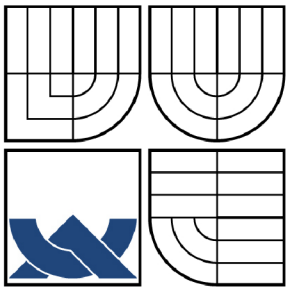
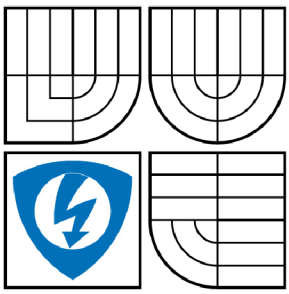


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV AUTOMATIZACE A MĚŘICÍ TECHNIKY

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF CONTROL AND INSTRUMENTATION

ROZPOZNÁVÁNI STANDARDNÍCH PILOT-CONTROLLER ŘÍDICÍCH POVELŮ V HLASOVÉ PODOBĚ

VOICE RECOGNITION OF STANDARD PILOT-CONTROLLER CONTROL COMMANDS

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

AUTOR PRÁCE — Bc. TOMÁŠ KUFA

AUTHOR

VEDOUCÍ PRÁCE — Ing. PETR HONZÍK, Ph.D.

SUPERVISOR

BRNO 2009

VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ
Fakulta elektrotechniky
a komunikačních technologií
Ústav automatizace a měřicí techniky

Diplomová práce

magisterský navazující studijní obor
Kybernetika, automatizace a měření

Student: Kufa Tomáš, Bc. ***ID:*** 83559
Ročník: 2 ***Akademický rok:*** 2008/09

NÁZEV TÉMATU:

Rozpoznáváni standardních PILOT-CONTROLLER řídicích povelů v hlasové podobě

POKYNY PRO VYPRACOVÁNÍ:

Vytvořte přehled týkající se problematiky hlasových ATC (Air Traffic Control) povelů. Zpracujte přehled přístupů použitelných pro jejich automatické rozpoznávání. Alespoň jeden z přístupů naprogramujte a v omezeném slovním rozsahu proveďte jeho vyhodnocení na skutečných zvukových záznamech. Srovnejte úspěšnost vlastního programu s nějakým jiným veřejně dostupným systémem.

DOPORUČENÁ LITERATURA:

Dle vlastního literárního průzkumu a doporučení vedoucího práce.

Termín zadání: 9.2.2009 ***Termín odevzdání:*** 25.5.2009

Vedoucí práce: Ing. Petr Honzík, Ph.D.
Konzultanti diplomové práce:

prof. Ing. Pavel Jura, CSc.
předseda oborové rady

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ Fakulta elektrotechniky a komunikačních technologií
Ústav automatizace a měřicí techniky

UPOZORNĚNÍ:

A b s t r a k t

Obsahem této práce je aplikace rozpoznávání řeči na ATC povely. Volba metod a přístupů k automatickému rozpoznávání ATC povelů vychází z podrobné studie letového provozu. Protože neexistuje jednoznačné řešení, zvlášť v tak obsáhlém oboru jako je rozpoznávaní řeči, je v této práci realizován rozpoznávač založený na porovnávání se vzory (DTW) a je srovnán s volně dostupným systémem HTK z University v Camridge založeném na statistických metodách využívajících skryté Markovovy modely. Míra vhodnosti obou metod je podložena praktickým testováním a vyhodnocením výsledku.

A b s t r a c t

The subject of this graduation thesis is an application of speech recognition into ATC commands. The selection of methods and approaches to automatic recognition of ATC commands rises from detailed air traffic studies. By the reason that there is not any definite solution in such extensive field like speech recognition, this diploma work is focused just on speech recognizer based on comparison with templates (DTW). This recognizor is in this thesis realized and compared with freely accessible HTK system from Cambrige University based on statistic methods making use of Hidden Markov models. The usage propriety of both methods is verified by practical testing and results evaluation.

K l í č o v á s l o v a

Rozpoznávání řeči, ATC povely, letecká frazeologie, řečový rozpoznávač, rozpoznávání podle vzoru, dynamické borcení časové osy-DTW, mel-kepstrální koeficienty-MFCC, skryté Markovovy modely-HMM, HTK toolkit

K e y w o r d s

Speech recognition, ATC commands, air phraseology, speech recognizer, template matching, dynamic time warping-DTW, mel-frequency ceptral coefficients, Hidden Markov Model-HMM, HTK toolkit

Bibliografická citace

KUFA, T. Rozpoznáváni standardních PILOT-CONTROLLER řídicích povelů v hlasové podobě. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2009. 51 s. Vedoucí diplomové práce Ing. Petr Honzík, Ph.D.

P r o h l á š e n í

„Prohlašuji, že svou diplomovou práci na téma Rozpoznávání standardních PILOT-CONTROLLER řídicích povelů v hlasové podobě 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 jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.“

V Brně dne: **25. května 2009** ……

podpis autora

P o d ě k o v á n í

Děkuji vedoucímu diplomové práce Ing. Petru Honzíkovi, Ph.D. za účinnou metodickou, pedagogickou pomoc a další cenné rady při zpracování mé diplomové práce. Martině Lyskové za korekturu a přátelům za ochotu a čas věnovaný tvorbě použitých nahrávek.

OBSAH

Seznam obrázků

Seznam tabulek

1. ÚVOD

Rozpoznávání zvukových signálů hlasu patří obecně mezi náročnou a velmi rozsáhlou vědeckou disciplínu. Je to obor stále se vyvíjející, a to hlavně v posledních čtyřiceti letech, přičemž první pokusy s jednoduchými zařízeními rozpoznávání se prováděly již v letech šedesátých. Zatím neexistuje žádný univerzální systém, který by dokázal rozpoznávat všechny možné vstupní hlasové signály tak, aby jeho úspěšnost byla vždy stoprocentní. Vysoká úspěšnost rozpoznávání hlasu spočívá v použití vhodných rozpoznávacích metod, filtrací, zvýrazňování apod. Je jich celá řada a jejich správné stanovení má zásadní vliv na konečný výsledek práce, proto je tento krok pravděpodobně nejdůležitější a je nutné mu věnovat patřičnou pozornost.

S tímto je velmi úzce spjato podrobné nastudování úlohy, kde má být hlasové rozpoznávání aplikováno. V našem případě se jedná o rozpoznávání ATC (Air Traffic Control) povelů. Je tedy nutné získat přehled způsobů komunikace mezi řídicí věží a letadlem, tedy znát nejen princip komunikace (pro určení kvality signálu, množství šumu atp.), ale také používané hlasové příkazy (určení velikosti databáze slov, pravidel výslovnosti, četnosti použití, priorit povelů atp.). Rozpoznání povelů, pro začátek alespoň v laboratořích, může být prvním krokem k vytvoření bezpilotního letadla-UAV (Unmanned Aerial Vehicle), které by se mohlo chovat přesně jako letadlo řízené živým pilotem. Není potřeba dodávat, že pouze v případě, bude-li výsledek dostatečně kvalitní a spolehlivý.

V této diplomové práci jsou shrnuty nejvhodnější postupy pro řešení úlohy rozpoznávání ATC povelů a také prakticky realizován rozpoznávač založený na metodě dynamického programování s omezenou slovní zásobou. Tento rozpoznávač je srovnán s volně dostupným toolkitem HTK z University v Camridge založeném na statistických metodách využívajících skryté Markovovy modely. Míra vhodnosti obou metod je podložena praktickým testováním a vyhodnocením výsledku.

2. LETECKÁ KOMUNIKACE

Z pohledu samotného rozpoznávání řečových signálů by se dalo pracovat bez informací z provozu, kde a jak řečové signály k rozpoznávání vznikají. Pokud je ale cílem rozpoznávat řečové signály s vysokou pravděpodobností úspěchů, je třeba se zaobírat i prostředím, kde vznikají, kudy a jakým způsobem vede datová cesta, a v neposlední řadě, co a v jakém rozsahu touto datovou cestou prochází.

2.1 LETECKÁ RÁDIOVÁ KOMUNIKACE

Komunikace mezi letadlem a pozemní stanicí probíhá na kmitočtech 118-137 MHz. Z bezpečnostních důvodů se používá AM modulace -při modulaci AM nedojde ke ztrátě informace při případném vysílání letadla a pozemní stanice najednou (obě vysílání jsou srozumitelná). Frekvenční modulace oproti AM modulaci slabší signál zcela odstraní. Při letu přes oceán je komunikace řešena digitálně přes satelit nebo na krátkých vlnách (ekonomicky slabší země).

Nejdůležitější je rozpoznávat přicházející informace z pozemních stanic, přičemž typů pozemních stanic je cela řada, jak je uvedeno níže, a pro potřebu samotného letu se stačí řídit jen několika.

Typy pozemních stanic[7]: CONTROL, APPROACH, RADAR, DEPARTURE, ARRIVAL, DIRECTOR, PRECISION, GROUND, DELIVERY, INFORMATION, INFO, DISPATCH, OPERATIONS.

Komunikace mezi letadlem a pozemní stanicí je čistě účelovou záležitostí a předmětem sdělovaných zpráv v leteckém provozu jsou:

1. Zprávy potřebné pro změny v navigaci letadla (změna kursu, výšky...)
2. Zprávy o poloze letadla či upřesnění nejbližší trasy (směrování na navigační body, vzdušné prostory...)
3. Meteorologické zprávy (VOLMET, ATIS, ...)
4. Provozní zprávy samotných aerolinií
5. Data přenášená mezi palubou letadla a pozemní stanicí (ACARS, HFDL, radar. odpovídač...)

Pro zjednodušení ale budeme považovat za nejnutnější jen hlasovou komunikaci v bodech 1 a 2.

2.2 FRAZEOLOGIE

Letecká frazeologie udává jasná pravidla, jak by měla komunikace vypadat a může do značné míry ovlivnit i výběr metod a algoritmů pro rozpoznávání řeči. Z tohoto důvodu je tato podkapitola velmi důležitá.

Požadavky techniky řeči pro spojovací postupy [11]:

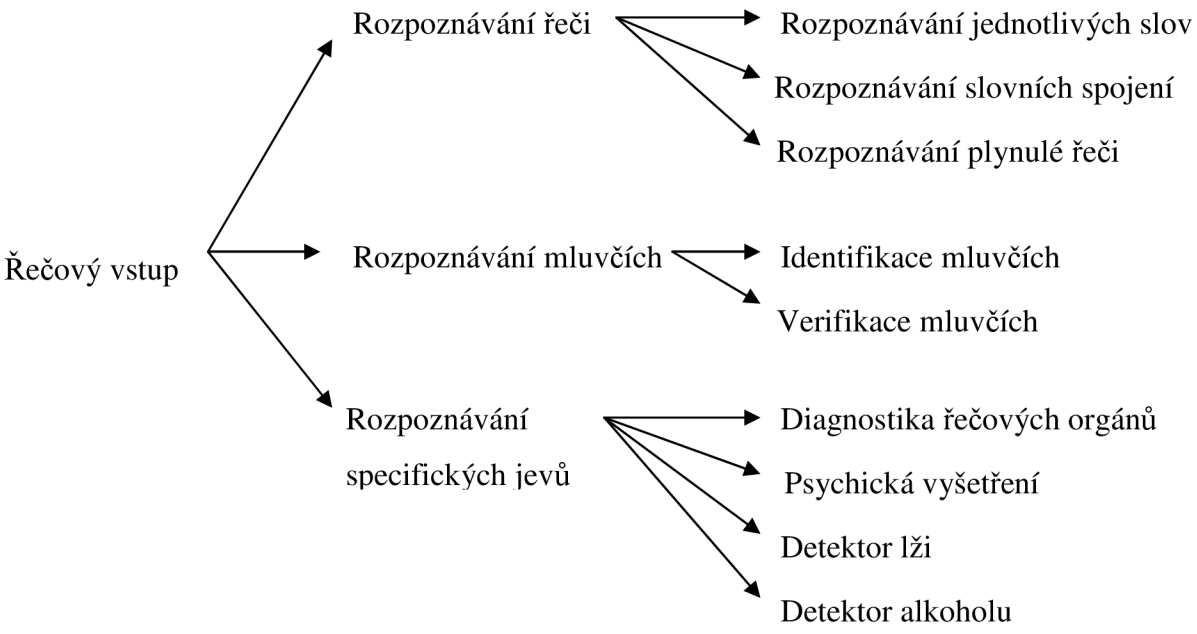
1. Vyslovovat každé slovo jasně a srozumitelně.
2. Udržovat stejnou rychlost hovoru, nepřekračující 100 slov za minutu. Je-li zpráva vysílána letadlu a je nutné provést záznam jejího obsahu, musí se přizpůsobením hovoru umožnit provedení písemného záznamu. Krátká přestávka před a po číslicích usnadňuje jejich srozumitelnost.
3. Zachovávat stejnou výši hlasu ve všech fázích hovoru.
4. Být seznámeni s provozní technikou mikrofonu ve vztahu k udržování konstantní vzdálenosti od mikrofonu, není-li používán modulátor s konstantní úrovní.
5. Přerušit hovor po dobu, kdy je nutné odvrátit hlavu od mikrofonu.

Předávané zprávy by měly být tedy krátké, obsahově jasné a srozumitelné. Jako doporučení se v literatuře [11] uvádí také vhodnost přizpůsobení techniky řeči převládajícím podmínkám spojení. Všechny tyto hlavní požadavky a doporučení (celkem je jich víc) značně ovlivňují koncovou úspěšnost rozpoznávání řeči.

Dalším aspektem týkajícím se frazeologie letecké komunikace je rozsah používaných povelů. Již bylo zmíněno, že povely jsou čistě účelové a proto ani počet používaných slov nemá zdaleka rozsah standardní mluvy. V případě bezproblémového (ideálního) letu by se mohl počet různých vyskytujících se slov dostat odhadem pod číslo 200. Tady lze začít uvažovat o algoritmech rozpoznávání řeči, které mají dobré výsledky na omezenou slovní zásobu.

3. ZPRACOVÁNÍ ŘEČOVÝCH SIGNÁLŮ

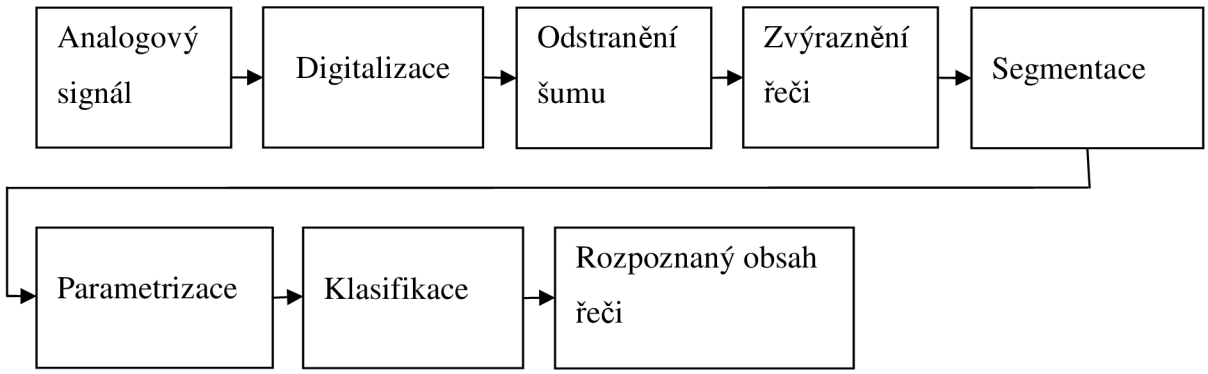
Oblast zpracování řeči je značně široká a v souvislosti se zde řešeným problémem není potřeba podrobně popisovat jednotlivá odvětví. Za zmínku ovšem stojí alespoň celkové rozdělení oblastí a cílů analýzy řečových signálů, viz. obrázek 3.1.



Obrázek 3.1 Oblasti a cíle analýzy řečových signálů [18]

Rozpoznávání řeči bude v následujících podkapitolách podrobněji rozvedeno, se zaměřením převážně na přístupy, které jsou pro rozpoznávání ATC povelů nejvhodnější.

Rozpoznávání řeči od vstupního signálu po konečný výsledek rozpoznání slova prochází několika fázemi. Jejich počet nemusí být vždy stejný, záleží na řešené problematice. Obecné schéma je vhodné uvést, neboť právě z něj se vychází pro konkrétní úlohu.



Obrázek 3.2 Postup zpracování a rozpoznávání řeči

3.1 DIGITALIZACE ANALOGOVÉHO SIGNÁLU

Byť to tak zřejmě na první pohled nevypadá, už digitalizací analogového signálu začíná práce na rozpoznávání řeči. Je nutné zvolit vhodný vzorkovací kmitočet tak, aby nedošlo ke ztrátám informací signálu, a zároveň, aby kmitočet nebyl příliš velký z důvodu objemného toku dat do vyšších úrovní celého systému. Bylo zjištěno, že podstatná informace v řečovém signálu je rozložena v kmitočtovém rozsahu do 8000Hz (srozumitelné používané přenosové telefonní pásmo je 300-3400Hz). Plně postačující vzorkovací kmitočet je tak 16kHz.

Po vzorkování dochází k operaci kvantování, to převádí diskrétní signál na kvantovaný signál. Ten je charakterizován konečným počtem možných diskrétních hodnot. Ty souvisí s maximální dosažitelnou dynamikou záznamu řeči. Ta by měla být kolem 50dB. Uveďme jen, že při použití lineárního kvantování je pro splnění této podmínky použito minimálně 10bitů, přičemž běžný standard je použití 16ti bitového lineárního kvantování. Negativní vlastností kvantování je degradace akustického signálu superpozicí kvantovacího šumu. Rozdíl mezi kvantovanou a přesnou hodnotou vzorku představuje kvantizační chybu vyjádřenou jako

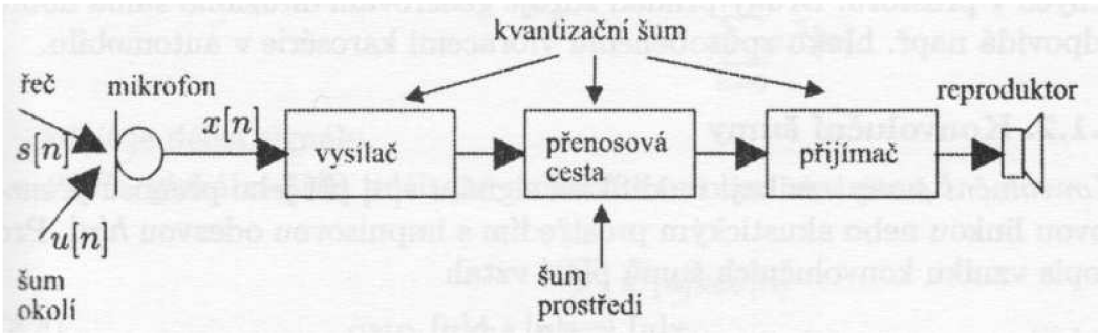
$$e(n) = x_q(n) - x(n) \qquad (3.1)$$

kde xq(n) velikost vzorku
x(n) všechny velikosti vzorků spadající do rozsahu jednoho kvantizačního stupně
e(n) kvantizační chyba

Údaje jako vzorkovací frekvence, počet bitů na jeden vzorek, typ kvantizéru a kódu dat, uspořádání souborů s hlasovým signálem, záhlavím souboru, pořadím bytů apod. tvoří tzv.*formát řečového signálu*.

3.2 VÝSKYT ŠUMU V ŘEČOVÉM SIGNÁLU

Každý řečový signál je při přenosu poznamenán šumem. Existují dva základní typy šumů: aditivní šum a konvoluční šum. Do skupiny aditivních šumů spadá kvantizační šum nebo šumové pozadí prostředí, popřípadě šum přenosové cesty. Konvoluční šum je pak způsoben přeslechy nebo změnami parametrů přenosové cesty. Působení šumu je názorně ukázáno na obrázku 3.3.



Obrázek 3.3 Výskyt šumu při přenosu řeči [19]

Řečový signál je značně zašuměný, a proto je třeba šum „vypreparovat“, tedy určit jeho míru v zašuměném signálu. To není jednoduchý úkol. Detekce šumu při jeho velké hodnotě se může stát nerealizovatelnou. Nejběžnějším kritériem pro měření úrovně šumu je používáno *odstup signál šum-SNR.* Existuje v několika modifikacích:

- GSNR-globální SNR počítané přes celý signál.
- SNR-globální SNR počítané z úseků s řečovou aktivitou.
- SNRi-lokální(krátkodobé SNR počítané v i-tém segmentu konečné délky při segmentaci řeči na kvazistacionární úseky(typicky 10-35ms)
- SSNR-segmentální SNR dané průměrem lokálních SNR v segmentech s řečovou aktivitou.
- ASNR-aritnetické segmentální SNR, dané průměrem lineárních lokálních SNR v segmentech s řečovou aktivitou a následným přepočtem na dB.

Každá z těchto uvedených modifikací má trochu jiné vlastnosti. Pro řečové signály používané v letectví, tedy dosti zašuměné, se hodí nejvíce SNR počítané z úseků s řečovou aktivitou, SSNR a ASNR. Všechna kritéria se počítají pro segmenty řeči (slovo nebo fragment). Nejlépe postihnout dynamiku řeči je schopno kritérium SSNR, ovšem na úkor mírně vychýlených hodnot. Protože je předpoklad použití právě tohoto kritéria v zadané úloze, bude uveden vzorec výpočtu:

$$SSNR = \frac{1}{K}\sum_{i=0}^{L-1} SNR_i \cdot VAD_i$$

$$SNR_i = 10\log\frac{\sigma_{s,i}^2}{\sigma_{n,i}^2} = 10\log\frac{\sum_{n=2}^{M-1} s_i^2[n]}{\sum_{n=2}^{M-1} n_i^2[n]} \qquad (3.2)$$

Kde σ_s^2 , σ_n^2 výkon řečového signálu a aditivního šumu

M délka segmentovaného signálu

$s_i[n]$, $n_i[n]$ segment signálu řeči a aditivního šumu délky M vybírané s krokem m

L celkový počet segmentů

K počet segmentů s řečovou aktivitou a následným přepočtem na dB

VAD_i hodnota 1 nebo 0 pro segmenty s řečovou aktivitou nebo bez ní

3.3 ZVÝRAZNĚNÍ ŘEČI

Zvýraznění řeči je do značné míry také potlačení šumu a pro rozpoznávání řeči je to část poměrně významná. Existuje několik technik, jak zvýraznění řeči dosáhnout. Patří zde

- Filtrace v časové a frekvenční oblasti
- Kompenzace v časové a frekvenční oblasti
- Modelování řeči
- Využití prostorové informace

V prvním případě (filtrace) jde o techniku, která potlačuje šum, ale také zkreslí řeč. Kompenzace v časové a frekvenční oblasti oproti tomu nezkreslí řeč a přitom potlačí šumy. Z těchto dvou technik je těžké vybrat, která by měla být použita, neboť vyvstává otázka, zdali dojde k výraznému zkreslení řeči filtrací v časové a frekvenční oblasti, a o jakou míru bude šum potlačen lépe (pokud se má o této technice vůbec uvažovat) než u techniky kompenzace v časové a frekvenční oblasti.

Modelování řeči provádí analýzu řeči se šumem, přičemž se snaží získat její model a pak původní signál opravit. Je to další technika, o které lze reálně uvažovat, a o tom, která bude mít nejlepší výsledky v řešeném problému, může rozhodnout až praktická realizace a srovnání.

Čtvrtá uvedená technika se týká umístění mikrofonu v prostoru. V našem konkrétním případě ale tato metoda se nedá použít (nelze ji ovlivnit) a navíc sama o sobě může být použita jen za omezujících podmínek.

3.4 SEGMENTACE A PARAMETRIZACE ŘEČI

Hlavní cíl segmentace a parametrizace řeči je snížit tok dat do vyšších úrovní systému.

Každý řečový signál se musí rozdělit na části (segmenty), aby bylo možné popsat (parametrizovat) jednotlivé segmenty, které se pak vyhodnocují. Velikost segmentů může být různá. Nejpoužívanější jsou framy, jejich délka se pohybuje mezi 10-35ms a to tak, aby frame byl kratší než trvání jedné hlásky. Frame se pak může považovat za stacionární a lze popsat méně parametry. Tento přístup se používá hlavně při větších kapacitách slovníků a je náročný na realizaci. Segment může tvořit také celé slovo, difón, trifón nebo slabiku. Nejlépe pro rozpoznávání ATC povelů se ale jeví přístup segment=slovo, vhodný pro malé nebo středně velké slovníky.

Parametrizace řeči pak musí (aby splňovala svůj význam redukce toku dat do vyšších úrovní systému), umožnit dostatečné odlišení framu (slov), potlačovat pokud možno charakteristické rysy řečníka a měla by být výpočetně dostupná.

3.5 KLASIFIKACE ŘEČI

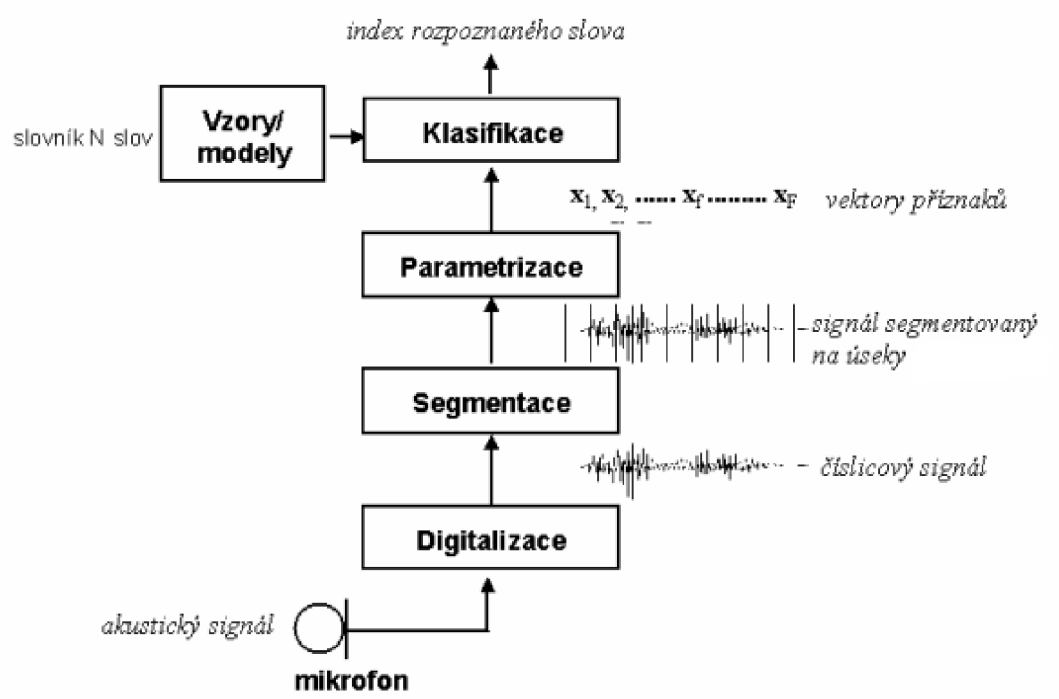
Klasifikace je poslední krok v rozpoznávání řeči. Přiřazené parametry segmentům se zde vyhodnocují a dle výsledku se určí nejlepší vhodné slovo. Existují tři metody automatického rozpoznávání: metody založené na porovnávání vzorů (template matching), metody využívající statistické rozhodování a metody uplatňující znalostní přístup.

Teplate matching je metoda klasifikace na základě referencí, u které by bylo obtížné realizovat nezávislost na mluvčím. Je zde používána metoda dynamického borcení času DTW (Dynamic Time Warping) pro „natažení“ nebo „zkrácení“ slova na hodnotu odpovídající referenčnímu vzoru. Klasifikace je pak obvykle prováděna na principu nejmenší vzdálenosti k některému obrazu vzorového slova uloženého ve slovníku. Tato metoda sice ztratila na aktuálnosti, ale pro specifické úlohy s malým slovníkem a závislostí na mluvčím se stále používá pro její výbornou účinnost.

Metody založené na statistikách mají také referenční slovník, ale jiný přístup k porovnávání. Klasifikace se zde děje pomocí parametrických modelů. Modelem nemusí být celé slovo, ale jen například fonémem (slovo je pak modelováno zřetězením menších jednotek). Nejvýznamnějším parametrickým přístupem ke klasifikaci jsou tzv. Markovovy skryté modely-HMM (Hidden Markov Model). Ty využívájí právě zřetězování menších subslovních jednotek, přičemž jsou popsány jak tyto jednotky, tak i přechody mezi nimi pravděpodobnostními koeficienty nebo funkcí hustoty rozložení pravděpodobnosti. Každou položku slovníku zastupuje jeden model, který je porovnáván s reprezentací neznámého slova. Při porovnávání u metody HMM se neměří vzdálenost, s jakou se model blíží své referenční podobě, ale pravděpodobnost. Ze všech referenčních modelů se pak vybere ten, jehož pravděpodobnostní hodnota je největší. HMM má jednu základní výhodu-umožňuje natrénovat slova slovníku takovým způsobem, že je téměř nezávislý na konkrétním mluvčím.

Oba přístupy jsou použitelné pro rozpoznávání ATC povelů a proto se v následujících kapitolách rozpoznávač založený na porovnávání se vzory (využívající DTW) realizuje a bude porovnán s volně dostupným toolkitem HTK založeném na

statistické metodě a využívající skryté Markovovy modely. Pro oba modely je platné schéma postupu zpracování a rozpoznávání řeči z obrázku 3.4.



Obrázek 3.4 Rozpoznávání izolovaných slov z malého slovníku [5]

4. REALIZACE ROZPOZNÁVAČE ZALOŽENÉHO NA POROVNÁVÁNÍ SE VZORY

V předchozí kapitole byl uveden teoretický postup zpracování reálného signálu z letového provozu a zvolení optimálního řešení. V této kapitole je rozvedeno řešení rozpoznávače založeného na porovnávání se vzory (template matching) s omezenou kapacitou slov, jež obnáší předzpracování dat, parametrizaci a klasifikaci. Realizace rozpoznávače probíhala v laboratorních podmínkách, proto signál není zašuměný vůbec nebo jen minimálně, a odstraňování šumu jako součást předzpracování dat není nutná. Obecně problém odstraňování šumu je komplikovaná část, která je tématem mnoha samostatných prací.

Důraz je kladen hlavně na předzpracování dat a funkčnost realizovaného rozpoznávače. Úspěšnost rozpoznávání rozpoznávače založeného na template matching bude porovnána s výsledky rozpoznávače založeného na statistickém přístupu z následující kapitoly, který je volně dostupný na http://htk.eng.cam.ac.uk/ pod názvem HTK. Bude tak ověřena nebo vyvrácena teorie z kapitoly 3.5 o vhodnosti a úspěšnosti jednotlivých přístupů.

4.1 DIGITALIZACE, POUŽITÁ DATA

Pro testování rozpoznávačů byla vytvořena malá databáze o 10 anglických slovech, 10 verzích každého slova a namluveno 5 mluvčími. Celkový počet slov tak činí 500. Všechny nahrávky byly upraveny na jednotný formát souborů *.wav o vzorkovacím kmitočtu 8000Hz a lineárním 16-bitovém kvantování. Uvedený formát byl zvolen ze dvou důvodů:

- cílem není rozlišit jednotlivé mluvčí, proto se volí takový vzorkovací kmitočet, který pokryje pásmo s užitečným signálem vypovídajícím o obsahu zvukového záznamu
- redukce objemu dat pro další zpracování (parametrizace, klasifikace) znamená urychlení výpočtu

K nahrávání nebyly použity stejné mikrofony, ale 4 odlišné, aby se situace přiblížila realitě a do jisté míry se znesnadnilo rozpoznávání.

Podrobný popis značení jednotlivých záznamů lze nalézt v příloze (A).

4.2 PŘEDZPRACOVÁNÍ DAT

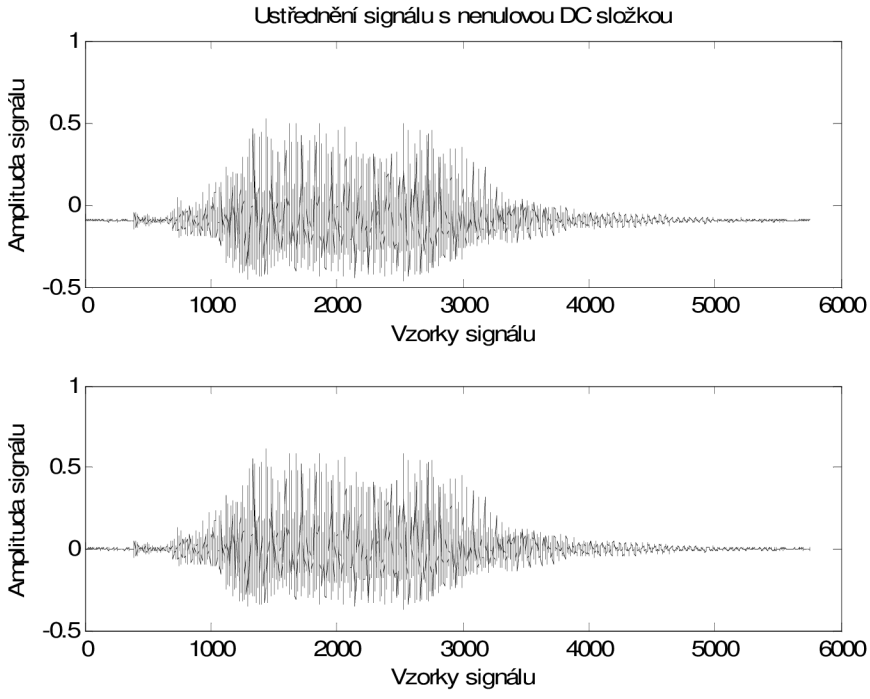
Následující podkapitoly jsou zaměřeny na úpravu signálu takovým způsobem, aby po jejich parametrizaci měly parametry co nejvyšší vypovídací hodnotu o informačním obsahu signálu.

4.2.1 Ustřednění

Jde o operaci odstraňující nenulovou stejnosměrnou složku, která v dalších krocích zpracování a následné parametrizaci může působit rušivě. Jedná se o jednoduchou nenáročnou úpravu a podle vzorce

$$s'[n] = s[n] - \mu_S \tag{4.1}$$

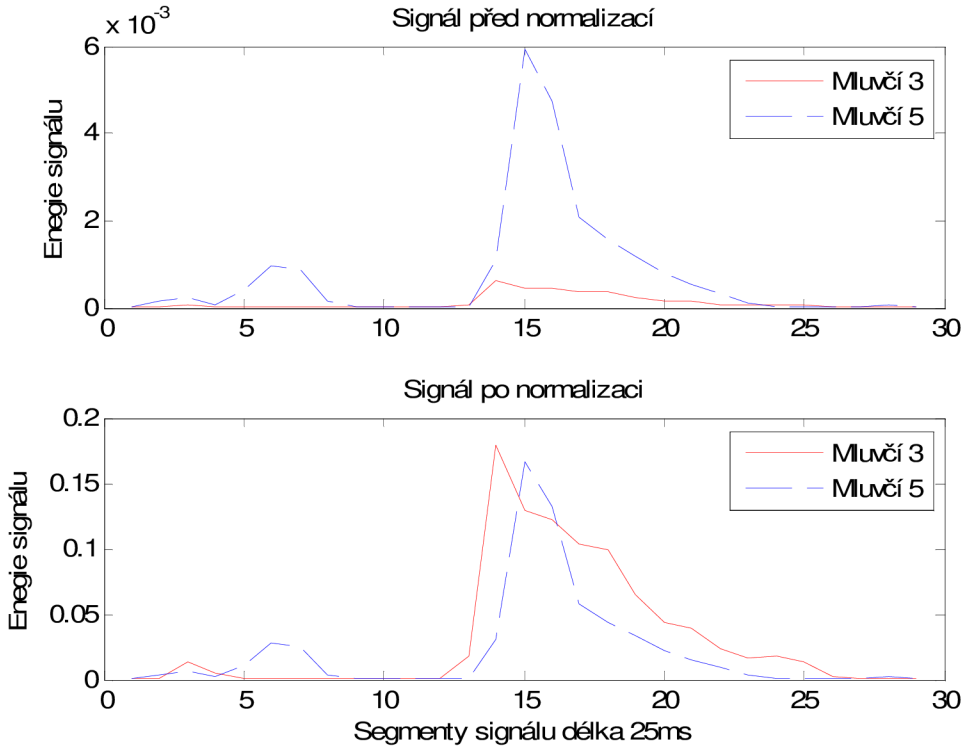
kde μ_S se musí odhadnout. Pro ustřednění v off-line režimu μ_S odpovídá střední hodnotě signálu.



Obrázek 4.1 Ukázka ustřednění signálu

4.2.2 Normalizace signálu

Normalizace signálu se provádí tam, kde jednotlivé signály (nahrávky) jsou nahrány s různými velikostmi amplitud (například díky citlivosti mikrofonu, vzdálenost mluvčího od mikrofonu). Střední krátkodobá energie se může pro stejná slova bez normalizace výrazně lišit. Samozřejmě nelze tvrdit, že po normalizaci bude pro stejná slova stejná krátkodobá energie. To neplatí ani pro stejné mluvčí. Normováním amplitudy se ale k sobě tato slova budou mnohem více podobat. Některé klasifikátory používají jako jeden z klasifikátoru právě střední krátkodobou energii, ale nejčastější využití střední krátkodobé energie je k určování hranic mezi slovy. Obrázek 4.2 svědčí o tom, že normalizovat signál má svůj význam i pro různé mluvčí



Obrázek 4.2 Význam normalizace signálu-dva mluvčí, slovo "descent"

4.2.3 Preemfáze

Při mluvě dochází k útlumu intenzity zvuku pro vyšší kmitočtové složky, amplituda klesá o 20dB na dekádu frekvence. Protože užitečné informace signálu

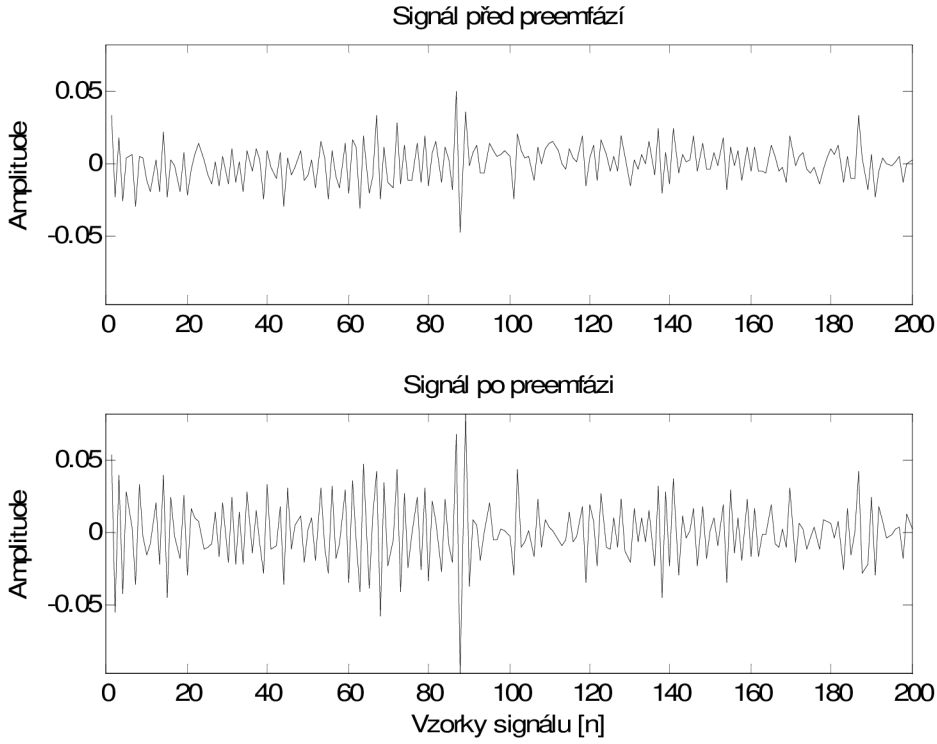
jsou obsaženy v pásmu nad 300Hz, je vhodné provést kompenzaci. Pro zajímavost uveďme, že podstatná část celkové energie signálu je obsažena v pásmu pod hranici 300Hz. Kompenzace útlumu je prováděna jednoduchým filtrem 1. řádu s horní propustí

$$H(z) = 1 - \kappa z^{-1} \tag{4.2}$$

kde κ se volí v rozmezí 0,9-1, standardně hodnota 0,97. V časové oblasti se preemfáze uplatňuje vztahem

$$s'(n) = s(n) - \kappa s(n-1) \tag{4.3}$$

Na obrázku 4.3 jsou vidět výraznější špičky a celkový výraz signálu je jakoby zašuměný, to ukazuje na větší podíl vyšších frekvencí.



Obrázek 4.3 Signál před a po preemfázi

4.2.4 Segmentace signálu a okenkování

Řečový signál se považuje za náhodný, což je nežádoucí, protože jej nelze nijak efektivně popsat. Proto se signál rozdělí na dostatečně malé segmenty, aby splňovaly dvě podmínky:

- segment musí jít popsat konstantními parametry, vybraný signál v rámci by měl vykazovat stacionaritu
- velikost segmentu musí být dostatečně velká pro bezchybný odhad parametrů

Bylo zjištěno, že splnění těchto podmínek nalezneme při volbě segmentu mezi 10 až 35ms. Všechny segmenty mají stejnou délku.

Existují dva způsoby segmentace, pro každou jsou charakteristické jiné vlastnosti

- Segmentace bez překrytí rámců
  - rychlý časový posun signálu
  - malé nároky na paměť/procesor
  - parametry následných rámců se mohou značně lišit
- Segmentace s překrytím rámců
  - pomalý časový posun
  - velké parametry na paměť/procesor
  - uhlazenější časové parametry signálu
  - rámce si mohou být příliš podobné (nežádoucí pro rozpoznávání pomocí HMM)

Výkony dnešních PC jsou dostatečně velké na to, aby bylo možno bez velkých časových prodlev použít druhý způsob segmentace signálu, navíc se s použitím segmentace s překrytím rámců očekávají lepší výsledky prohledávací metody DTW (viz. kapitola 4.4.2), neboť prohledávací cesta je přesnější.

Segmentace signálu patří k teorii, která určuje, jakým způsobem mají být aplikovány okénkové funkce na signál. Jak bude ukázáno dále, okénková funkce vybírá podle předepsaných parametrů z celého signálu takovou část, která odpovídá jednotlivému segmentu a navíc jednotlivým vzorkům přidělí určitou váhu. Existuje celá řada okénkových funkcí, které jsou vhodnější pro různé signály, patří zde pravoúhlé okno, trojúhelníkové okno, kosinové okno, Gaussovo okno, Hammingovo

okno, Hannovo okno, Blackmanovo okno a Kaiserovo okno. Nejpoužívanější okénkovou funkcí pro řečové signály je Hammingovo a pravoúhlé okno.

Matematický popis pro obě okna představují rovnice 4.4 a 4.5. Pravoúhlé okno je nejjednodušší, vybere pouze daný počet vzorků odpovídající velikosti zvoleného segmentu. Jeho aplikací na signál dochází ke dvěma nežádoucím jevům- rozmazání a rozptylu spektra signálu. Oba tyto jevy lze odstranit použitím Hammingova okna.

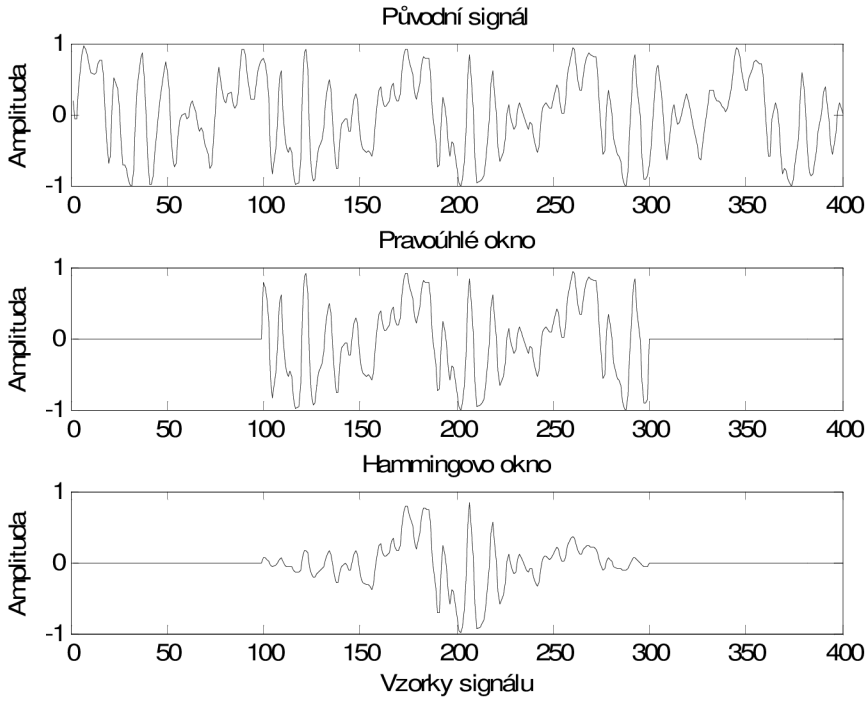
- ***Pravoúhlé okno***

$$w[n] = \begin{cases} 1 & pro\ 0 \leq n \leq l_{ram} - 1 \\ 0 & jinde \end{cases} \tag{4.4}$$

- ***Hammingovo okno***

$$w[n] = \begin{cases} 0{,}54 - 0{,}46cos\frac{2\pi n}{l_{ram}-1} & pro\ 0 \leq n \leq l_{ram} - 1 \\ 0 & jinde \end{cases} \tag{4.5}$$

Názornou ukázku aplikace obou popisovaných oken lze vidět na obrázku 4.4.



Obrázek 4.4 Výběr segmentu aplikací oken na signál, foném "á"

Pro realizovaný rozpoznávač bylo použito Hammingovo okno.

4.3 PARAMETRIZACE

Parametrizace je posledním krokem před samotnou klasifikací, kdy se jednotlivé segmenty řečového signálu popíší vhodnými parametry. Úkolem parametrizace je tedy přizpůsobit vstupní data potřebám rozpoznávače. Nejvýznamnější parametry pro zpracování řeči jsou:

- MFCC (mel-kepstrální koeficienty)
- RASTA (RelAtive SpectTrAl processing)
- PLP (Perceptual Linear Prediction)

Pro parametrizaci bylo použito MFCC.

4.3.1 Výpočet Mel-kepstrálních koeficientů

Obecně se spektrální analýza používá kvůli lepšímu popisu signálu, neboť bylo dokázáno, že odlišnosti v řečových signálech představující stejné slovo/foném nejsou ve spektrální oblasti, ale v časovém členění. To je důvodem tvorby parametrizace ve spektrální oblasti.

4.3.1.1 Diskretní fourierova transformace

Řečový signál, který chceme klasifikovat, se musí z výše uvedených důvodů převést do spektrální oblasti. Pro rychlejší výpočty se tak děje rychlou fourierovou transformací (FFT).

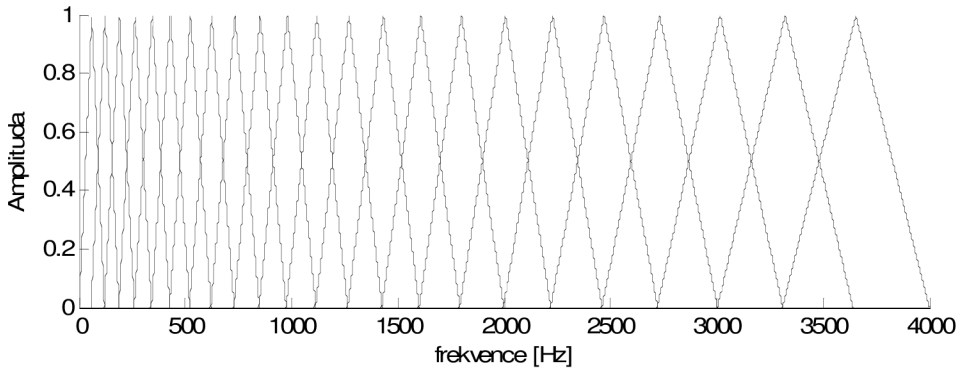
4.3.1.2 Mel-spektrum

Lidský sluch je charakteristický větším rozlišením na nižších frekvencích. U MFCC postupujeme tak, že na frekvenční osu rozmístíme nelineárně filtry. Při konstrukci filtrů můžeme nelineárně upravit frekvenční osu a na této upravené ose pak filtry rozmístit rovnoměrně. Používaná nelineární úprava využívá převodu Hertzů na Mely [3]:

$$f_{mel} = 2595\, log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \tag{4.6}$$

Na takto upravenou osu frekvence se aplikuje tzv. melovská banka filtrů (viz. obrázek 4.5), kde počet bank se volí standardně 23-26. Banky představují trojúhelníkové okno se vzájemným sousedním 50-ti procentním překrytím.

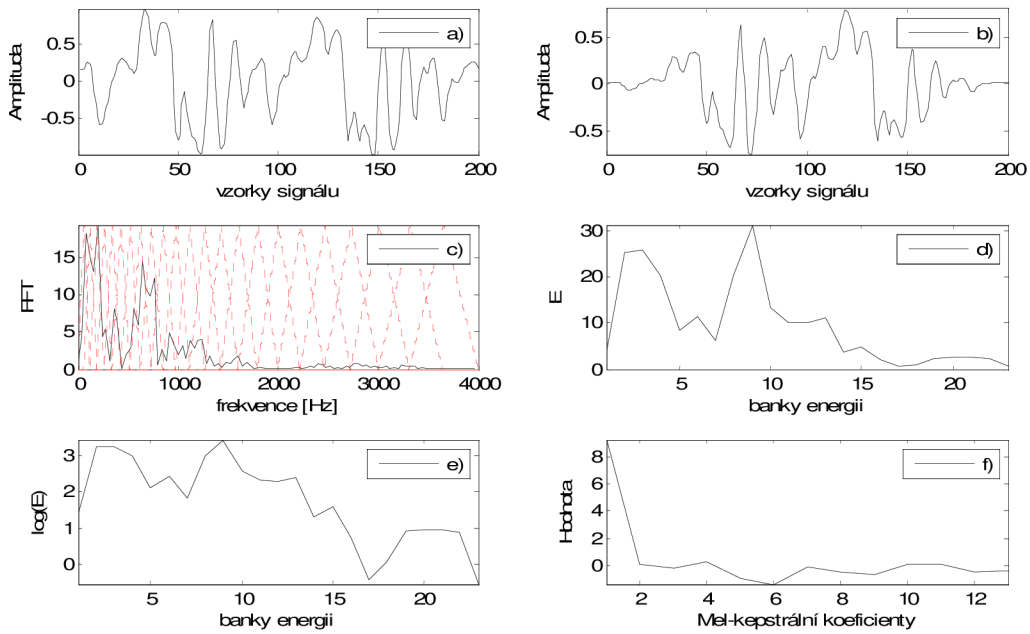
Spektrální hodnoty parametrizovaného signálu obsažený v každém okně se tímto oknem váhují a sečtou. Výstupní hodnoty představují mel-spektrum.



Obrázek 4.5 Melovská banka filtrů

4.3.1.3Mel-kepstrum

Výstup z banky filtrů se logaritmuje z důvodu logaritmického vnímání hlasitosti lidmi a poté se jednotlivé koeficienty dekorelují pomocí cosinovy transformace. Výsledné hodnoty představují mel-kepstrální koeficienty. Pro menší nároky na výpočetní techniku se parametry redukují při dekorelaci z původních 23 hodnot na 13. Pro větší přehlednost je celý postup vidět na obrázku 4.6.



Obrázek 4.6 Postup výpočtu MFCC

Jednotlivé obrázky představují: a) původní signál, b) preemfázovaný signál, c) DFT spektrum a jeho váhování filtry, d) energie na výstupech jednotlivých filtrů, e) log této energie, f) MFCC.

4.4 KLASIFIKACE

V předchozích kapitolách byl ukázán postup parametrizace segmentu (25ms záznamu). Každé slovo je tvořeno desítkami až stovkami segmentů. Po parametrizaci je slovo popsáno maticí parametrů, kde každý řádek odpovídá jednomu segmentu a má 13 sloupců-parametrů (viz kapitola 4.3). Jelikož ani stejný mluvčí neřekne jedno slovo stejně několikrát za sebou, mají stejná slova rozdílný počet segmentů. Rozpoznávač založený na porovnávání se vzory může být řešen lineární nebo nelineární transformací. Vlastnosti a princip obou transformací jsou vysvětleny v následujících podkapitolách.

4.4.1 Lineární transformace

Lineární transformace je metodou velice jednoduchou a v praxi již příliš nepoužívanou, přesto v laboratorních podmínkách nemá špatné výsledky. Dvě slova jsou srovnávány tak, že parametrizační matici kratšího slova přizpůsobí velikosti parametrizační matice delšího slova. V našem případě tedy přidá do parametrizační matice kratšího slova tolik řádků matice, kolik odpovídá rozdílu řádků obou matic. Toto se realizuje použitím transformační funkce času:

$$D(K,D) = \sum_{i=1}^{K} d\left[k\left(w(i)\right), d(i)\right] \qquad (4.7)$$

kde w(i) je hledaná cesta v rovině (i,j) definována tak, aby srovnání bylo lineární ***K*** je parametrizační matice kratšího slova a ***D*** je parametrizační matice delšího slova.

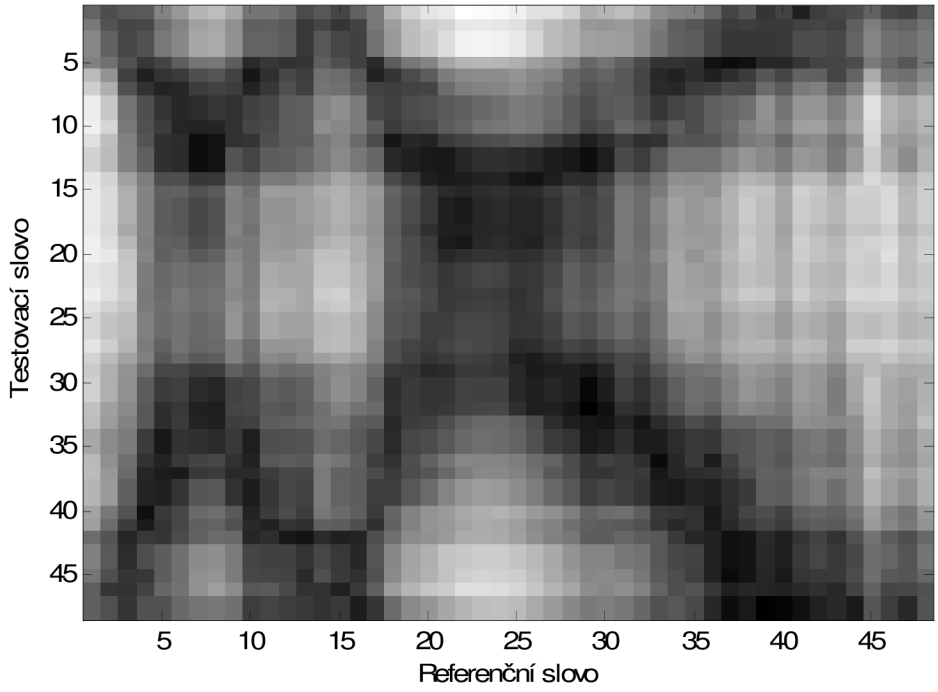
Nevýhodou této transformace je neúspěšné přiřazení slova v případě, kdy je špatně určena hranice slova při jeho nahrávání. Spolu s tímto slovem se tak v záznamu objeví například hluk pozadí, šum. Parametrizovaná matice tohoto chybného záznamu se pak porovnává s referencemi ve slovníku a nemůže být správně určena.

4.4.2 Nelineární transformace, DTW

Nevýhodu lineární transformace řeší nelineární transformace, jejíž matematickou realizací je optimalizační algoritmus známý pod názvem Dynamic Time Warping-dynamické borcení časové osy (DTW).

Nelineární transformace přizpůsobuje nejen celkovou délku slov (stejně jako lineární transformace), ale rovněž jeho jednolivé vnitřní části.

Z parametrizačních matic dvou srovnávajících se slov se vypočte matice vzdálenosti o velikosti LxN, kde L je počet segmentů neznámého slova a K je počet segmentů referenčního slova. Pro zjednodušení se před výpočtem matice vzdálenosti menší parametrizační matice přizpůsobí té vetší. Taková matice může vypadat, jak je ukázáno na obrázku 4.7, kde tmavší místa odpovídají větší lokální vzdálenosti.



Obrázek 4.7 Matice vzdálenosti pro slova "descent" a "ground"

Jednotlivé vzdálenosti mohou být počítány několika způsoby a podílejí se na úspěšnosti rozpoznávání. Předem nelze určit, který způsob výpočtu vzdáleností bude mít nejlepší výsledky. Literatura [18] nicméně uvádí, že praktický význam při měření v reálných podmínkách má Euklidova vzdálenost.

Nejběžnější typy vzdálenosti:

- Absolutní rozdíl vzdálenosti

$$d_b(j) \sum_{n=1}^{N} |r_n(j) - t_n(j)| \quad (4.8)$$

- Čebyševova vzdálenost (míra)

$$d_M(j) = max_{n=1,..,N} |r_n(j) - t_n(j)| \quad (4.9)$$

- kvadrát Euklidovy vzdálenosti

$$d_{KE}(j) = \sum_{n=1}^{N} [r_n(j) - t_n(j)]^2 \quad (4.10)$$

- Euklidova vzdálenost

$$d_E(j) = \sqrt{\sum_{n=1}^{N} [r_n(j) - t_n(j)]^2} \quad (4.11)$$

kde $r_n(j)$ … n-tý příznak j-tého segmentu slova vzorového slova

$t_n(j)$ … n-tý příznak j-tého segmentu slova testovaného slova

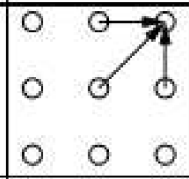
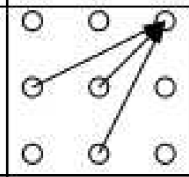
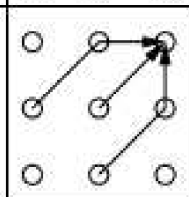
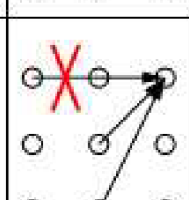
N …… celkový počet příznaků

Matice vzdálenosti vzniklá použitím jednou z výše uvedených vzorců je poté prohledávána k určení optimální cesty maticí pomocí DTW. Optimální cesta je taková, která ve výsledku zajistí nejmenší součet vzdálenosti podél této cesty. Optimální cesta musí respektovat dvě základní pravidla prohledávání

- počátek cesty musí začínat v matici vzdálenosti v prvku (1,1)
- konec cesty je v prvku (j,i), kde *j* je počet segmentů testovaného slova a *i* je počet segmentů referenčního slova

Hledání optimální cesty pouze za těchto podmínek by bylo výpočetně velice náročné. Proto se zavedlo několik typů funkcí DTW. Starší literatura [17] těchto typů uvádí sedm. Od tří nejsložitějších se upustilo, protože plně dostačují čtyři základní, viz Tabulka 1.

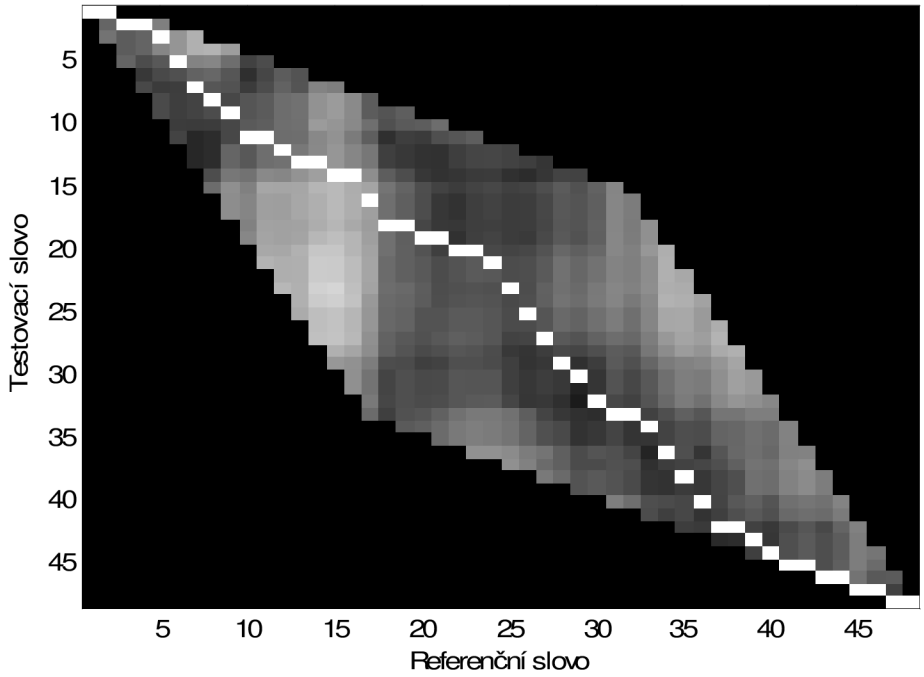
Tabulka 1 Typy lokálních omezení cesty DTW [3]

Typ DTW		α	β	Typ $w(k)$	$g(n,m)$
I.		0	∞	a	$\min\begin{cases} g(n,m-1)+d(n,m) \\ g(n-1,m-1)+2d(n,m) \\ g(n-1,m)+d(n,m) \end{cases}$
				d	$\min\begin{cases} g(n,m-1)+d(n,m) \\ g(n-1,m-1)+d(n,m) \\ g(n-1,m)+d(n,m) \end{cases}$
II.		$\frac{1}{2}$	2	a	$\min\begin{cases} g(n-1,m-2)+3d(n,m) \\ g(n-1,m-1)+2d(n,m) \\ g(n-2,m-1)+3d(n,m) \end{cases}$
				d	$\min\begin{cases} g(n-1,m-2)+d(n,m) \\ g(n-1,m-1)+d(n,m) \\ g(n-2,m-1)+d(n,m) \end{cases}$
III.		$\frac{1}{2}$	2	a	$\min\begin{cases} g(n-1,m-2)+2d(n,m-1)+d(n,m) \\ g(n-1,m-1)+2d(n,m) \\ g(n-2,m-1)+2d(n-1,m)+d(n,m) \end{cases}$
IV.		$\frac{1}{2}$	2	b1	$\min\begin{cases} g(n-1,m)+kd(n,m) \\ g(n-1,m-1)+d(n,m) \\ g(n-1,m-2)+d(n,m) \end{cases}$ kde $k=1$ pro $r(k-1)\neq r(k-2)$ $k=\infty$ pro $r(k-1)=r(k-2)$

Použitý typ pro realizovaný rozpoznávač je IV. Pokud se podíváme na obrázek 4.7 nekoresponduje s ukázkou procházení cesty z tabulky 1. Jde pouze o formalitu, která se programátorsky dá snadno vyřešit. Důležitější je pro typ DTW IV vysvětlit logiku procházení matice vzdálenosti:

- $bod(i-1,j) \xrightarrow{přechod} bod(i,j)$ jeden segment testovaného slova je přiřazen dvěma po sobě jdoucím segmentům-je zopakován.
- $bod(i-1,j-1) \xrightarrow{přechod} bod(i,j)$ přiřazení dvou po sobě jdoucích segmentů slov-zachování segmentu.
- $bod(i-1,j-2) \xrightarrow{přechod} bod(i,j)$ dva segmenty testovaného slova jsou přiřazeny jednomu segmentu vzorového slova-segment je vypuštěn.

Při dodržení podmínek pro typ DTW IV se matice vzdálenosti z obrázku 4.7 omezí na prostor znázorňující obrázek 4.8. Přičemž černý prostor představuje nekonečně velké lokální vzdálenosti. Zcela bílá barva je již vyznačená optimální cesta pro hledané slovo.



Obrázek 4.8 Omezení matice vzdálenosti pro nalezení optimální cesty pro DTW IV a nalezená optimální cesta

4.5 VÝSLEDKY ROZPOZNÁVAČE

Postup návrhu rozpoznávače a nastavení byl uveden v předchozích kapitolách a bylo zdůvodněno zvolení těchto nastavení, přesto zde bude pro přehlednost a lepší dohledání uvedeno znovu celé nastavení rozpoznávače a způsobu předzpracování.

- Signál
  - vzorkovací frekvence 8000Hz
  - kvantování 16-ti bitové lineární
- Předzpracování
  - Konstanta preemfáze κ 0,97
  - délka okna 25ms
  - překrytí oken 10ms
- Parametrizace
  - parametry MFCC
  - počet bank filtrů 23
  - počet mel-frekvenčních koef. 13

V Teoretickém úvodu bylo zmíněno, že metoda porovnávání se vzory založena na prohledávání matice vzdálenosti algoritmem DTW je poměrně úspěšná pro stejné mluvčí.

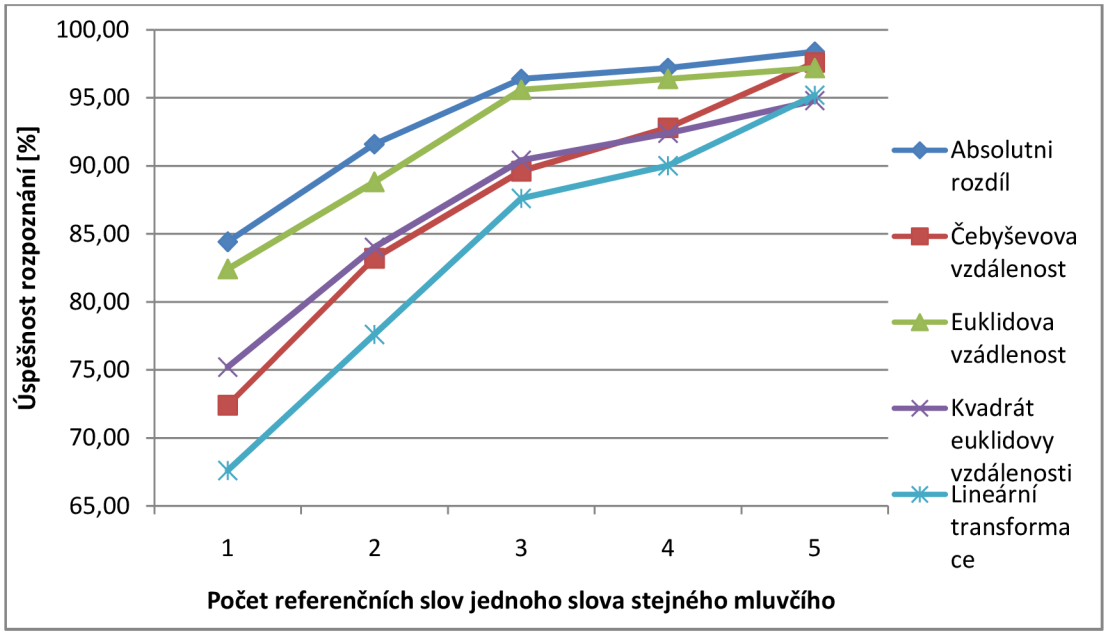
Tabulka 2 2 tuto teorii podkládá. Pro jednotlivé mluvčí byly jako referenční vzory postupně do slovníku zakomponovány verze slov 1-5 stejného mluvčího, přičemž testovacích slov bylo vždy 5 od stejného mluvčího (verze slova 6-10).

Tabulka 2 Úspěšnost [%] DTW s použitím absolutního rozdílu vzdálenosti a obsahem 1-5 referenčních slov každého slova stejného mluvčího

Mluvčí	Počet referenčních slov stejného mluvčího				
	1	2	3	4	5
Mluvčí 1	82	88	100	100	100
Mluvčí 2	78	96	96	98	98
Mluvčí 3	94	94	96	94	98
Mluvčí 4	86	96	96	98	100
Mluvčí 5	82	84	94	96	96
Průměr	84	92	96	97	98

Uvedená tabulka byla vybrána protože DTW zde dosahuje nejlepších výsledků s použitím absolutního rozdílu vzdáleností. Tabulky pro ostatní typy vzdálenosti jsou uvedeny v příloze (B). Všechny následující vyhodnocení se týkají právě DTW a výpočtu matic vzdáleností absolutním rozdílem vzdáleností. Je vhodné uvést, jak moc se použití výpočtu matic vzdáleností liší. Graf úspěšností rozpoznávání je vidět na obrázku 4.9, přičemž je zde uvedena pro srovnání také lineární transformace.

Pokud v referenčním slovníku nejsou obsažena slova mluvčího, jehož obsah řeči je rozpoznáván, dojde k razantnímu propadu úspěšnosti jak je vidět v tabulce 3. Přitom v tomto případě je počet referenčních slov čtyřnásobný, neboť jako reference ke každému slovu je bráno 1-5verzí daného slova od všech mluvčích, kromě toho, jež se rozpoznává. Úspěšnost se tak pohybuje v mezích 50-60% a příliš nezáleží na velikosti referenčního slovníku. Délka srovnávání slov se slovníkem se s takto rostoucím počtem referencí značné zvyšuje.



Obrázek 4.9 Srovnání úspěšnosti rozpoznávání nelineární transformace (DTW) s různými výpočty matic vzdáleností a lineární transformace

Tabulka 3 Úspěšnost [%] DTW s použitím absolutního rozdílu vzdálenosti a obsahem 1-5 referenčních slov z každého slova od každého mluvčího mimo rozpoznávaného

Rozpoznávaný Mluvčí	Počet referenčních slov x 4				
	1	2	3	4	5
Mluvčí 1	66	60	60	58	60
Mluvčí 2	38	32	32	30	30
Mluvčí 3	64	58	62	68	68
Mluvčí 4	60	76	72	74	68
Mluvčí 5	48	52	50	50	62
Průměr	55	56	55	56	58

5. HTK TOOLKIT

HTK toolkit je pro trénování parametrů Markovových modelů. Nejde čistě o toolkit pro rozpoznávání řeči, ale právě pro tuto část je primárně určen. Jde o statistickou metodu, která je vhodná k porovnání úspěšnosti s realizovaným rozpoznávačem.

HTK je dostupné zdarma a pochází z University v Cambridge. Protože HTK je určeno k trénování skrytých Markovových modelů, v následujících podkapitolách se s těmito modely seznámíme.

5.1 SKRYTÉ MARKOVOVY MODELY – HMM

Markovovy modely se dají použít jak pro malé slovníky (stovky slov), tak pro velké slovníky (desítky tisíc slov). Rozdíl je v použitých Markovových modelech. U malých slovníků se modeluje každé slovo zvlášť, srovnání vstupního slova pak probíhá s každým modelem slova. Z výpočetních důvodů toto neleze realizovat pro velké slovníky. Proto se u nich používají modely menších slovních jednotek jako je například foném nebo kontextově závislý foném. Tento přístup lze také zvolit při řešení našeho problému, a to odhadem s dobrou úspěšností, ale lepších výsledků se jistě dosáhne kvalitním zpracováním rozpoznávače pro malý slovník.

Následující podkapitoly se budou týkat právě skrytých Markovových modelů pro jednotlivá slova.

5.2 ZÁKLAD HMM

Metoda HMM je statistická a určování pravděpodobnosti shody vzorů s modelem vychází z Bayesova vztahu

$$P(\omega_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_j)P(\omega_j)}{p(\boldsymbol{x})} \tag{5.1}$$

kde $P(\omega_j|\boldsymbol{x})$ je posterior třídy ω_j, známe-li datový vektro **x**,

$p(\boldsymbol{x}|\omega_j)$je likehood datového vektrou x známe-li třídu ω_j

$P(\omega_j)$je prior třídy ω_j

$p(\boldsymbol{x})$ je evidence (normalizační funkce)

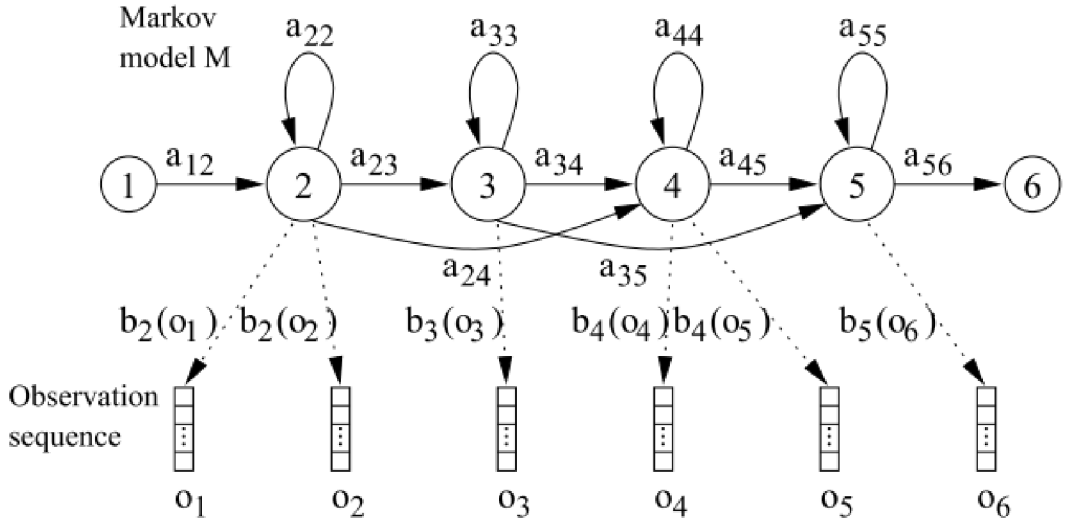
Pro nalezení slova nám přitom stačí nalézt maximální likehood podle níže uvedeného vzorce, neboť evidence $p(\boldsymbol{x})$ je stejná pro všechny, třídy a u jednoduchých rozpoznávačů se předpokládá, že priory všech tříd budou stejné, $P(\omega_j) = \omega_j$.

$$\omega_j^* = \max\, p(\boldsymbol{x}|\omega_j) \tag{5.2}$$

Třídy se modelují Gaussovými rozděleními (nebo směsí Gaussových rozdělení) tak, aby bylo možné rozpoznávat jednotlivé vektory slov při srovnávání. K natrénování modelů je ale zapotřebí dostatečné množství trénovacích dat (desítky až stovky slov pro jeden model).

5.3 STRUKTURA HMM

Struktura HMM je podobná konečnému stavovému automatu s tím rozdílem, že přechody mezi stavy jsou s pravděpodobnostním ohodnocením a každý stav vyhodnocuje vstupní data podle implementované funkce hustoty rozložení pravděpodobnosti. Názorná ukázka skrytého Markovova modelu je na obrázku 5.1.



Obrázek 5.1 Struktura HMM [3]

První a poslední stavy slouží pro připojení modelů, ostatní se nazývají vysílací a reprezentují nějaký vstupní vektor. Počet stavů závisí na délce vektoru parametrů, běžně se používá 39 vysílacích stavů.

Pravděpodobnosti přechodů a_{ij} zavádí, jak je pravděpodobné v modelu přeskočit ze stavu i do stavu j. Pro jednotlivé pravděpodobnosti vycházející ze stavu musí platit $\sum a_{ij} = 1$. Pravděpodobnosti setrvání ve stavu a_{ii} jsou zde z důvodu nestejné délky rozpoznávaného slova. Čím delší je stejné slovo, tím je větší pravděpodobnost, že se některé vektory parametrů budou opakovat. Naproti tomu pravděpodobnosti, že se nějaký stav modelu přeskočí, pravděpodobnosti jsou docla malé, proto se někdy volí implicitně jako nulové.

Hodnota funkce hustoty rozdělení pravděpodobnosti-likelihood b_j je dána většinou směsí Gaussových rozložení. Záleží na počtu prvků vstupní sekvence vektorů $\boldsymbol{o}(t)$. Běžně je ale vektor $\boldsymbol{o}(t)$ P-rozměrný (39x39). Výpočet likelihoodu je pak

$$b_j[\boldsymbol{o}(t)] = N\left(\boldsymbol{o}(t); \mu_j; \Sigma_j\right) = \frac{1}{\sqrt{(2\pi)^P |\Sigma_j|}} e^{-\frac{1}{2}\left(\boldsymbol{o}(t)-\boldsymbol{\mu}_j\right)^T \Sigma_j^{-1} \left(\boldsymbol{o}(t)-\boldsymbol{\mu}_j\right)} \quad (5.3)$$

Výpočet podle uvedeného vzorce je poměrně náročný, proto se zavádí zjednodušení rozdělení pravděpodobnosti a to takové, že se předpokládá, že parametry nejsou korelované. Pak kovarianční matice o rozměrech PxP je matice diagonální. Výpočet se zjednoduší na

$$b_j[o(t)] = \prod_{i=1}^{P} N\left(o(t); \mu_{ji}; \sigma_{ji}\right) = \prod_{i=1}^{P} \frac{1}{\sigma_{ji}\sqrt{2\pi}} e^{-\frac{\left[o(t)-\mu_{ji}\right]^2}{2\sigma_{ji}^2}} \quad (5.4)$$

Dalším zjednodušením výpočtů je použití stavů sdílených mezi modely tzn., že více modelů používá stejné nebo velmi podobné stavy. Ušetří se tak více paměti a HMM budeme mít méně parametrů.

5.4 NASTAVENÍ HTK

S HTK lze realizovat nejen rozpoznávače s omezeným slovníkem, ale mnohem náročnější a tomu také odpovídá rozsah manuálu **Chyba! Nenalezen zdroj odkazů.** a jednotlivých nastavení nástrojů. Pro účely rozpoznávání omezeného slovníku se tato nastavení zjednodušují.

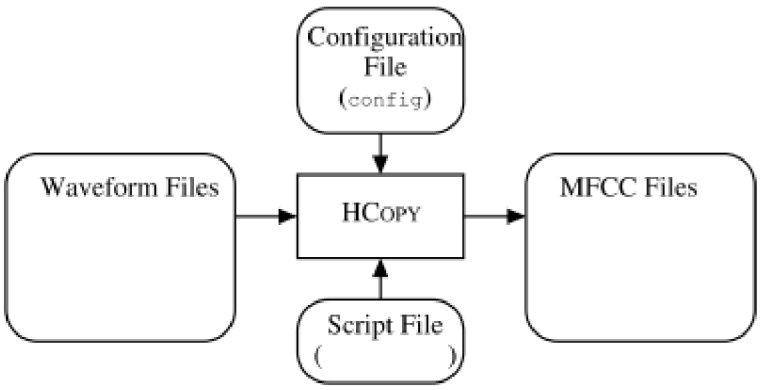
Většina parametrů pro předzpracování a parametrizaci je stejná jako použité

Většina parametrů pro předzpracování a parametrizaci je stejná jako použité parametry u realizovaného rozpoznávače

Aby řečové signály byly parametrizovány, jak je požadováno, použije se konfigurační soubor natavení.conf z podadresáře "cfg". Tento soubor obsahuje tyto specifikace

BYTEORDER=VAX	pořadí byte - Intel-PC
SOURCEKIND=WAVEFORM	druh zdroje WAV
SOURCEFORMAT=WAV	hlavička souboru WAV
SOURCERATE=1250	vzorkovací perioda zdroje ve 100ns
ZMEANSOURCE=TRUE	ustřednění signálu
TARGETKIND=MFCC_E_D_A	typ výstupních parametrů
TARGETFORMAT=HTK	výstupní formát HTK
TARGETRATE=100000	vzorkovací perioda výstupních vektorů ve 100ns
WINDOWSIZE=250000.0	délka okna 25ms (ve 100ns)
NUMCHANS=23	počet trojúhelníkových filtrů pro MFCC
ENORMALISE=TRUE	normování energie.

Rozdíl je pouze v parametrech popisujících jednotlivé segmenty. Každý segment je popsán nejen 13 mel-kepstrálními koeficienty, ale také 13 derivacemi a 13 druhými derivacemi těchto koeficientů. Parametrizace řeči v HTK probíhá, jak je ukázáno na obrázku 5.2.



Obrázek 5.2 Parametrizace v HTK [19]

Dalším krokem je definování trénovacích a testovacích dat. To se děje za pomoci Matlabu a spuštěním souboru se zvoleným nastavením *ZapisProHTK.m.* Tři podadresáře v adresáři Matlab jsou určeny pro nástroje HTK. Adresáře `Matlab/mlf` a `Matlab/scripts` se musí po vykonání programu *ZapisProHTK.m* zkopírovat do stejnojmenných podadresářů v adresáři HTK, tedy do `HTK/mlf` a `HTK/scripts`.

Třetím adresářem je `Matlab/Lab`. V něm jsou obsaženy soubory popisující délku jednotlivých nahrávek a jejich význam (slovo). Tyto soubory se generují po spuštění souboru *ZapisProHTK.m* pro všechny používané trénovací a testovací data. Protože jsou již obsažena implicitně pro všechna slova v adresáři `HTK/data` (`pozor, ne HTK/Lab`), tak jejich kopírování je nutné pouze při smazání z tohoto adresáře nebo pokud data jsou pozměněna, případně přidána nová.

Parametrizace, trénování a rozpoznávání se děje pomocí příkazů v příkazovém řádku nebo spuštěním dávkového souboru. Potřebné příkazy zde pro možnost zopakování testů jsou uvedeny [2]:

- Parametrizace

  HCopy -T 1 -C cfg\nastaveni.conf -S scripts\trainm.scp

  HCopy -T 1 -C cfg\nastaveni.conf -S scripts\testm.scp

- Trénování modelů, ukázka pro slovo approach (nutné udělat pro všechna slova!)
  - inicializace modelů

HCompV -T 7 -I mlf\trainm.mlf -l approach -m -S scripts\trainm_htk.scp -M hmm0 proto\approach

  - přetrénování modelů

HRest -T 7 -I mlf\trainm.mlf -l approach -S scripts\trainm_htk.scp -M hmm1 hmm0\approach

- Rozpoznávání (toto je jeden příkaz!)

  HVite -T 1 -d hmm1 -S scripts\testm_htk.scp -i mlf\testout.mlf -w net\network dics\dictionary lists\models

- Vyhodnocení

  Hresults -I mlf\testm.mlf lists\models mlf\testout.mlf

Výhodné je si všechny tyto příkazy dát do dávkového souboru a spouštět najednou. Úspěšnost rozpoznávání je po správném použití nástroje *HResults* vypsána v příkazovém řádku jako *ACC* (word accuracy).

5.5 VÝSLEDKY ROZPOZNÁVAČE HTK

Pro rozpoznávání obsahu slov stejného mluvčího s jeho vlastními trénovacími slovy rozpoznávač nebyl schopen natrénovat model, nicméně, při třech trénovacích slovech každého rozpoznávaného slova byly výsledky téměř bezchybné, viz. tabulka 4.

Tabulka 4 Úspěšnost [%] HTK s 1-5 trénovacími slovy každého slova stejného mluvčího

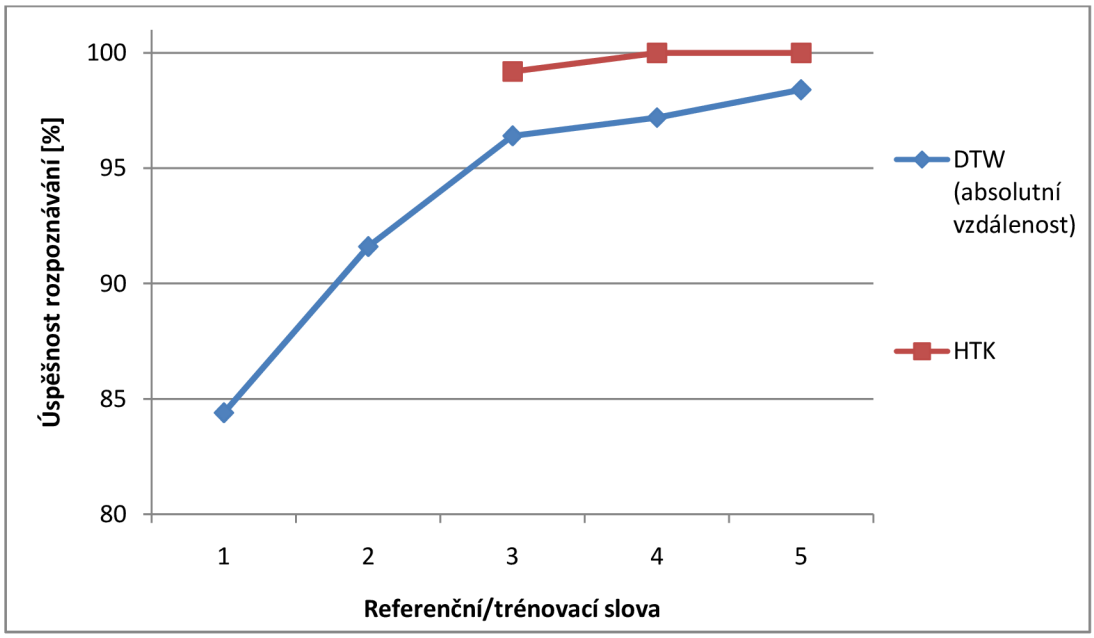
Mluvčí	Počet trénovacích slov stejného mluvčího				
	1	2	3	4	5
Mluvčí 1	---	---	100	100	100
Mluvčí 2	---	---	100	100	100
Mluvčí 3	---	---	100	100	100
Mluvčí 4	---	---	96	100	100
Mluvčí 5	---	---	100	100	100
Průměr	---	---	99	100	100

Pokud rozpoznávač nebyl trénován se slovy rozpoznávaného mluvčího, úspěšnost se snížila, ale ne tak razantně, jako u rozpoznávače DTW. Odpovídá tomu tabulka 5. Trénovací slova zde byla od 1-5 a to od každého mluvčího kromě rozpoznávaného. Počet trénovacích slov tedy byl minimálně čtyři a maximálně 20 pro jedno slovo. V případě většího počtu mluvčích a slov by úspěšnost rozpoznávání rostla.

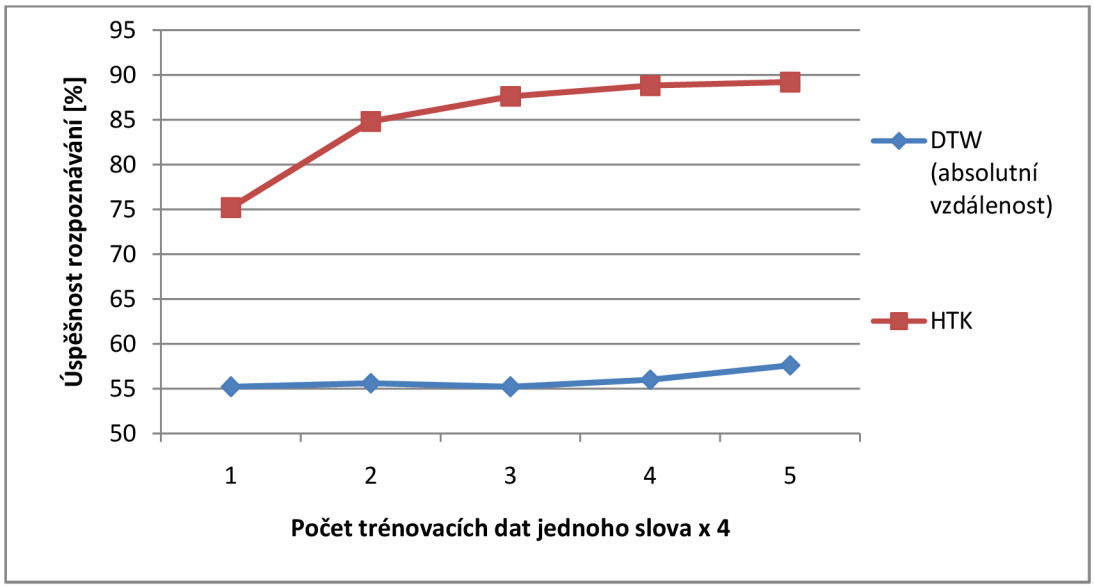
Tabulka 5 Úspěšnost [%] HTK s 1-5 trénovacími slovy pro každé slovo od každého mluvčího mimo rozpoznávaného

Rozpoznávaný Mluvčí	Počet referenčních slov x 4				
	1	2	3	4	5
Mluvčí 1	94,00	94,00	98,00	96,00	96,00
Mluvčí 2	50,00	62,00	60,00	68,00	74,00
Mluvčí 3	88,00	92,00	94,00	92,00	92,00
Mluvčí 4	74,00	90,00	92,00	92,00	94,00
Mluvčí 5	70,00	86,00	94,00	96,00	90,00
Průměr	75,20	84,80	87,60	88,80	89,20

6. GRAFICKÉ POROVNÁNÍ ÚSPĚŠNOSTI OBOU ROZPOZNÁVAČŮ



Obrázek 6.1 Úspěšnost rozpoznávání s 1-5 referenčními/trénovacími slovy každého slova ve slovníku u stejného mluvčího



Obrázek 6.2 Úspěšnost rozpoznávání s 1-5 referenčními/trénovacími slovy všech mluvčích (x4) mimo rozpoznávaného

7. ZÁVĚR

Po získání přehledu z oblasti letového provozu a rozpoznávání řeči byly vybrány dva řečové rozpoznávače, které by požadavky rozpoznávání ATC povelů mohly splňovat. Jedná se o rozpoznávač založený na porovnávání se vzory (DTW) a volně dostupný toolkit HTK založený na statistických metodách a využívající skryté Markovovy modely.

Rozpoznávač DTW dosahuje poměrně vysoké úspěšnosti s vhodně zvoleným referenčním slovníkem pro všechny způsoby výpočtů matic vzdáleností. Nejlepší výsledky jsou ovšem dosaženy při výpočtu těchto matic vzdáleností absolutním rozdílem vzdálenosti. Proto byla s touto volbou provedena testování pro srovnání s toolkitem HTK. DTW se podle testů řadí k velice úspěšné metodě pro rozpoznávání slov stejného mluvčího. Průměrná hodnota pro pět testovacích mluvčích přesahuje 84% již při jednom referenčním vzoru každého slova. Pro počet pěti referenčních slov tato úspěšnost přesáhne 98%. Bohužel pokud nejsou ve slovníku slova mluvčího, jehož slova jsou rozpoznávána, úspěšnost se pohybuje v rozmezí 50-60%. Přidáváním slov jiného mluvčího do referenčního slovníku se výsledky nijak výrazně nemění. Pokud slovník obsahuje nejen slova od rozpoznávaného mluvčího, ale také od mluvčích ostatních, úspěšnost se zvýší na 93%.

HTK toolkit pracuje na základě statistických metod. Při testech úspěšnosti pro jednoho mluvčího bylo dosaženo téměř 100%, ale je nutné použít jako trénovací data alespoň tři referenční slova. Pokud není použito k trénování modelu slov mluvčího, jehož slova jsou rozpoznávána, výsledky dosahují 75-89%, podle počtu trénovacích dat. Trénovací data byla vytvořena použitím jednoho až pěti slov každého z mluvčích s výjimkou mluvčího rozpoznávaného. V případě trénovacích dat od všech mluvčích, včetně rozpoznávaného, bylo dosaženo kvality přes 99%.

Z výše uvedeného lze shrnout, že výsledky pro rozpoznávání slov u jednoho mluvčího jsou u obou metod téměř stejné, ale k velké odlišnosti s rozpoznáváním dochází při chybějících slovech v referenčním slovníku (trénovacích slovech) rozpoznávaného mluvčího. Statistické metody rozpoznávání dosahují lepších výsledků a pro praktické zavedení jsou vhodnější. Jediný požadavek kladený na tyto

metody je mít dostatečný počet trénovacích dat. Statistické metody jsou navíc vhodné nejen pro malé slovníky, ale i pro značně obsáhlé. DTW se dá použít pouze tam, kde je požadavek jednoho mluvčího a dobrá úspěšnost rozpoznávání. Za jiných podmínek by musel referenční slovník obsahovat neúnosné množství slov a rozpoznávání by již bylo časově příliš náročné.

LITERATURA

[1] Adamec, M. – *Moderní rozpoznávání řečové aktivity.* Diplomová práce, VUT Brno 2008.

[2] Černocký, J.: *Skryté Markovovy modely (HTK)*, cvičení předmětu ZRE, fakulta FIT, VUT, 2008 http://www.fit.vutbr.cz/study/courses/ZRE/public/labs/09_htk/09_htk.pdf

[3] Černocký, J.: *Zpracování řečových signálů-studijní opora*. Elektronická skripta předmětu ZRE, VUT Brno 2007.

[4] *HTK book*, 2008 [online]. http://htk.eng.cam.ac.uk/prot-docs/htk_book.shtml

[5] Chaloupka, J.: *Signály a informace, Přednáška č.13 – Měření podobnosti signálů, úvod do rozpoznávání.*, 2007 [online]. http://www.fm.vslib.cz/~kes/data/sgipr_13.pdf

[6] Kepka, J. – Psutka, J.: *Kombinace příznakových a strukturálních metod rozpoznávání,* ZČU (Západočeská univerzita), Plzeň, 1994, ISBN 80-7082-131-0.

[7] Kmitocty.cz: *Letecká radiová komunikace a radionavigace (základy)* [online]. http://www.kmitocty.cz/technika/menu.htm

[8] Kotek, Z. – Mařík, V. – Hlaváč, V. – Psutka, J. – Zdráhal, Z.: *Metody rozpoznávání a jejich aplikace.*, Academia, Praha, 1993, ISBN 80-200-0297-9.

[9] Kulčák, J. : *Přípravný kurz z letecké frazeologie s využitím internetu.*, CERM, Brno, 2007, ISBN 978-80-7204-531-0.

[10] Malenovský, V.: *Adaptivní filtrace zašuměných řečových signálů-Web page* [online]. http://www.elektrorevue.cz/clanky/02063/index.html

[11] Ministerstvo dopravy České republiky: *Letecký předpis, Radiotelefonní postupy a letecká frazeologie a terminologie pro poskytování letových provozních služeb a provádění letů, L frazeologie* [online]. http://lis.rlp.cz/predpisy/predpisy/dokumenty/L/L-Frazeologie/index.htm

[12] Nouza, J.: Počítačové rozpoznávání řeči [online]. http://archiv.computerworld.cz/cwarchiv.nsf/clanky/75BCB9905F09B63FC12569B000520254?OpenDocument

[13] Plšek, M. – Vondra, M.: *Detekce základního tónu v zašumělých řečových nahrávkách* [online]. http://www.elektrorevue.cz/clanky/02025/index.html

[14] Pollák, P. – Čmejla, R.: *Analýza a zpracování řečových a biologických signálů, Sborník prací.*, ČVUT, 2007 [online]. http://noel.feld.cvut.cz/sbornik07/data/sbornik-2007.pdf

[15] Prchal, J. – Šimák, B.: *Digitální zpracování signálu v telekomunikaci, kapitola 3-Vzorování, obnova a kvantování.*, 2004 [online]. http://www.comtel.cz/files/download.php?id=3361

[16] Psutka, J. – Muller, L. – Matoušek, J. – Radová, V.: *Mluvíme s počítačem česky.*, Academia, Praha, 2006, ISBN 80-200-1309-1.

[17] Psutka, J.: *Komunikace počítače mluvenou řečí.,* Academia, Praha, 1995, ISBN 80-200-0203-0.

[18] Sigmund, M.: *Rozpoznávání řečových signálů-přednášky.,* VUT FEKT, Brno, 2007, ISBN 978-80-214-3526-1.

[19] Uhlíř, J. – Sovka, P. – Pollák, P. – Hanžl, V. – Čmejla, R.: *Technologie hlasových komunikací.*, Nakladatelství ČVUT, 2007, ISBN 978-80-01-03888-8.

[20] Vladimír, M. – Štěpánková, O. – Lažanský, J., a kol.: *Umělá inteligence (2)*, Academia, Praha, 1997, ISBN 80-200-0504-8.

SEZNAM POUŽITÝCH ZKRATEK A SYMBOLŮ

ATC - Air Traffic Control

DTW - Dynamic Time Warping, dynamické "borceni" času

HMM - Hidden Markov Model, skryté Markovovy řetězce

HTK - Hidden Markov Model Toolkit

LPC – Linear Predictive Coding, lineární predikce

MFCC - Mel-frequency cepstral coefficients, Mel-frekvenční kepstrální koeficienty

PLP – Perceptual Linear Prediction

RASTA - RelAtiveSpectTrAl processing

SEZNAM PŘÍLOH

PŘÍLOHA (A) - Značení zvukových záznamů

Každý záznam je zapsán ve formátu Mluvčí_Slovo_VerzeSlova.wav. Kódování mluvčích a slov naleznete v tabulce 6 a tabulce 7.

Tabulka 6 Kódování mluvčích

Mluvčí	Kód
Mluvčí 1	001
Mluvčí 2	002
Mluvčí 3	003
Mluvčí 4	004
Mluvčí 5	005

Tabulka 7 Kódování slov

Slovo	Kód
approach	1
climb	2
contact	3
descent	4
flight	5
ground	6
level	7
position	8
report	9
request	10

PŘÍLOHA (B) - dodatečné tabulky ke kapitole 4.5

Tabulka 8 Úspěšnost [%] DTW s použitím Čebyševovy vzdálenosti a obsahem 1-5 referenčních slov každého slova stejného mluvčího

Mluvčí	Počet referenčních slov stejného mluvčího				
	1	2	3	4	5
Mluvčí 1	66	76	88	96	100
Mluvčí 2	64	86	88	92	96
Mluvčí 3	82	86	94	92	96
Mluvčí 4	70	84	92	94	98
Mluvčí 5	80	84	86	90	98
Průměr	72	83	90	93	98

Tabulka 9 Úspěšnost [%] DTW s použitím Euklidovy vzdálenosti a obsahem 1-5 referenčních slov každého slova stejného mluvčího

Mluvčí	Počet referenčních slov stejného mluvčího				
	1	2	3	4	5
Mluvčí 1	78	86	100	100	100
Mluvčí 2	72	90	92	94	96
Mluvčí 3	90	90	96	94	98
Mluvčí 4	92	94	96	96	98
Mluvčí 5	80	84	94	98	94
Průměr	82	89	96	96	97

Tabulka 10 Úspěšnost [%] DTW s použitím kvadrátu Euklidovy vzdálenosti a obsahem 1-5 referenčních slov každého slova stejného mluvčího

Mluvčí	Počet referenčních slov stejného mluvčího				
	1	2	3	4	5
Mluvčí 1	74	86	90	100	100
Mluvčí 2	72	86	92	94	96
Mluvčí 3	76	86	90	94	98
Mluvčí 4	76	82	94	86	86
Mluvčí 5	78	80	86	88	94
Průměr	75	84	90	92	95

PŘÍLOHA (C) - Obsah datového disku

Dokumentace	- elektronická verze diplomové práce
Matlab	- zdrojové kódy a adresáře pro realizaci rozpoznávače DTW
	- zdrojový kód a adresáře k podpoře HTK (viz kapitola 5.4)
HTK	- kompilované nástroje pro rozpoznávač založený na statické metodě, využívající skryté Markovy modely