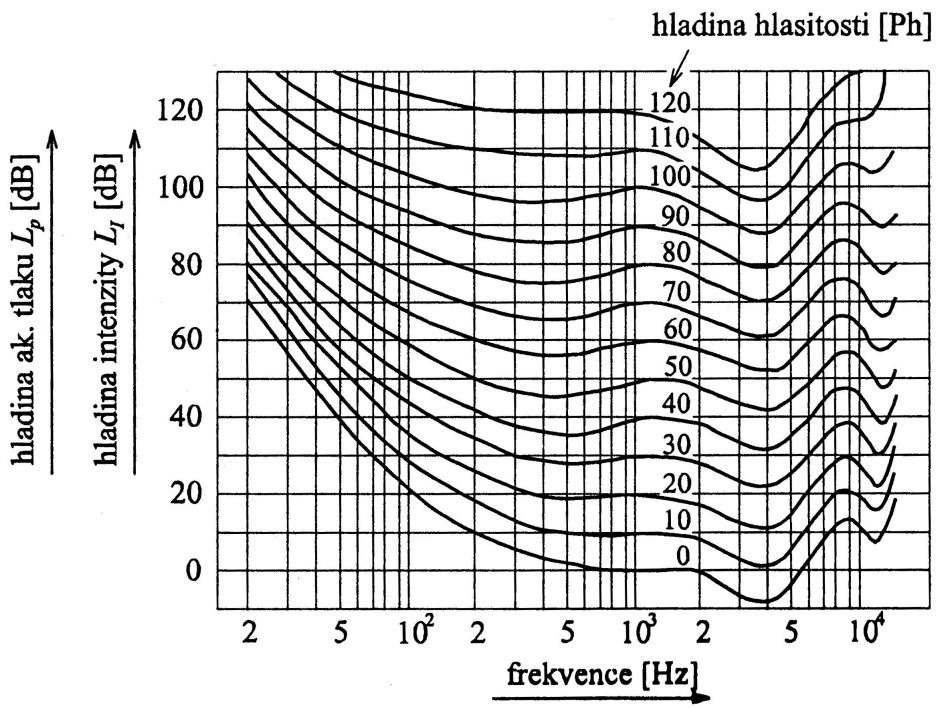
Sledovanými parametry jsou koeficienty číslicového filtru a_i a koeficient zesílení G. Vyjdeme-li z předpokladu, že signál je na sledovaném časovém intervalu stacionární,

můžeme pro výpočet koeficientů použít metodu nejmenších čtverců.
Koeficienty a_i slouží také pro výpočet spektra signálu, které má charekter vyhlazené spektrální obálky skutečného spektra původního diskretizovaného signálu $s(k)$, Pro frekvenční přenos modelu vytváření akustického signálu platí po substituci $z = e^{j\omega}$

$$H(j\omega) = \frac{G}{1 + \sum_{i=1}^{Q} a_i e^{-j\omega i}} \tag{2.5}$$

2.2.2 Perceptivní lineární prediktivní kódování (PLP)

Popis spektrálních vlastností řečového signálu pomocí lineární preditivní analýzy (LPC) je sice velmi efektivní, ale příliš neodpovídá způsobu vnímání řeči člověkem. Lidský sluch vnímá intenzitu a frekvence tónů nelineárně viz. obr. 2.3.



Obr. 2.3: Křivky stejné hlasitosti [1]

Metoda LPC nezohledňuje také jev nazývaný maskování zvuků, ze kterého vyplývá zavedení kritických pásem spektrální citlivosti. Jako reakce na tyto nedostatky byla vyvinuta metoda perceptivní lineární prediktivní analýzy (PLP). Tato metoda využívá pro transformaci spektra řečového signálu do odpovídajícího sluchového spektra kombinací tří složek psychofyziky slyšení: kritické pásmo spektrální citlivosti, křivky stejné hlasitosti a závislost vnímané hlasitosti na intenzitě zvuku.

2.3 Homomorfní zpracování řečového signálu

Jedním z nelineárních způsobů zpracování řečového signálu, které je založené na zobecněném proncipu superpozice, je homomorfní analýza. Tyto metody jsou vhodné pro analýzu a oddělování signálů vzniklých konvolucí dvou nebo více signálů. Protože modelování řeči je založeno na konvoluci budící funkce a impulsní odezvy hlasového ústrojí, je homomorfní analýza vhodnou metodou pro zpracování řečových signálů. Pro homomorfní zpracování signálu vycházíme z předpokladu, že je dána diskrétní posloupnost $x(n)$, která vznikla konvolucí posloupností $x_1(n)$ a $x_2(n)$

$$x(n) = x_1(n) * x_2(n) \tag{2.6}$$

Pak je možné popsat funkci charakteristického systému D_* následujícími rovnicemi:

$$X(z) = Zx(n) = Zx_1(n) * x_2(n) = X_1(z)X_2(z) \tag{2.7}$$

$$\hat{X}(z) = \log(X(z)) = \log(X_1(z)) + \log(X_2(z)) = \hat{X}_1(z) + \hat{X}_2(z) \tag{2.8}$$

$$\hat{x}(n) = Z^{-1}\{\hat{X}(z)\} = Z^{-1}\{\hat{X}_1(z) + \hat{X}_2(z)\} = \hat{x}_1(z) + \hat{x}_2(z) \tag{2.9}$$

Pomocí charakteristického systému D_* převedeme konvolutorní součin vstupních signálů na součet modifikovaných vstupních signálů. Vyčíslíme-li tuto transformaci na jednotkové kružnici ($z = e^{j\omega}$ - Fourierova transformace) dostaneme z rovnice 2.8

$$\hat{X}(e^{j\omega}) = \log|X(e^{j\omega})| + j\arg(X(e^{j\omega})) \tag{2.10}$$

Rovnice 2.10 musí být jednoznačně definována, proto je nutné zavést předpoklad, že $\arg(X(e^{j\omega}))$ je lichá spojitá funkce ω . Potom můžeme určit

$$\hat{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\hat{X}(e^{j\omega})e^{j\omega n}d\omega \tag{2.11}$$

Výraz $\hat{x}(n)$ se nazývá komplexní kepstrum a je to zpětná Fourierova transformace logaritmu Fourierova obrazu vstupního signálu $x(n)$. Pokud použijeme při výpočtu reálnou část $\hat{X}(e^{j\omega}$ dostaneme kepstrum $c(n)$

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log|X(e^{j\omega})|e^{j\omega n}d\omega \tag{2.12}$$

Pro výpočet komplexního kepstra se využívá diskrétní Fourierova transformace a inverzní diskrétní Fourierova transformace.

2.3.1 Kepstrální koeficienty lineárního prediktivního kódování (LPCC)

Hlasový trakt modelovaný lineárním systémem je také možné popsat pomocí kepstrálních koeficientů. Pro jejich výpočet musíme nejprve určit logaritmus přenosové funkce, který vyjádříme za pomoci Taylorova rozvoje, když předpokládáme, že polynom $A(z)$ je Q-tého řádu a všechny kořeny tohoto polynomu jsou uvnitř jednotkové kružnice

$$\log(H(z)) = \log\left(\frac{G}{A(z)}\right) = c(0) + c(1)z^{-1} + \ldots = \sum_{k=0}^{\infty} c(k)z^{-k} \tag{2.13}$$

kde $c(k)$ jsou kepstrální koeficienty LPC. Po derivaci a úpravě dostáváme rovnici

$$-\sum_{i=1}^{Q} i a_i z^{-i} = \left[\sum_{k=1}^{\infty} k c(k) z^{-k}\right]\left[\sum_{i=0}^{Q} a_i z^{-i}\right] \tag{2.14}$$

Uvažujeme-li $a_0 = 1$, pak můžeme odvodit vztahy pro výpočet kepstrálních koeficientů LPC

$$c(1) = -a_1 \tag{2.15}$$

$$c(k) = -a_k - \sum_{i=1}^{k-1}\left(\frac{i}{k}\right) c(i) a_{k-i} \quad pro \quad 2 \leq k \leq Q \tag{2.16}$$

$$c(k) = -\sum_{i=1}^{Q}\left(\frac{k-i}{k}\right) c(k-i) a_i \quad pro \quad k = Q+1, Q+2, \ldots \tag{2.17}$$

Pro správnou reprezentaci spektrální obálky analyzovaného mikrosegmentu je třeba vyčíslovat vždy $k = 1, 2, \ldots, Q^*$ kepstrálních koeficientů LPC, kde $Q^* \geq Q$ často se však volí dolní mez tzn. $Q^* = Q$. Hlavní výhoda použití kepstrálních koeficientů LPC pro analýzu řečových signálů je, že tyto koeficienty nekorelují.

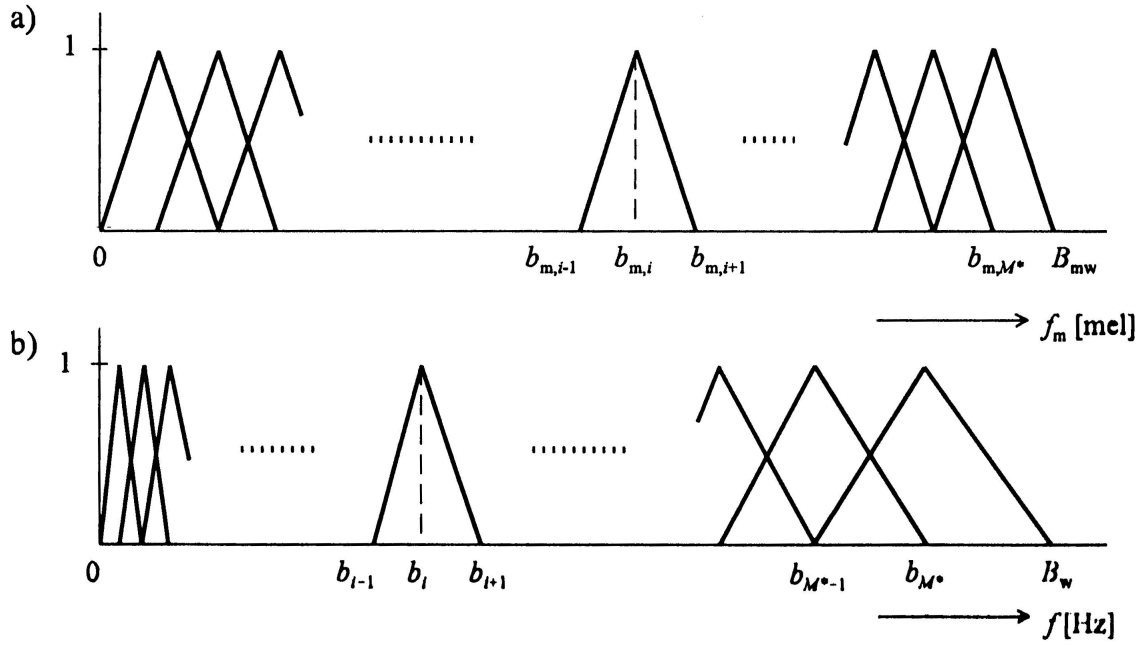
2.3.2 Melovské kepstrální koeficienty (MFCC)

Zpracování řečových signálů analýzou Melovských kepstrálních koeficientů (MFCC) se snaží respektovat nelineární vlastnosti vnímání řečových signálů lidským sluchem především nelineární vnímání frekvencí. Tato metoda analýzy využívá banku trojúhelníkových pásmových filtrů, které mají lineární rozložení frekvencí v melovské frekvenční škále definované

$$f_m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.18}$$

kde f $[Hz]$ je frekvence v lineární škále a f_m $[mel]$ je frekvence v nelineární melovské škále.

V preemfázi jsou u řečového signálu zdůrazněny amplitudy složek s vyššími frekvencemi. Poté je na mikrosegmenty signálu nejčastěji aplikováno Hammingovo okénko, jehož délka se volí jako mocnina dvou (výhodné pro zpracování FFT) a které se obvykle posouvá v čase o 10 ms. Dále se provede pomocí FFT výpočet amplitudového spektra, na ktré navazuje melovská filtrace. Počet pásem banky melovských filtrů (počet trojúhelníkových filtrů v bance filtrů) je vhodné volit v závislosti na počtu a umístění kritických pásem a zohlednit také velikost vzorkovací frekvence a celkové šířky přenášeného pásma B_w $[Hz]$ respektive B_{mw} $[mel]$ viz. obr. 2.4.



Obr. 2.4: Banka trojúhelníkových filtrů a) v melovské škále (a) a v původní netransformované škále (b) [1]

V melovské frekvenční škále mají odezvy jednotlivých filtrů tvar rovnoramenných trojúhelníků, které jsou rovnoměrně rozděleny ve frekvenčním spektru. Při průchodu signálu filtrem je každý koeficient FFT násoben odpovídajícím ziskem filtru a výsledky jsou pro příslušné filtry akumulovány. V dalším kroku analýzy jsou vypočteny logaritmy výstupů jednotlivých filtrů, kterými dojde k omezení dynamiky signálu. A nakonec výpočtu melovských kepstrálních koeficientů (MFCC) $c_m(j)$, je provedena zpětná diskrétní Fourierova transformace (IDFT), která může být díky reálnému a symetrickému výkonovému spektru redukována na diskrétní kosinovou transfor-

maci (DCT)

$$c_m(j) = \sum_{i=1}^{M^*} \log y_m(i) \cos\left(\frac{\pi j}{M^*}(i - 0,5)\right) \quad pro \quad j = 0, 1, \ldots, M \tag{2.19}$$

kde M^* je počet pásem melovského pásmového filtru a M je počet melovských kepstrálních koeficientů. Počet koeficientů může být podstatně menší než počet pásem melovského pásmového filtru a obvykle postačuje 10 až 13 koeficientů.

3 METODY ROZPOZNÁNÍ ŘEČI

Při sestavování systému pro rozpoznávání mluvené řeči je třeba vyřešit několik základních problémů, které souvisí s variabilitou řečníka, variabilitou prostředí, ve kterém sledovaný řečník mluví, a také se složitostí řešené úlohy. Systémy pro rozpoznávání řeči mohou být na řečníku závislé (jsou uzpůsobeny parametrům řeči jednotlivce nebo malé skupiny) nebo nezávislé (jsou vytvořeny univerzálně a využívají velké množství různých hlasů - řádově stovky až tisíce). Možnosti rozpoznání řeči jsou také velmi ovlivněny akustickými vlastnostmi prostředí hlavně přítomností a velikostí okolního šumu. V závislosti na složitosti řešené úlohy můžeme systémy rozpoznávání řeči rozdělit na systémy pro rozpoznávání izolovaných slov (jednodušší) a systémy pro rozpoznávání souvislé řeči (složitější).
Metody rozpoznávání řeči jsou buď založeny na principu porovnávání se vzory (např. dynamické borcení času), nebo využívají statistické metody (skryté Markovovy modely).

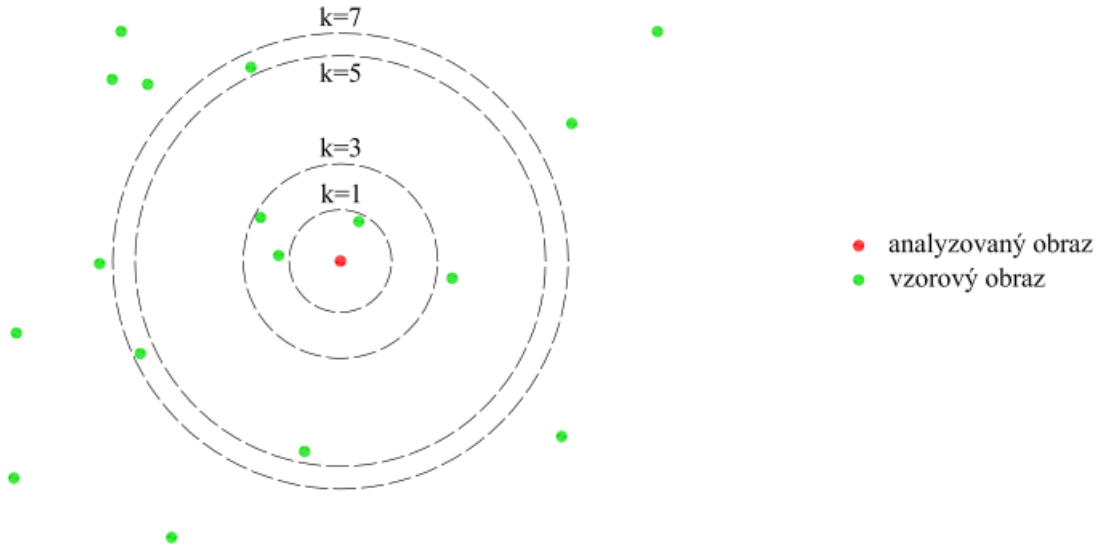
3.1 Dynamické borcení času (DTW)

Metody založené na principu porovnávání se vzory jsou použitelné hlavně pro rozpoznávání izolovaných slov. Vzorové obrazy jsou tvořeny vzorovými slovy, které jsou definovány posloupností příznakových vektorů. Při rozpoznávání, tzn. určování třídy tato metoda hledá nejmenší vzdálenost mezi analyzovaným řečovým signálem a vzorovým obrazem. Tato vzdálenost se obvykle určuje pomocí metody dynamického programování, která řešení hledá pomocí nelineární transformace časové osy jednoho z obrazů (analyzovaný nebo vzorový obraz), při které je vzdálenost obrazů nejmenší. Na nelineární časové normalizaci je založena metoda dynamického borcení času (DTW).
Tato metoda vznikla na základě podrobné analýzy několika nahrávek stejnoho slova vysloveného stejným řečníkem. Analýza prokázala, že odlišnosti mezi odpovídajícími signály nespočívají v oblasti spektra, ale v časovém členění. Odlišnosti v časovém členění vznikají na základě nestejné délky vyslovování slov a také v nepoměru odpovídajících si částí uvnitř slova (fonémy, hlásky).

3.2 K nejbližších sousedů (KNN)

Nejjednodušším způsobem klasifikace je metoda K nejbližších sousedů (KNN). Třída analyzovaného obrazu je určena na základě vzdálenosti k vzorovým obrazům (obr. 3.1). Analyzovaný obraz zařadíme do třídy, která má nejvíce vzorových obrazů mezi

K nejbližšími obrazy k analyzovanému obrazu. Počet hledaných nejbližších obrazů K volíme obvykle lichý, čímž omezíme pravděpodobnost nerozhodného výsledku mezi dvěma třídami.



Obr. 3.1: Princip metody KNN

3.3 Skryté Markovovy modely (HMM)

Statistické metody rozpoznání řeči jsou založeny na skrytých Markovových modelech (HMM) a jsou využitelné pro rozpoznávání jednotlivých slov i celých promluv. Slova mohou být modelována jako celek, nebo mohou být samostatně modelovány subslovní jednotky (slabiky, fonémy, trifony, ...), což je častější. Rozpoznávání řeči probíhá ve dvou krocích pomocí metody dekódování s maximální aposteriorní pravděpodobností. Nejdříve akustický procesor převádí řečový signál na posloupnosti vektorů příznaků a poté lingvistický dekodér převádí posloupnost příznaků na řetězec slov.

Máme-li posloupnost N slov $W = \{w_1, w_2, \ldots, w_N\}$ a jí odpovídající posloupnost vektorů příznaků akustické informace $O = \{o_1, o_2, \ldots, o_T\}$, potom hledáme posloupnost slov $\hat{W}$, která je nejpravděpodobnější posloupností slov pro danou akustickou informaci O. Hledáme tedy maximální podmíněnou pravděpodobnost $P(W|O)$ a s využitím Bayesova pravidla dostaneme rovnici

$$\hat{W} \doteq argmax_W P(W|O) = argmax_W \frac{P(W)P(O|W)}{P(O)} \tag{3.1}$$

kde $P(O|W)$ je pravděpodobnost, že při vyslovení posloupnosti slov W vznikne

posloupnost výstupních vektorů příznaků O. $P(W)$ je apriorní pravděpodobnost posloupnosti slov W (tj. pravděpodobnost, že řečník bude číst posloupnost slov W) a $P(O)$ je apriorní pravděppodobnost posloupnosti výstupních vektorů. Protože pravděpodobnost $P(O)$ není funkcí W, lze ji při hledání maxima rovnice 3.1 ignorovat. Hledanou posloupnost slov $\hat{W}$ lze tedy určit maximalizací sdružené pravděpodobnosti $P(W, O)$

$$\hat{W} = argmax_W P(W, O) = argmax_W P(W)P(O|W) \tag{3.2}$$

Z rovnice 3.2 vyplývá, že problém stanovení nejlepší posloupnosti slov k danému akustickému signálu vyjádřenému posloupností pozorovaných vektorů příznaků $O = \{o_1, o_2, \ldots, o_T\}$ lze řešit pomocí dvou oddělených pravděpodobností $P(O|W)$ a $P(W)$, které mohou být modelovány a trénovány nezávisle na sobě. Podmíněné rozdělení pravděpodobnosti $P(O|W)$ nese informaci o akustickém modelu (model řečníka) a apriorní rozdělení pravděpodobnosti $P(W)$ nese informaci o jazykovém modelu. Tyto dva zdroje znalostí je třeba určit ještě před vlastním rozpoznáváním, a to nejčastěji na základě trénování z řečových a jazykových dat. [1]

3.3.1 Akustické modelování pomocí HMM

Akustický model je sestaven pro co nejpřesnější a nejrychlejší odhad podmíněné pravděpodobnosti $P(O|W)$ pro kteroukoliv možnou posloupnost vektorů příznaků O odpovídající každé uvažované posloupnosti slov W. Hlavními požadavky na tyto modely jsou flexibilita (nezávislost na hlasu, artikulaci, tempu řeči, akustickém pozadí), přesnost a účinnost (rozpoznávání v reálném čase). Efektivním způsobem na řešení této úlohy je využití skrytých Markovových modelů (HMM).
Modelování řeči pomocí HMM je založeno na interpretaci způsobu vytváření řeči člověkem a vychází z předpokladu, že hlasové ústrojí se během krátkého časového úseku (tzv. mikrosegmentu) nachází v jednom z konečného počtu stavů a generuje signál, který můžeme popsat pomocí spektrálních charakteristik (vektorem příznaků). Při modelování řečového signálu pomocí HMM jsou vytvářeny dvě časově svázané posloupnosti náhodných proměnných podpůrný Markovův řetězec (posloupnost konečného počtu stavů) a řetězec vektorů příznaků (posloupnost spektrální charakteristik mikrosegmentů řečového signálu).

4 DATABÁZE

4.1 Timit

Timit je databáze čtené řeči určená pro studium akusticko-fonetických jevů a vývoj a testování systémů automatického rozpoznání řeči. Rozděluje Spojené státy americké na 8 regionů dle dialektů americké angličtiny. Obsahuje celkem 6300 vět, přičemž každý z 630 mluvčích čte 10 vět z různých regionů. Z 630 mluvčích tvoří 70% muži, 30% ženy. Detailní přehled rozdělení je v tabulce (viz. tab. 5.1). Zvukové soubory jsou nahrány v kvalitě 16 000 Hz, 16 bit. Každá věta je pak charakterizována soubory obsahující časový fonetický přepis, časový slovní přepis a prostý textový přepis. Databáze byla vyvinuta ve spolupráci Massachusetts Institute of Technology, SRI International a Texas Instruments, Inc.[2].

Tab. 4.1: Rozdělení mluvčích podle dialektů

Oblast dialektu	**Muži**		**Ženy**		**Celkem**	
New England	31	63%	18	27%	49	8%
Northern	71	70%	31	30%	102	16%
North Midland	79	67%	23	23%	102	16%
South Midland	69	69%	31	31%	100	16%
Southern	62	63%	36	37%	98	16%
New York City	30	65%	16	35%	46	7%
Western	74	74%	26	26%	100	16%
Army Brat (moved around)	22	67%	11	33%	33	5%
Celkem	438	70%	192	30%	630	100%

4.2 Aurora 5

Pro realizaci byla použita číslicová databáze Aurora 5 verze c6, která obsahuje 2388 nahrávek číslic v anglickém jazyce, které namluvili různí mužští a ženští mluvčí. Tyto nahrávky obsahují také šumy na pozadí o různé intenzitě. Nahrávky byly pořízeny v International Computer Science Institute in Berkeley, mají vzorkovací frekvencí 8 kHz a jsou uloženy v *.raw formátu.

5 PRAKTICKÁ ČÁST

Praktická část je realizována ve spolupráci s firmou Honeywell sekcí Aerospace. V rámci rozvoje nabízených služeb a vzhledem k vývoji nové generace multi-modálně ovládaného kokpitu se objevuje možnost zapojení hlasového ovládání. Cílem je vytvoření prototypu systému pro rozpoznání izolovaných slov, který by mohl být následně implementován ve vyvíjené aplikaci pro další testování. Prototyp je určen pro kokpit dopravních letadel typu Boeing 737 nebo Airbus A320. Hlasové ovládání by mělo být primárně zaměřeno na muže Američana rodilého mluvčího. V současné době nemá žádná společnost aplikaci s hlasovým ovládáním certifikovanou pro letecký průmysl. Jedním z významných problémů je specifický slang pilotů, ale samožřejmě největším zůstává šum. Různé úrovně hlasitosti dle Federal Aviation Administration (FAA) jsou uvedeny níže (tab. 5.1) a podrobněji popsány [3].

Tab. 5.1: Různé úrovně hlasitosti dle FAA [3]

Zdroj	**Hlasitost [dB]**
Šeptání	20 – 30
Běžná kancelář	40 – 60
Průměrná mužská konverzace	60 – 65
Hlučná kancelář	60 – 80
Kokpit proudového letadla	70 – 90
Rušná městská ulice	80 – 100
Kokpit helikoptéry	80 – 102
Motorová pila	100 – 110
Sněžný skůtr	110 – 120
Proudový motor	130 – 160

Následujícím krokem praktické části je návrh a následná implementace systému pro rozpoznání izolovaných slov založeném na DTW s KNN s využitím libovolné řečové databáze v prostředí Matlab. Systém by měl být schopen pracovat v reálném prostředí, kde se mohou vyskytnout různé šumy a odrazy. Při návrhu je zapotřebí porovnat různé metody pro extrakci příznaků a klasifikaci z hlediska úspěšnosti klasifikace, robustnosti a výpočetní náročnosti.
Dále přejdeme k využití statistických metod rozpoznání řeči a systémům nezávislých na mluvčím zvlášť pro ženy a muže založenému na HMM v prostředí Matlab s využitím HTK. Poté k tvorbě vlastních databází (Basic a Speech4EFB) a testování zohledňující různé úrovně šumu s přihlédnutím ke specifickým podmínkám v kokpitu.

5.1 DTW

5.1.1 Výběr tréninkové a testovací množiny

Z databáze Aurora 5 verze c6 bylo vybráno 60 vzorků pro každou z číslic one, two, three, four, five, six, seven, eight a nine. Následně byl počet vzorků pro jednotlivé číslice náhodně rozdělen na 50 referenčních vzorků a 10 testovacích vzorků s přihlédnutím ke statistickému rozložení ženských a mužských mluvčích ve vybraných vzorcích (původně vybraných 60 vzorcích). Výsledkem je databáze 450 referenčních vzorků a 90 testovacích vzorků (tab. 5.2).

Tab. 5.2: Informace o databázi Aurora a výběru nahrávek pro další práci

Počet nahrávek celkem	2388
Počet nahrávek čísel 1-9	559
Počet zvolených nahrávek 1-9	540
Počet referenčních vzorů pro každé z čísel	50
Počet testovacích vzorků pro každé z čísel	10
Počet referenčních vzorů celkem	450
Počet testovacích vzorků celkem	90

5.1.2 Úspěšnost rozpoznávání pro různé délky rámce

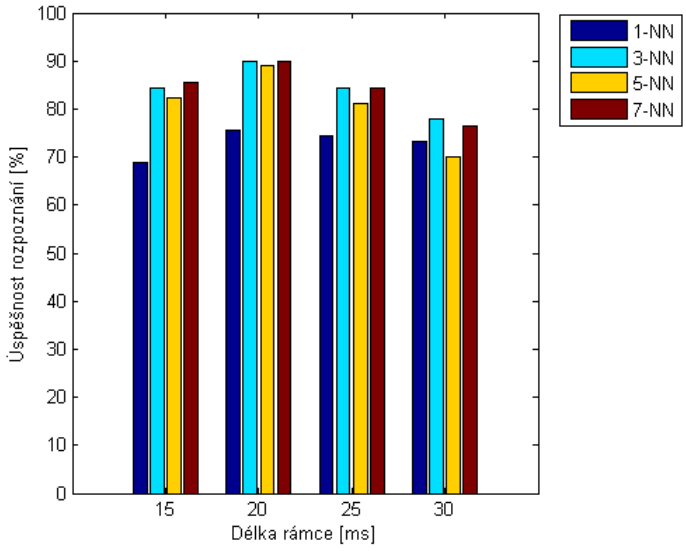
MFCC

Nejprve bylo pro každý vzorek z databáze vypočítáno 12 MFCC koeficientů s konstantní délkou překrytí 12 ms a variabilní délkou rámce. Délka rámce se měnila od 15 ms do 30 ms po 5 ms. Pak byla vypočtena pomocí DTW vzdálenost referenčních vzorků od vzorků testovacích. Na závěr proběhlo vyhodnocení metodou KNN dle počtu 1, 3, 5 a 7 nejbližších referenčních vzorků a porovnání výsledku rozpoznávání s předem známou skutečně vyslovenou číslicí. Dosažené úspěšnosti rozpoznávání jsou uvedeny v tabulce 5.3.

V následujících grafech (viz. obr. 5.1 a 5.2) vidíme, že nejhorší úspěšnost rozpoznání odpovídá délce rámce 15 ms. Je to způsobeno pomalým časovým posunem a tím pádem velmi podobnými parametry. Naopak nejlépe dopadlo rozpoznání pro délky rámce 20 ms a 30 ms, což odpovídá délce fonémů v anglickém jazyce. Rozpoznávání dle referenčních vzorů pomocí KNN nám v tomto případě ukazuje, že nejvyšší úspěšnosti dosáhneme při využití 3-NN, kdy nám 3 čísla umožní lépe určit danou číslici. Naopak nejhorší úspěšnosti dosáhneme při 7-NN, kdy je prostor vzorů již příliš široký a zvyšuje se tím závislost na mluvčím.

Tab. 5.3: Úspěšnost rozpoznání při proměnných délkách rámce pro MFCC

K-NN	**MFCC – délka rámce [ms]**			
	15	20	25	30
1-NN	68,9%	84,4%	82,2%	85,6%
3-NN	75,6%	90,0%	88,9%	90,0%
5-NN	74,4%	84,4%	81,1%	84,4%
7-NN	73,3%	77,8%	70,0%	76,6%

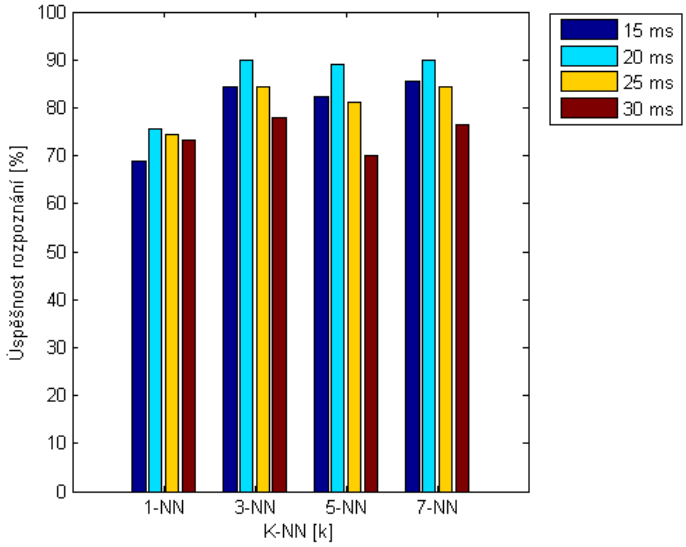


Obr. 5.1: Úspěšnost rozpoznání při proměnných délkách rámce pro MFCC

LPCC

Následně bylo pro každý vzorek z databáze vypočítáno 12 LPCC koeficientů s konstantní délkou překrytí 12 ms a variabilní délkou rámce. Délka rámce se následně měnila od 15 ms do 30 ms po 5 ms. Pak byla vypočtena pomocí DTW vzdálenost referenčních vzorků od vzorků testovacích. Na závěr proběhlo vyhodnocení metodou KNN dle počtu 1, 3, 5 a 7 nejbližších referenčních vzorků a porovnání výsledku rozpoznávání s předem známou skutečně vyslovenou číslicí. Dosažené úspěšnosti rozpoznávání jsou uvedeny v tabulce 5.4.

V grafech (viz. obr. 5.3 a 5.4) a z uvedené tabulky (viz. tab. 5.4) vidíme, že úspěšnosti rozpoznání jsou pro různou délku rámců téměř totožné. Ta nejhorší je pro délku rámce 15 ms a nejlepší pro délku rámce 30 ms. Rozpoznávání dle referenčních vzorů pomocí K-NN nám ukazuje, že jejich počet má na úspěšnost rozpoznání



Obr. 5.2: Úspěšnost rozpoznání při proměnných délkách rámce pro MFCC

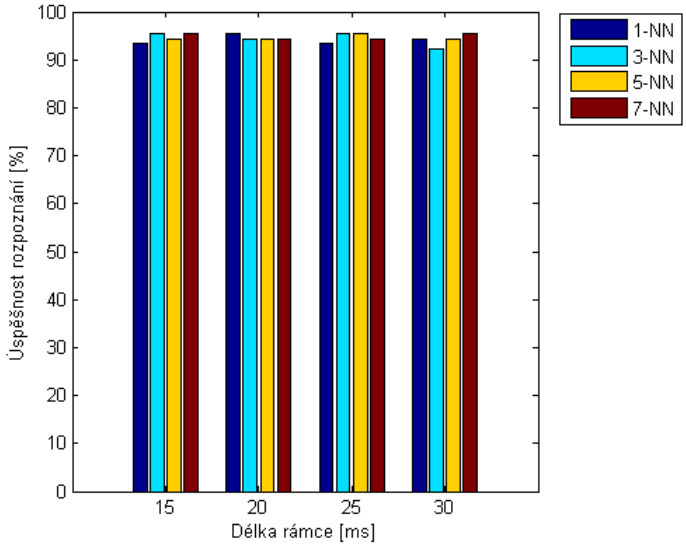
Tab. 5.4: Úspěšnost rozpoznání při proměnných délkách rámce pro LPCC

K-NN	**LPCC – délka rámce [ms]**			
	15	20	25	30
1-NN	93,3%	95,6%	94,4%	95,6%
3-NN	95,6%	94,4%	94,4%	94,4%
5-NN	93,3%	95,6%	95,6%	94,4%
7-NN	94,4%	92,2%	94,4%	95,6%

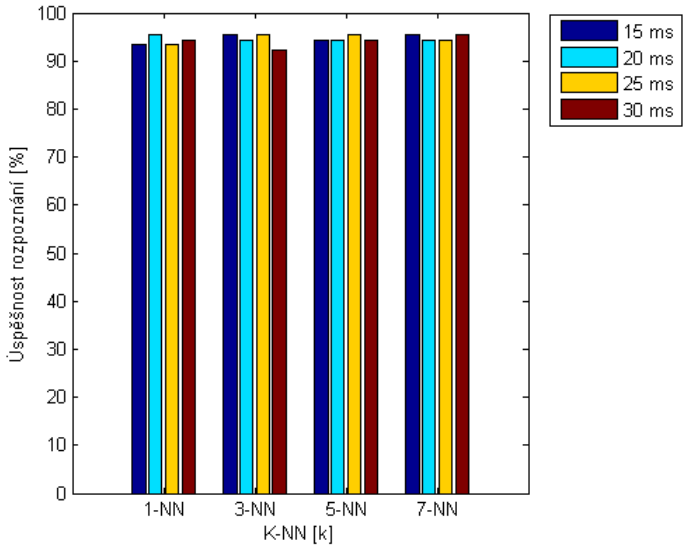
minimální vliv (maximální rozdíl 3,4%).

5.1.3 Porovnání výsledků

Výsledné srovnání výsledků předchozích postupů je uvedeno v tabulce (viz. tab. 5.5) a grafech (viz. obr. 5.5 a 5.6). Vyplývá z ní závislost volby délky rámce a počtu K-NN u MFCC pro ovlivnění úspěšnosti rozeznání na rozdíl od LPCC, kde ve srovnání s MFCC tyto parametry mění úspěšnost rozeznání minimálně. Avšak ani nejlepší volbou parametrů MFCC nedosáhneme úspěšnosti u LPCC.



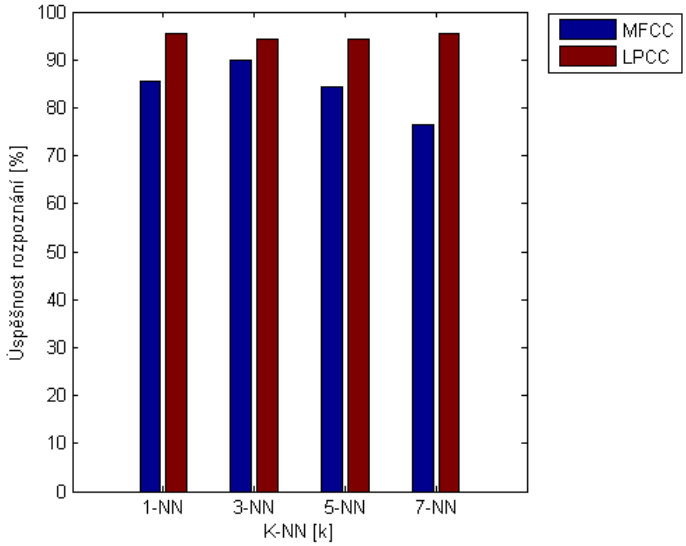
Obr. 5.3: Úspěšnost rozpoznání při proměnných délkách rámce pro LPCC



Obr. 5.4: Úspěšnost rozpoznání při proměnných délkách rámce pro LPCC

Tab. 5.5: Srovnání úspěšnosti rozpoznání při proměnných délkách rámce pro MFCC a LPCC

K-NN	**MFCC – délka rámce [ms]**				**LPCC – délka rámce [ms]**			
	15	20	25	30	15	20	25	30
1-NN	68,9%	84,4%	82,2%	85,6%	93,3%	95,6%	94,4%	95,6%
3-NN	75,6%	90,0%	88,9%	90,0%	95,6%	94,4%	94,4%	94,4%
5-NN	74,4%	84,4%	81,1%	84,4%	93,3%	95,6%	95,6%	94,4%
7-NN	73,3%	77,8%	70,0%	76,6%	94,4%	92,2%	94,4%	95,6%



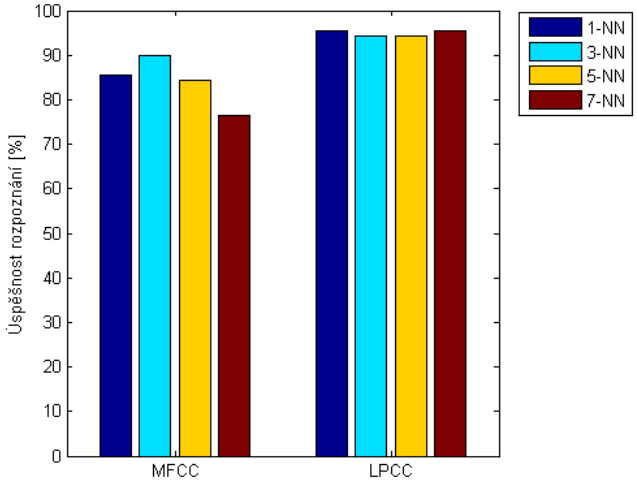
Obr. 5.5: Srovnání úspěšnosti rozpoznání při délce rámce 30 ms pro MFCC a LPCC

5.1.4 Úspěšnost rozpoznávání pro různé počty kepstrálních koeficientů

MFCC

Nyní definujme pro každý vzorek z databáze konstantní délku rámce 20 ms a překrytí 12 ms. Variabilním parametrem pak je počet MFCC koeficientů, konkrétně 6, 12 a 18. Poté byla pomocí DTW určena vzdálenost referenčních vzorků od vzorků testovacích. Na závěr proběhlo vyhodnocení metodou KNN dle počtu 1, 3, 5 a 7 nejbližších referenčních vzorků a porovnání výsledku rozpoznávání s předem známou skutečně vyslovenou číslicí. Dosažené úspěšnosti rozpoznávání jsou uvedeny v tabulce (viz. tab. 5.6).

V následujícím grafu (viz. obr. 5.7 a 5.8) vidíme, že úspěšnosti rozpoznání dosahují



Obr. 5.6: Srovnání úspěšnosti rozpoznání při délce rámce 30 ms pro MFCC a LPCC

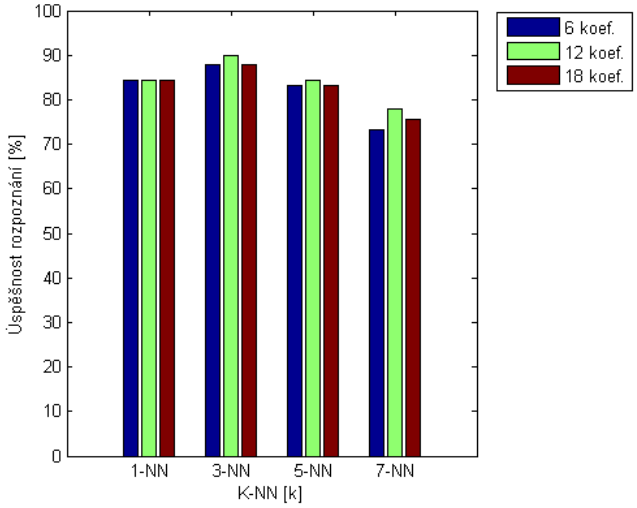
Tab. 5.6: Úspěšnost rozpoznání při různém počtu MFCC koeficientů

K-NN	**MFCC – počet koeficientů [n]**		
	6	12	18
1-NN	84,4%	84,4%	84,4%
3-NN	87,8%	90,0%	87,8%
5-NN	83,3%	84,4%	83,3%
7-NN	73,3%	77,8%	75,6%

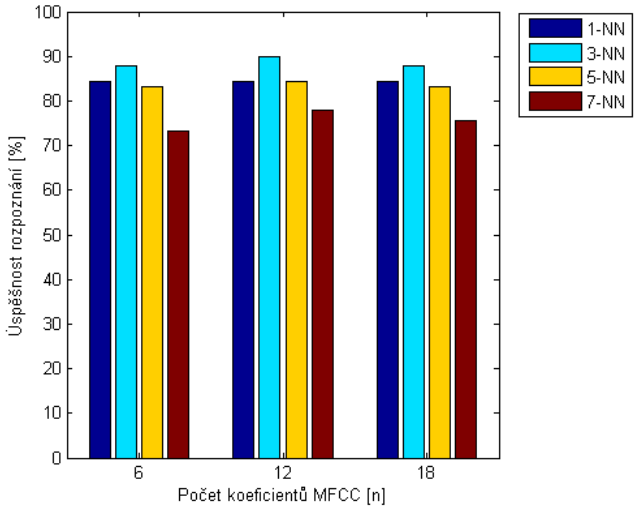
nejnižších hodnot pro 6 a 18 MFCC koeficientů a jsou téměř totožné. Naopak nejlépe dopadlo rozpoznání pro 12 MFCC koeficientů. Čímž bylo potvrzeno, že optimální počet koeficientů je 10 až 13. Rozpoznávání dle referenčních vzorů pomocí K-NN nám v tomto případě ukazuje, že nejvyšší úspěšnosti dosáhneme při využití 3-NN, kdy nám 3 čísla umožní lépe určit danou číslici. Naopak nejhorší úspěšnosti dosáhneme při 7-NN, kdy je prostor vzorů již příliš široký a zvyšuje se tím závislost na mluvčím.

LPCC

V dalším kroku definujme pro každý vzorek z databáze konstantní délku rámce 20 ms a překrytí 12 ms. Variabilním parametrem bude počet LPCC koeficientů, konkrétně 6, 12 a 18. Poté byla pomocí DTW určena vzdálenost referenčních vzorků od vzorků testovacích. Na závěr proběhlo vyhodnocení metodou KNN dle počtu 1, 3,



Obr. 5.7: Úspěšnost rozpoznání při různém počtu MFCC koeficientů



Obr. 5.8: Úspěšnost rozpoznání při různém počtu MFCC koeficientů

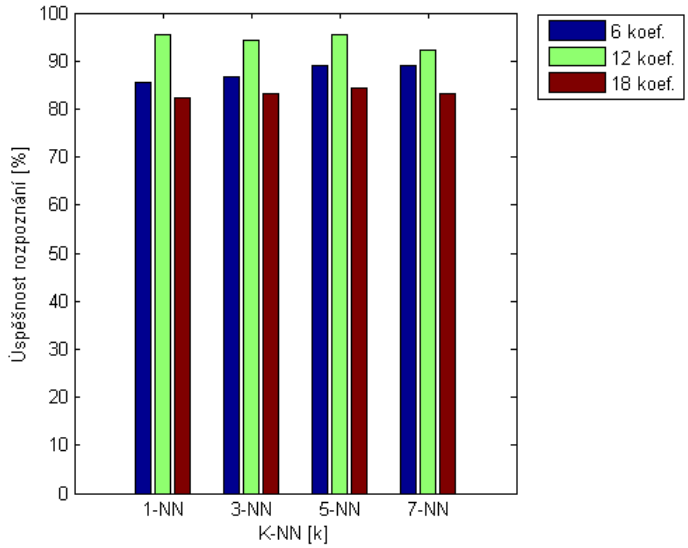
5 a 7 nejbližších referenčních vzorků a porovnání výsledku rozpoznávání s předem známou skutečně vyslovenou číslicí. Dosažené úspěšnosti rozpoznávání jsou uvedeny v tabulce (viz. tab. 5.7).

Z grafů (viz. obr. 5.9 a 5.10) vidíme, že úspěšnost rozpoznání pro 18 LPCC koeficientů dosahuje nejnižších hodnot, následována s malým odstupem úspěšností pro 6 LPCC koeficientů. Naopak nejlépe dopadlo rozpoznání pro 12 MFCC koeficientů, čímž bylo potvrzeno, že se jedná o optimální počet koeficientů. Rozpoznávání dle referenčních vzorů pomocí KNN nám v tomto případě ukazuje, že nejvyšší úspěš-

Tab. 5.7: Úspěšnost rozpoznání při různém počtu LPCC koeficientů

K-NN	**LPCC – počet koeficientů [n]**		
	6	12	18
1-NN	85,6%	95,6%	82,2%
3-NN	86,7%	94,4%	83,3%
5-NN	88,9%	95,6%	84,4%
7-NN	88,9%	92,2%	83,3%

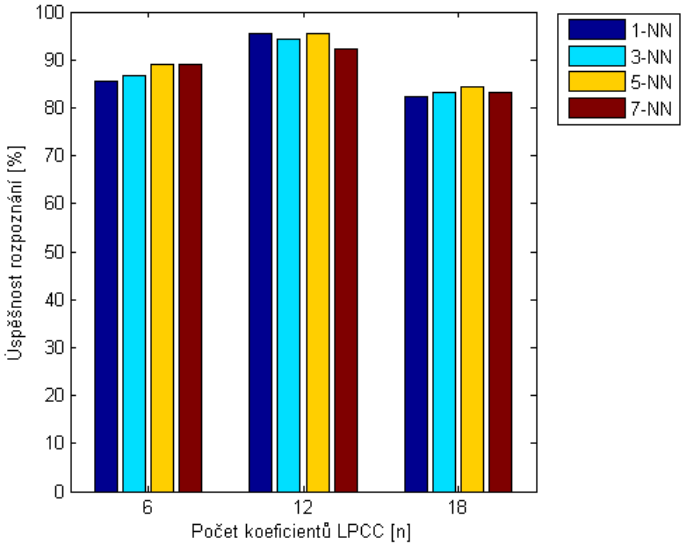
nosti dosáhneme při využití 5-NN, kdy nám 5 čísel umožní lépe určit danou číslici. Vzhledem k hodnotám úspěšnosti rozeznání je však vliv volby rozsahu menší a to do rozsahu maximálně 3,4%.



Obr. 5.9: Úspěšnost rozpoznání při různém počtu LPCC koeficientů

5.1.5 Porovnání výsledků

Výsledné srovnání výsledků předchozích postupů je uvedeno v tabulce (viz. tab. 5.7) a grafech (viz. obr. 5.11 a 5.12). Vyplývá z ní závislost volby počtu koeficientů pro MFCC a LPCC pro ovlivnění úspěšnosti rozeznání a zároveň ji lze u MFCC zvýšit významně volbou K-NN, u LPCC pak méně. Volba nesprávného počtu parametrů u MFCC i LPCC může vést k velkým rozdílům v úspěšnosti.



Obr. 5.10: Úspěšnost rozpoznání při různém počtu LPCC koeficientů

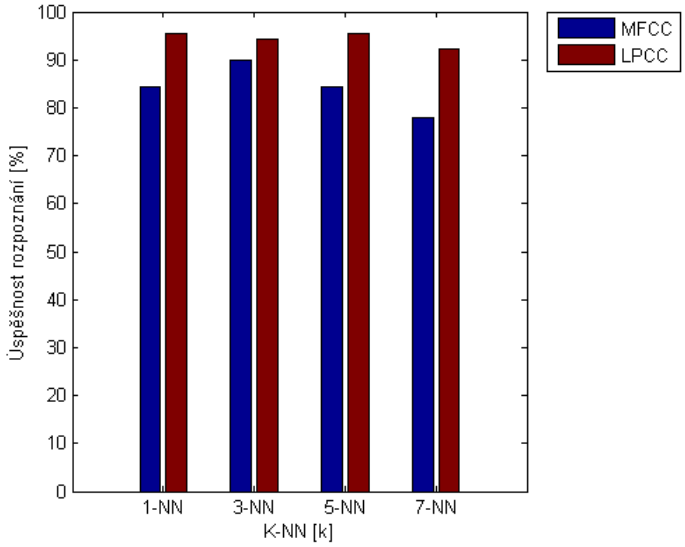
Tab. 5.8: Srovnání úspěšnosti rozpoznání MFCC a LPCC pro různý počet koeficientů

K-NN	**MFCC – počet koeficientů [n]**			**LPCC – počet koeficientů [n]**		
	6	12	18	6	12	18
1-NN	84,4%	84,4%	84,4%	85,6%	95,6%	82,2%
3-NN	87,8%	90,0%	87,8%	86,7%	94,4%	83,3%
5-NN	83,3%	84,4%	83,3%	88,9%	95,6%	84,4%
7-NN	73,3%	77,8%	75,6%	88,9%	92,2%	83,3%

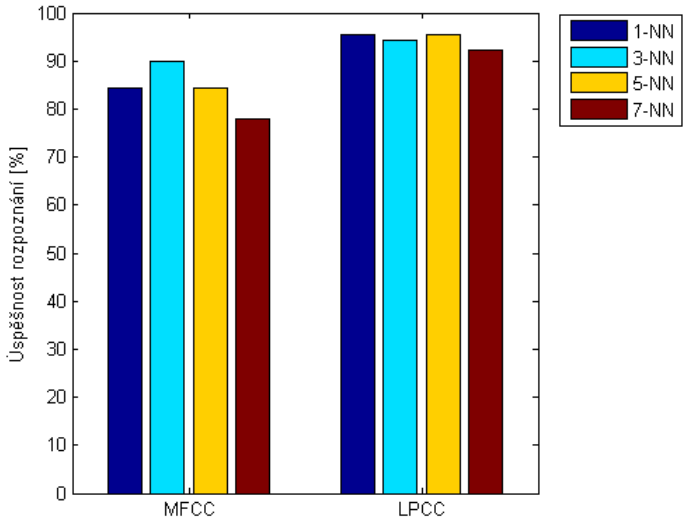
5.2 HMM

5.2.1 Tvorba modelů

Nyní přistoupíme k tvorbě systému rozpoznání izolovaných slov s využití statistických metod rozpoznání řeči založenému na HMM v prostředí Matlab s využitím HTK [4]. Zaměříme se na tvorbu systémů nezávislých na mluvčím zvlášť pro ženy a muže. Akustický model bude založen na shlukování trifonémů s fonetickými rozhodovacími stromy. Jazykový model bude řešen přes kombinaci gramatiky a výslovnostního slovníku. Pro trénink bude využit Baum-Welchův algoritmus, pro rozpoznání pak Viterbiho algoritmus. Výsledný systém nám umožní snadnou změnu množiny rozpoznávaných slov. Dále se budeme zabývat vývojem systému pro muže, neboť



Obr. 5.11: Srovnání úspěšnosti rozpoznání MFCC a LPCC pro 12 koeficientů



Obr. 5.12: Srovnání úspěšnosti rozpoznání MFCC a LPCC pro 12 koeficientů

postup pro ženy je stejný. Následuje nastíněný postup, konkrétní příkazy s parametry jsou uvedeny ve zdrojových kódech na přiloženém DVD.
Databáze Timit obsahuje zvlášť data určená pro trénování a testování, které rozdělíme podle pohlaví. Dále si upravíme výslovnostní slovník databáze Timit pro práci s HTK. Následně přistoupíme k redukci fonémové sady z originálních 61 na 47 na základě transkripce. Redukci lze provést i na 39 fonémů [5]. Vytvoříme potřebné

trénovací skripty a fonémové přepisy jednotlivých vět. Následně přistoupíme k extrakci příznaků. Pro parametrizaci byly s přihlédnutím k článku [6] zvoleny MFCC s logaritmem energie a odhadem první a druhé diference. Sestavíme 5-ti stavové levo-pravé HMM modely pro jednotlivé fonémy. Ty sloučíme a inicializujeme. Poté výsledný model několikrát přetrénujeme, zavedeme model pauzy, opět přetrénujeme. Svážeme fonémy do trifonémů a přetrénujeme. Na závěr vytvoříme fonetické rozhodovací stromy a opět přetrénujeme. Nyní již zbývá pouze definovat jednoduchou gramatiku jako množinu rozpoznávaných slov a vybrat testovací data, ze kterých je třeba extrahovat příznaky. A na závěr provést rozpoznání a analýzu výsledků.
Takto vytvořený systém bude testován na dvou vytvořených databázích, které vznikly na základě požadavků následného využití pro hlasové ovládání v kokpitu letadla.

5.2.2 Tvorba databáze Basic

Databáze Basic byla vytvořena za účelem získání zvukových náhrávek pro vývoj a vyhodnocení pro systém KWD. Nahrávání čtené řeči v anglickém jazyce probíhalo v zasedací místnosti.

Popis nahrávání

Šum v zasedací místnosti byl 40 dB, jeho úroveň byla měřena hlukoměrem CEM DT-8852 s rozsahem 30 – 130 dB a odchylkou měření ± 1,4 dB. Pro nahrávání byla zvolena sluchátka Plantronics Blackwire C320M.

Popis databáze

Databáze Basic obsahuje celkem 100 nahrávek. Každá z 10 frází je řečena každým z 10 mluvčích ze 2 států: Česko a Slovensko. Rozdělení podle pohlaví je 6 mužů a 4 ženy. Více informací je v tabulkách (viz. tab. 5.9, tab. 5.10)

Tab. 5.9: Rozdělení mluvčích

Státní příslušnost	**Muži**		**Ženy**		**Celkem**	
Česko	4	67%	2	33%	6	100%
Slovensko	2	50%	2	50%	4	100%
Celkem	6	60%	4	40%	10	100%

Po konzultaci s piloty bylo vybráno 10 frází, aby pokryly požadavky pro následné využití v kokpitu. Seznam frází:

1. Runway blocked

Tab. 5.10: Údaje o jednotlivých mluvčích

Mluvčí	Stát	Věková skupina	Pohlaví
E	Slovensko	21-25	žena
G	Česko	21-25	žena
J	Česko	26-30	žena
M	Slovensko	26-30	muž
R	Slovensko	26-30	muž
S	Slovensko	21-25	žena
T	Česko	21-25	muž
U	Česko	21-25	žena
V	Česko	31-35	muž
Z	Česko	26-30	muž

2. Start up not approved
3. Ready for take off
4. Climb to flight level
5. Continue present heading
6. After departure turn left
7. Make full stop
8. Hold your position
9. Report current weather
10. Maintain own separation

Databáze Basic obsahuje soubory uložené ve formátu *.wav nahrané jako mono, 16 000 Hz, PCM signed 16 bit, 256 kbps. Adresářová struktura a soubory v nich jsou organizovány následovně:
/<Gender>/<File_type>
kde,
Gender == Male | Female
File_type == <Phrase>_<Name>
kde,
Phrase == 1 | 2 | ... | 9 | 10
Name == E | G | ... | V | Z

5.2.3 Testování a analýza

Pro účely evaluace byla použita celá databáze Basic, která byla rozdělena podle pohlaví mluvčích. Tím vznikly dvě množiny po 50 vzorcích (5 mluvčích * 10 frází),

které byly následně otestovány pro na ně upravenou gramatiku. Výsledky úspěšnosti rozpoznání jsou uvedeny v tabulce (viz. tab. 5.11).

Tab. 5.11: Úspěšnost rozpoznání frází z databáze Basic

Model	**Úspěšnost rozpoznání**
Muži	100%
Ženy	100%

Vzhledem k 100% úspěšnosti rozpoznání není analýza nutná.

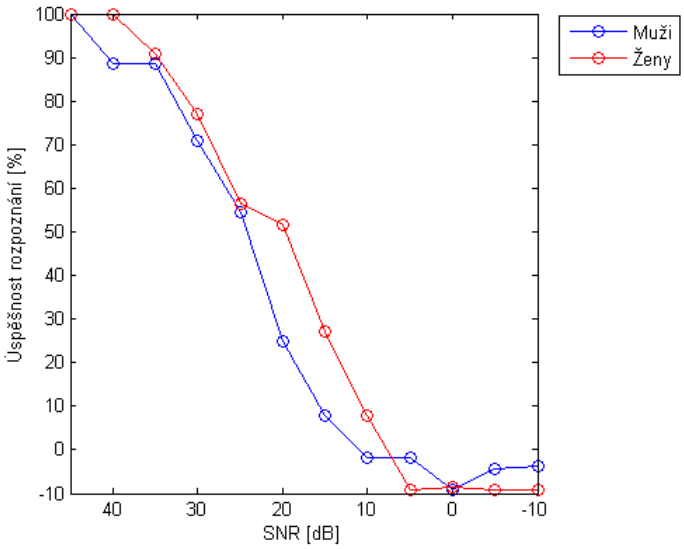
5.2.4 Testování a analýza uměle zašumělých dat

Dalším krokem je přidání bílého gaussovského šumu do nahrávek a změna SNR, aby bylo částečně simulováno prostředí kokpitu dopravního letadla [7]. Výsledky úspěšnosti rozpoznání jsou uvedeny v tabulce (viz. tab. 5.12).

Tab. 5.12: Úspěšnost rozpoznání frází ze zašumělé databáze Basic

SNR [dB]	**Model**	
	Muži	Ženy
40	88,48%	100%
35	88,48%	90,91%
30	70,91%	76,97%
25	54,55%	56,36%
20	24,85%	51,52%
15	7,88%	27,27%
10	-1,82%	7,88%
5	-1,82%	-9,09%
0	-9,09%	-8,48%
-5	-4,24%	-9,09%
-10	-3,64%	-9,09%

Výsledky pro větší přehlednost znázorníme v grafu (obr. 5.13), kde přidáme bod SNR = $+\infty$ dB, který odpovídá originálním nahrávkám.
Z grafu vyplývá předpokládaný průběh, kdy se snižujícím se SNR klesá úroveň rozpoznání. Z výsledků ženského modelu jsou patrné dvě zlomové úrovně. První u hodnoty SNR 30 dB, kde dochází k poklesu na spodní hranici přijatelnosti úrovně rozpoznání. U druhé hodnoty SNR 5 dB je již úroveň rozpoznání záporná, tudíž je použití za



Obr. 5.13: Úspěšnost rozpoznání frází ze zašumělé databáze Basic

daných podmínek de facto vyloučeno. Na rozdíl od ženského modelu nemá křivka mužského modelu tak jednoznačný průběh a druhý zlomový bod se posouvá již na hodnotu SNR 10 dB. Za zmínku pak stojí nepředpokládaný výsledek při SNR 40 dB. Po podrobnější analýze vyplynulo možné vysvětlení v kombinaci horší výslovnosti a přeci jen malém počtu mluvčích. Závěrem lze konstatovat, že vzhledem k dosaženým výsledkům pro danou databázi nerodilých mluvčí jsou výsledné křivky téměř v souladu s předpokladem.

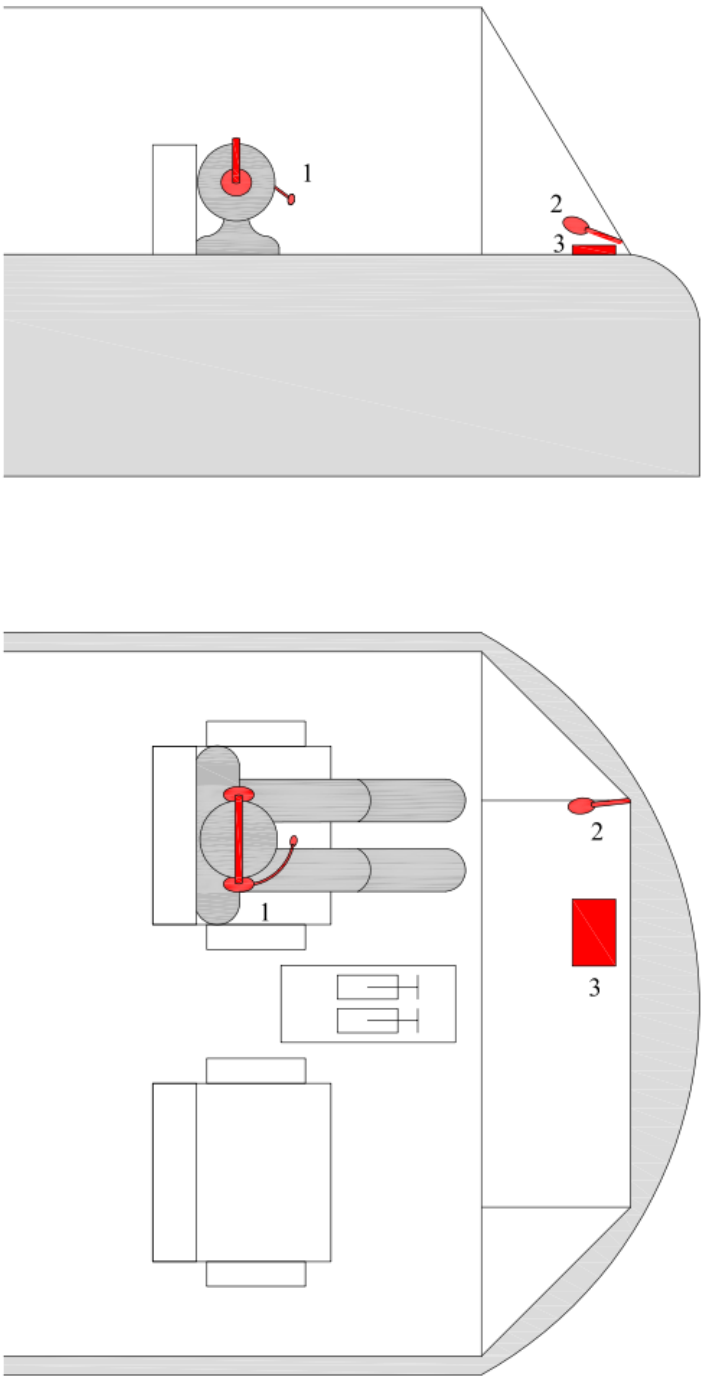
5.2.5 Tvorba databáze Speech4EFB

Databáze Speech4EFB byla vytvořena za účelem získání zvukových náhrávek pro vývoj a vyhodnocení pro systém KWD. Nahrávání čtené řeči v anglickém jazyce probíhalo v leteckém simulátoru firmy Honeywell.

Popis nahrávání

Na základě specifikací byl pro simulaci reálného prostředí využit softwarový model letounu Airbus A320. Ten byl upraven, aby úroveň šumu v kokpitu na základě [7] byla 78 dB. Úroveň šumu byla měřena hlukoměrem CEM DT-8852 s rozsahem 30 – 130 dB a odchylkou měření ± 1,4 dB. Pro nahrávání byla zvolena tři různá zařízení. Na hlavu mluvčího byla umístěna sluchátka Plantronics Blackwire C620M a na palubní desce byly ve vzdálenosti 75 cm od mluvčího umístěny směrový supercardioid mikrofon Rode Videomic a tablet Acer Iconia s vestavěným mikrofonem

(obr. 5.14).



Obr. 5.14: Rozmístění nahrávacích zařízení v leteckém simulátoru: 1 - sluchátka, 2 - mikrofon, 3 - tablet

Popis databáze

Databáze Speech4EFB obsahuje celkem 2372 nahrávek. Každá z 18 frází je řečena dvakrát každým z 22 mluvčích z 5 států: Česko, Indie, Izrael, Mexiko a Slovensko. Rozdělení podle pohlaví je 8 žen a 14 mužů. Více informací je v tabulkách (viz. tab. 5.13, tab. 5.14)

Tab. 5.13: Rozdělení mluvčích

Státní příslušnost	**Muži**		**Ženy**		**Celkem**	
Česko	9	60%	6	40%	15	100%
Slovensko	2	50%	2	50%	4	100%
Mexiko	1	100%	0	0%	1	100%
Izrael	1	100%	0	0%	1	100%
Indie	1	100%	0	0%	1	100%
Celkem	14	64%	8	36%	22	100%

Celkem 18 frází bylo vybráno, aby pokryli požadavky pro využití v kokpitu. Seznam frází:

1. Turbulence situation
2. Load company route
3. Center aircraft
4. Center on destination
5. Center on alternate
6. Follow track
7. Update data
8. Clear screen
9. Increase range
10. Display terrain
11. Hide terrain
12. Flight level two four zero
13. Flight level three six zero
14. Flight level four eight zero
15. Show me weather for destination
16. Show me weather for alternate
17. Display relevant sigmets
18. Display relevant pireps

Databáze Speech4EFB obsahuje soubory uložené ve formátu *.wav nahrané jako mono, 16 000 Hz, PCM signed 16 bit, 256 kbps. Adresářová struktura a soubory

Tab. 5.14: Údaje o jednotlivých mluvčích

Mluvčí	**Stát**	**Věková skupina**	**Pohlaví**
AC	Česko	26-30	žena
CC	Mexiko	26-30	muž
DK	Česko	26-30	muž
DN	Česko	26-30	muž
GB	Česko	21-25	žena
JC	Česko	26-30	žena
JS	Česko	31-35	muž
KD	Česko	21-25	žena
LM	Slovensko	21-25	žena
MH	Slovensko	26-30	muž
MM	Slovensko	26-30	žena
MN	Česko	21-25	muž
MO	Česko	31-35	muž
PT	Česko	31-35	muž
RL	Slovensko	26-30	muž
TK	Izrael	31-35	muž
TB	Česko	21-25	muž
US	Indie	26-30	muž
VH	Česko	21-25	žena
VP	Česko	26-30	muž
VV	Česko	26-30	muž
ZM	Česko	21-25	žena

v nich jsou organizovány následovně:
/<Device>/<Gender>/<File_type>
kde,
Device == Headphones | Microphone | Tablet
Gender == Male | Female
File$_type$ ==$< Name > \_ < Device > \_ < Phrase > \_ < Number >$
kde,
$Name == AC \mid CC \mid \ldots \mid VV \mid ZM$
$Device == M \mid H \mid T1$
$Phrase == 1 \mid 2 \mid \ldots \mid 17 \mid 18$
$Number == 1 \mid 2$

5.2.6 Testování a analýza

Databáze Speech4EFB obsahuje i celé věty, proto byly pro účely evaluace vybrány následující krátké fráze:

1. Turbulence situation
3. Center aircraft
4. Center on destination
6. Follow track
7. Update data
8. Clear screen
9. Increase range
10. Display terrain
11. Hide terrain

Ty byly rozděleny podle pohlaví mluvčích na dvě množiny vzorků, které byly následně otestovány pro na ně upravenou gramatiku. Konkrétně bylo pro každé zařízení 292 mužských vzorků (14 mluvčích * 9 frází * 2 vzorky) a 144 ženských vzorků (8 mluvčích * 9 frází * 2 vzorky). Výsledky úspěšnosti rozpoznání jsou uvedeny v tabulce (viz. tab. 5.15).

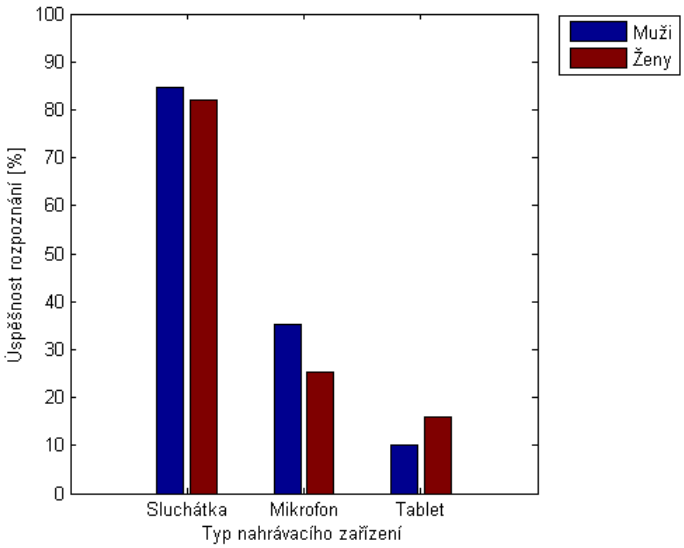
Tab. 5.15: Úspěšnost rozpoznání frází z databáze Speech4EFB

Model	**Úspěšnost rozpoznání**		
	Sluchátka	Mikrofon	Tablet
Muži	84,59%	35,15%	25,13%
Ženy	81,91%	25,33%	15,79%

Výsledky pro větší přehlednost znázorníme v grafu (obr. 5.15).
Z výsledků plyne, že pro další vývoj a testování se je třeba zaměřit na využití kvalitních sluchátek, které částečně eliminují šum na pozadí. Ani kvalitní směrový mikrofon ani tablet nedosahují přijatelných výsledků. Jedním z důvodů je vzdálenost od mluvčího, kterou ovšem nelze měnit. Dalším je pak určitá deformovanost hlasové nahrávky ze směrového mikrofonu. V případě tabletu se jedná o problém levného zabudovaného všesměrového mikrofonu, kde je již šum pozadí příliš velký.

5.2.7 Testování a analýza uměle zašumělých dat

Z předchozích výsledků plyne, že další vývoj bude vhodné zaměřit na kvalitní sluchátka. V praxi se ovšem často mohou vyskytovat i jiná méně kvalitní. Je proto důležité ověřit další možnosti. Dalším krokem je tudíž přidání bílého gaussovského



Obr. 5.15: Úspěšnost rozpoznání frází z databáze Speech4EFB

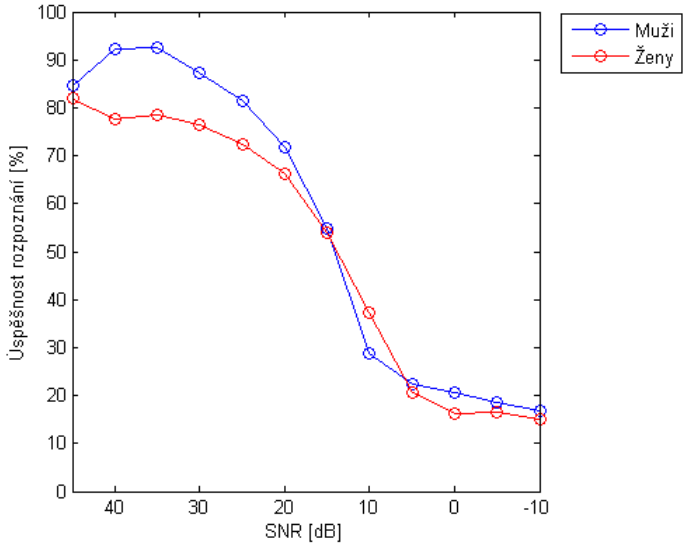
šumu do nahrávek ze sluchátek a změna SNR, aby bylo částečně simulováno využití sluchátek nižší kvality v prostředí kokpitu dopravního letadla. Výsledky úspěšnosti rozpoznání jsou uvedeny v tabulce (viz. tab. 5.16).

Tab. 5.16: Úspěšnost rozpoznání frází ze zašumělé databáze Speech4EFB pro nahrávky ze sluchátek

SNR [dB]	**Model**	
	Muži	Ženy
40	92,11%	77,63%
35	92,67%	78,62%
30	87,41%	76,32%
25	81,57%	72,37%
20	71,80%	66,12%
15	54,89%	53,95%
10	28,68%	37,17%
5	22,37%	20,72%
0	20,49%	16,12%
-5	18,61%	16,45%
-10	16,92%	15,13%

Výsledky pro větší přehlednost znázorníme v grafu (obr. 5.16), kde přidáme bod

SNR = $+\infty$ dB, který odpovídá originálním nahrávkám.



Obr. 5.16: Úspěšnost rozpoznání frází ze zašumělé databáze Speech4EFB pro nahrávky ze sluchátek

Z grafu vyplývá předpokládaný průběh, kdy se snižujícím se SNR klesá úroveň rozpoznání. Z výsledků ženského modelu je patrná jedna zlomová úroveň u hodnoty SNR 25 dB, kde dochází k poklesu na spodní hranici přijatelnosti úrovně rozpoznání. Mnohem zajímavější z hlediska analýzy je křivka mužského modelu. Zlom v úspěšnosti rozpoznání se zde nachází u SNR 20 dB. Na rozdíl od ženského modelu nemá křivka tak jednoznačný průběh. Za zmínku pak stojí růst úspěšnosti rozpoznání při SNR více než 30 dB, kde je úroveň úspěšnosti rozpoznání vyšší než u originálních nahrávek. Toto je pravděpodobně způsobeno změnou originálního „vyhlazeného“ signálu na signál s minimálním šumem, ale dostatečným například pro lepší rozlišení mezi blízkými frázemi „hide terrain“ a „display terrain“. Závěrem lze konstatovat, že vzhledem k dosaženým výsledkům pro robustnější databázi nerodilých mluvčích z 5-ti různých států jsou výsledné křivky téměř v souladu s předpokladem.

6 ZÁVĚR

V rámci diplomové práce byly v teoretické části řešeny metody analýzy a rozpoznávání řečových signálů.
Praktická část byla realizována ve spolupráci s firmou Honeywell sekcí Aerospace, jejíž požadavky byly v práci zohledněny.
Nejprve byl vytvořen systém pro rozpoznání izolovaných slov v prostředí Matlab založený na DTW a KNN. Tento byl následně testován na databázi Aurora 5 verze c6 pro rozpoznání číslic 1-9. Pro každý vybraný vzorek z databáze vypočítáno 12 MFCC a 12 LPCC koeficientů s konstantní délkou překrytí 12 ms a variabilní délkou rámce. Délka rámce se měnila od 15 ms do 30 ms po 5 ms. Pak byla vypočtena pomocí DTW vzdálenost referenčních vzorků od vzorků testovacích. Na závěr proběhlo vyhodnocení metodou KNN dle počtu 1, 3, 5 a 7 nejbližších referenčních vzorků a porovnání výsledku rozpoznávání s předem známou skutečně vyslovenou číslicí. Výsledné srovnání výsledků předchozího postupu je uvedeno v tabulce (viz. tab. 5.5) a grafech (viz. obr. 5.5 a 5.6). Vyplývá z ní závislost volby délky rámce a počtu K-NN u MFCC pro ovlivnění úspěšnosti rozeznání na rozdíl od LPCC, kde ve srovnání s MFCC tyto parametry mění úspěšnost rozeznání minimálně. Avšak ani nejlepší volbou parametrů MFCC nedosáhneme úspěšnosti u LPCC. Následně jsme definovali pro každý vzorek z databáze konstantní délku rámce 20 ms a překrytí 12 ms. Variabilním parametrem byl počet MFCC koeficientů, konkrétně 6, 12 a 18. Poté byla pomocí DTW určena vzdálenost referenčních vzorků od vzorků testovacích. Na závěr proběhlo vyhodnocení metodou KNN dle počtu 1, 3, 5 a 7 nejbližších referenčních vzorků a porovnání výsledku rozpoznávání s předem známou skutečně vyslovenou číslicí. Výsledné srovnání výsledků předchozích postupů je uvedeno v tabulce (viz. tab. 5.7) a grafech (viz. obr. 5.11 a 5.12). Vyplývá z ní závislost volby počtu koeficientů pro MFCC a LPCC pro ovlivnění úspěšnosti rozeznání a zároveň ji lze u MFCC zvýšit významně volbou K-NN, u LPCC pak méně. Volba nesprávného počtu parametrů u MFCC i LPCC může vést k velkým rozdílům v úspěšnosti.
Dále přejdeme k využití statistických metod rozpoznání řeči a systémům nezávislých na mluvčím zvlášť pro ženy a muže založenému na HMM v prostředí Matlab s využitím HTK. Poté k tvorbě vlastní databáze Basic tvořenou 100 nahrávkami od 10 různých mluvčích, následnému testování a analýze zohledňující různé úrovně šumu s přihlédnutím ke specifickým podmínkám v kokpitu. Výsledky jsou uvedeny v tabulkách (viz. tab. 5.11 a 5.12). Jasnější pohled dává grafické vyjádření, kde z grafu (viz. obr. 5.13) vyplývá předpokládaný průběh, kdy se snižujícím se SNR klesá úroveň rozpoznání. Z výsledků ženského modelu jsou patrné dvě zlomové úrovně. První u hodnoty SNR 30 dB, kde dochází k poklesu na spodní hranici přijatelnosti úrovně rozpoznání. U druhé hodnoty SNR 5 dB je již úroveň rozpoznání záporná,

tudíž je použití za daných podmínek de facto vyloučeno. Na rozdíl od ženského modelu nemá křivka tak jednoznačný průběh a druhý zlomový bod se posouvá již na hodnotu SNR 10 dB. Za zmínku pak stojí nepředpokládaný výsledek při SNR 40 dB. Po podrobnější analýze vyplynulo možné vysvětlení v kombinaci horší výslovnosti a přeci jen malém počtu mluvčích. Závěrem lze konstatovat, že vzhledem k dosaženým výsledkům pro danou databázi nerodilých mluvčí jsou výsledné křivky téměř v souladu s předpokladem. Dále byla vytvořena robustní databáze Speech4EFB, která obsahuje celkem 2372 nahrávek od 22 mluvčích a byla vytvořena v leteckém simulátoru speciálně pro podmínky využití v leteckém průmyslu. Z testování a analýzy, uvedené v tabulce (viz. tab. 5.15) a grafu (obr. 5.15), vyplynulo, že pro další vývoj a testování se je třeba zaměřit na využití kvalitních sluchátek, které částečně eliminují šum na pozadí. Ani kvalitní směrový mikrofon ani tablet nedosahují přijatelných výsledků. Jedním z důvodů je vzdálenost od mluvčího, kterou ovšem nelze měnit. Dalším je pak určitá deformovanost hlasové nahrávky ze směrového mikrofonu. V případě tabletu se jedná o problém levného zabudovaného všesměrového mikrofonu, kde již šum pozadí příliš velký. V praxi se ovšem často mohou vyskytovat i jiná méně kvalitní. Je proto důležité ověřit další možnosti. Po přidání šumu do nahrávek ze sluchátek po evaluaci vyplynuly výsledky uvedené v tabulce (viz. tab. 5.16) a grafu (obr. 5.16). Vyplývá z nich předpokládaný průběh, kdy se snižujícím se SNR klesá úroveň rozpoznání. Z výsledků ženského modelu je patrná jedna zlomová úroveň u hodnoty SNR 25 dB, kde dochází k poklesu na spodní hranici přijatelnosti úrovně rozpoznání. Mnohem zajímavější z hlediska analýzy je křivka mužského modelu. Zlom v úspěšnosti rozpoznání se zde nachází u SNR 20 dB. Na rozdíl od ženského modelu nemá křivka tak jednoznačný průběh. Za zmínku pak stojí růst úspěšnosti rozpoznání při SNR více než 30 dB n úroveň vyšší než u originálních nahrávek. Toto je pravděpodobně způsobeno změnou originálního „vyhlazeného“ signálu na signál s minimálním šumem, ale dostatečným například pro lepší rozlišení mezi blízkými frázemi „hide terrain“ a „display terrain“. Závěrem lze konstatovat, že vzhledem k dosaženým výsledkům pro robustnější databázi nerodilých mluvčích z 5-ti různých států jsou výsledné křivky téměř v souladu s předpokladem.

LITERATURA

[1] PSUTKA, J., L. MÜLLER, J. MATOUŠEK a V. RADOVÁ. *Mluvíme s počítačem česky.* 1. vyd. Praha: Academia, 2006. ISBN 80-200-1309-1.

[2] GAROFOLO, J., et al. *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus.* CD-ROM: NIST Speech Disc 1–1.1. Gaithersburg: U. S. Department of Commerce, 1993.

[3] Melchor, J. A. and J. P. Spanyers. *MEDICAL FACTS FOR PILOTS.* FAA Civil Aerospace Medical Institute, 2008. Dostupné z: http://www.faa.gov/pilots/safety/pilotsafetybrochures/media/hearing_brochure.pdf

[4] HTK, *Speech Recognition Toolkit.* Dostupné z: http://htk.eng.cam.ac.uk/

[5] Lopes, C. and F. Perdigao. *Phone Recognition on the TIMIT Database.* Universidade de Coimbra, Portugal, 2011. Dostupné z: http://cdn.intechopen.com/pdfs/15948/InTech-Phoneme_recognition_on_the_timit_database.pdf

[6] Mandal, A., et. al. *Strategies for High Accuracy Keyword Detection in Noisy Channels.* Speech Technology and Research Laboratory, SRI International, USA, 2013. Dostupné z: http://www.sri.com/sites/default/files/publications/strategies_for_high-accuracy_keyword_detection_in_noisy_channels-final.pdf

[7] Lower, M. C. and M. Bagshaw. *NOISE LEVELS AND COMMUNICATIONS ON THE FLIGHT DECKS OF CIVIL AIRCRAFT.* University of Southampton, UK, 1996. Dostupné z: http://www.isvr.co.uk/reprints/Inter96air.pdf

[8] HOLMES, J. and W. HOLMES. *Speech synthesis and recognition.* 2nd ed. London: Taylor Francis, 2001. ISBN 0-7484-0856-8.

[9] MCLOUGHLIN, I. *Applied speech and audio processing: with Matlab examples.* 1st pub. Cambridge: Cambridge University Press, 2009. ISBN 978-0-521-51954-0.

[10] RABINER, L. and B.-H. JUANG. *Fundamentals of speech recognition.* Upper Saddle River: Prentice Hall, 1993. ISBN 0-13-015157-2.

[11] ČERNOCKÝ, J. *Zpracování řečových signálů – studijní opora.* VUT FIT v Brně, 2006.

[12] JURAFSKY, D. and J. H. MARTIN. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.* 2nd ed. Upper Saddle River: Pearson Education, 2008. ISBN 978-0-13-187321-6.

[13] KESHET, J. and S. BENGIO. *Automatic speech and speaker recognition: large magin and kernel methods.* 1st ed. Chichester: John Wiley Sons, 2009. ISBN 978-0-470-69683-5.

[14] MADISETTI, V. *The digital signal processing handbook.* 2nd ed. Boca Raton: CRC Press, 2010. ISBN 978-1-4200-4608-3.

[15] AURORA, *Speech recognition experimental framework.* Dostupné z: http://aurora.hsnr.de.

[16] MATLAB, *The language of technical computing.* Dostupné z: http://www.mathworks.com/products/matlab.

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

a_i koeficienty číslicového filtru

$A(z)$ polynom Q-tého řádu

B počet bitů v binárním kódu

B_{mw} celková šířka přenášeného pásma [mel]

B_w celková šířka přenášeného pásma [Hz]

$c(k)$ kepstrální koeficienty LPC

$c_m(j)$ koeficienty MFCC

$c(n)$ kepstrální koeficienty

D_* charakteristický systém

DCT diskrétní kosinová transformace – Discrete Cosine Transform

DPCM diferenční pulsní kódová modulace – Differential Pulse-Code Modulation

DTW dynamické borcení času – Dynamic Time Warping

EFB elektronická letecká taška – Electronic Flight Bag

f frekvence v lineární škále

FFT rychlá Fourierova transformace – Fast Fourier Transform

f_m frekvence v nelineární melovské škále

F_m horní hranice frekvenčního pásma signálu

F_v frekvence vzorkování

G koeficient zesílení

HMM skryté Markovovy modely – Hidden Markov Model

$H(z)$ přenosová funkce modelu

IDFT zpětná diskrétní Fourierova transformace – Inverse Discrete Fourier Transform

K počet nejbližších sousedů

KNN k nejbližších sousedů – k Nearest Neighbor

KWD detekce klíčových slov – Keyword Detection

LPC lineární prediktivní kódování – Linear Predictive Coding

LPCC kepstrální koeficienty lineárního prediktivního kódování – Linear Predictive Cepstral Coefficients

M počet koeficientů MFCC

M^* počet pásem melovského pásmového filtru

MFCC melovské kepstrální koeficienty – Mel-Frequency Cepstral Coefficients

PCM pulsní kódová modulace – Pulse-Code Modulation

PIREP zpráva pilota – Pilot Report

PLP perceptivní lineární prediktivní kódování – Perceptual Linear Predictive coding

Q řád modelu LPC

Q^* počet kepstrálních koeficientů LPC

SIGMET významná meteorologická situace – Significant Meteorological Information

$s(k)$ k-tý vzorek signálu

S_{max} maximální úroveň vzorkovaného signálu

s_n signál vyjádřený diskrétní posloupností

SNR odstup signálu od šumu – Signal to Noise Ratio

$s(t)$ časově spojitý signál

T perioda vzorkování

$u(k)$ buzení

$x(n)$ diskrétní posloupnost

$\hat{x}(n)$ komplexní kepstrum

Δ kvantizační krok