VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

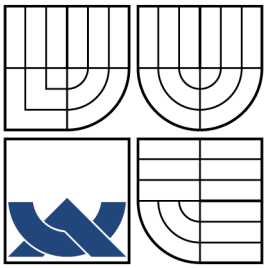
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

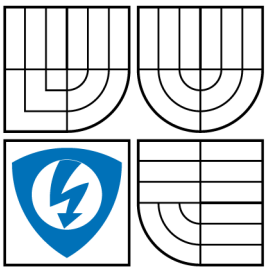
SYSTÉM PRO DIARIZACI MLUVČÍCH

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE JOSEF BRADÁČ
AUTHOR



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

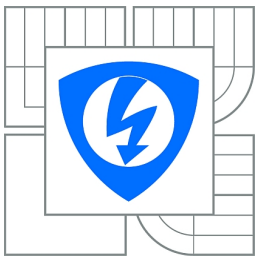
SYSTÉM PRO DIARIZACI MLUVČÍCH
SPEAKER DIARIZATION SYSTEM

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — JOSEF BRADÁČ
AUTHOR

VEDOUCÍ PRÁCE — Ing. IVAN MÍČA
SUPERVISOR

BRNO 2011



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Bakalářská práce

bakalářský studijní obor
Teleinformatika

Student:	Josef Bradáč	***ID:***	125374
Ročník:	3	***Akademický rok:***	2011/2012

NÁZEV TÉMATU:

Systém pro diarizaci mluvčích

POKYNY PRO VYPRACOVÁNÍ:

Seznamte se s technikami segmentace mluvčích v neznámých řečových nahrávkách. V jazyce Octave nebo Matlab pak implementujte jednoduchý systém pro diarizaci mluvčích. Vhodnou metodou a s pomocí databáze řečových nahrávek vyhodnoťte účinnost navrženého systému.

DOPORUČENÁ LITERATURA:

[1] ANGUERA, X., et al. Speaker diarization: A review of Recent Research. IEEE Transactions on acoustics speech and language processing 2011.

[2] NISHIDA, M., YAMAMOTO, S. Speaker clustering based on non-negative matrix factorization. Proceedings of Interspeech 2011. September 2011. ISSN 1990-9772.

Termín zadání:	6.2.2012	***Termín odevzdání:***	31.5.2012

Vedoucí práce: Ing. Ivan Míča
Konzultanti bakalářské práce:

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

ABSTRAKT

Systém pro diarizaci mluvčích má široké uplatnění na poli zpracování a analýzy řečových signálů. V této práci je rozebrán úvod do problematiky a následný postup pro navržení systému. Výsledkem práce je implementace samotného systému a jeho vyhodnocení na základě databáze nahrávek rozhovorů.

KLÍČOVÁ SLOVA

diarizace mluvčích, VAD, ROC, energie, průchody nulou, základní tón hlasu, MFCC, delta koeficienty, k-nn, k-means, segmentace, clustering

ABSTRACT

Speaker diarization system has wide application in the field of processing and analysis speech signals. This work is broken down to introduction and follow for designing the system. Result of this work is an implementation of the system itself and its evaluation based on interview´s database.

KEYWORDS

speaker diarization, VAD, ROC, energy, zero passages, basic tone of voice, MFCC, delta coefficient, k-nn. k-means, segmentation, clustering

BRADÁČ, Josef *Systém pro diarizaci mluvčích*: bakalářská práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2011. 39 s. Vedoucí práce byl Ing. Ivan Míča

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma „Systém pro diarizaci mluvčích“ jsem vypracoval samostatně pod vedením vedoucího bakalářské 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 bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

Brno ....

(podpis autora)

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu semestrální práce panu Ing. Ivanu Míčovi za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Brno ....

(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Cílem bakalářské práce je nastínit problematiku systémů pro diarizaci mluvčích a z těchto znalostí poté navrhnout a naprogramovat systém pro rozpoznávání mluvčích. Po krátkém úvodu se zaměříme na výpočet příznaků v časové oblasti, jako jsou velikost energie, počet průchodů nulou a základní tón hlasu. Dále budeme pokračovat do frekvenční oblasti, kde se zaměříme na Mel frekvenční kepstrální koeficienty a znich určíme jejich dynamické koeficienty. Po této fázi se nabízí více možností. Nastíníme základní postupy klasifikace a následné selekce příznaků. Jako poslední bod systému bude clustering a segmentace mluvčích.

Z tohoto přehledu vybereme jeden způsob a podle něj se pokusíme vytvořit systém pro diarizaci mluvčích. Na závěr systém otestujeme na databázi nahrávek, abychom mohli vyhodnotit jeho úspěšnost.

1 DIARIZACE MLUVČÍCH

Systém pro diarizaci mluvčích byl poprvé představen v roce 2003 v National Institute of Standards and Technology (NIST) [1]. Základem tohoto systému je určit kdy kdo mluví v audio záznamu, o kterém nevíme, kolik množství řeči obsahuje a také kolik je v něm aktivních mluvčích.

Původně byl navržen jako pomůcka pro automatické rozpoznávání řeči, ale časem se začal používat ve více oblastech. Dnes se tento systém využívá např.:

- Televizní vysílání
- Aplikační domény
- Přednášky
- Besedy
- Telefonní hovory

Jak se také můžeme dočíst v článku [2], diarizaci lze rozdělit na:

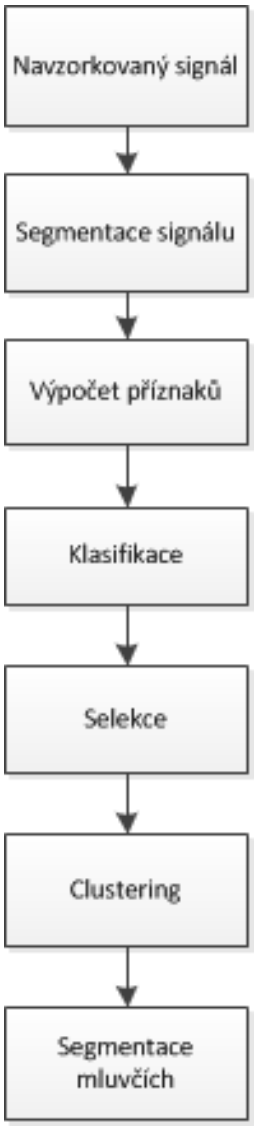
- Rozpoznávání z audio signálu
- Rozpoznávání z video signálu
- Kombinace obou

Další rozdělení by se mohlo provést podle toho, jestli systém používáme za běhu, tzv. online a nebo z nahrávky - offline. V našem případě so budeme zabývat pouze systémy pro rozpoznání mluvčích z audio signálu při offline režimu.

Systém má velké využití a zároveň se liší pro jednotlivé použití. Někdy je potřeba ze signálu vyfiltrovat ruch okolí, např. člověk, který stojí na hlučném nádraží a potřebuje vyřídit telefonní hovor, tak aby mu člověk na druhé straně spojení rozuměl, musí mobil potlačit šum a zvýraznit řečový signál. Pokud ovšem budeme používat diarizaci mluvčích na besedě s více mluvčími, a budeme pořizovat záznam, můžeme šum okolí zanedbat, protože se nepředpokládá, že posluchači budou hluční. Naopak problém nastává, když si mluvčí skáčou do řeči. Systém má pak za úkol označit jednotlivé mluvčí a zpracovat signál tak, aby se řeč více lidí neprolínala, ale byla oddělená.

Z toho plyne, že existuje spoustu možností, jak navrhnout a vytvořit systém pro diarizaci mluvčích. Zatím není možné vytvořit univerzální systém, který by měl 100 % účinnost. Vždy to bude něco na úkor něčeho. Jelikož lidský hlas je jedinečný a každý člověk ho má trochu jiný, nedá se předpokládat, že systém, který pracuje s digitálním signálem, dokáže rozpoznat hlasy jednotlivých lidí lépe, než samotný člověk.

Návrh systému je poměrně komplikovaná záležitost a nelze proto pokaždé postupovat stejným způsobem. Přesto je spoustu rysů a úprav stejných a je vhodné



Obr. 1.1: Návrh systému pro diarizaci mluvčích

je provádět pro úspěšné sestavení systému. Postup, kterého se v této práci budeme z větší části držet, je znázorněn na obrázku 1.1.

2 PŘÍZNAKY

2.1 Energie signálu

Pokud budeme předpokládat, že okolní hluk nebude příliš výrazný a jeho energie bude výrazně menší než energie řečového projevu mluvčích, je tento příznak vhodný. Energie signálu [4] se počítá v každém segmentu zvlášť. Pro správné určení řečových a neřečových segmentů se stanoví mezní práh, který určuje právě hranici mezi těmato dvěma stavy. Velikost energie vypočítáme ze vztahu:

$$E[i] = \frac{1}{N} \sum_{n=0}^{N-1} (x[n])^2. \tag{2.1}$$

Je to rychlý způsob pro detekci řečové aktivity (VAD), ale je dobré k němu přidat ještě další příznaky, které výpočet zpřesní. Velikost energie signálu se také používá pro výpočet základního tónu hlasu (2.3), kde potřebujeme rozpoznat znělý a neznělý úsek.

2.2 Počet průchodů nulou

U tohoto příznaku se nedá předpokládat velká přesnost. Rozhodnutí se dělá tak, že pokud má segment málo průchodů nulou, je považován za řečový [4]. Pokud má mnoho průchodů, jedná se o šum. Z části je to pravda, ovšem platí to pouze u znělých úsecích v řeči. Neznělé hlásky mají také velký počet průchodů nulou a tak nelze předpokládat velkou úspěšnost tohoto příznaku při použití u VAD. Jeho hlavní využití v této práci je spolu s příznakem velikosti energie v počítání základního tónu hlasu (2.3) taktéž pro rozpoznání znělého a neznělého úseku. Pro výpočet počtu průchodu nulou je dán vztah:

$$ZCR[i] = \sum_{n=0}^{N-1} |\mathrm{sgn}(s[n]) - \mathrm{sgn}(s[n-1])| \tag{2.2}$$

2.3 Základní tón hlasu

Jedním z důležitých parametrů pro diarizaci mluvčích je základní tón hlasu, o kterém pojednává článek [4]. Je to základní parametr řečového signálu v kmitočtové oblasti. Během mluvení se tón hlasu mění a frekvence hlasu se pohybuje v rozmezí jedné oktávy. Vypočítáním průměrné hodnoty základního tónu jsme schopni určit pohlaví mluvčího, případně o kterého mluvčího se jedná.

Základní kmitočet se udává 60–400 Hz, ale liší se u mužů, žen i dětí. U mužů je průměrná hodnota základního tónu 132 Hz, u žen 223 Hz a u dětí se pohybuje

v rozmezí 200–600 Hz. Základní kmitočet lidského hlasu můžeme vypočítat ze vztahu

$$F_0 = 1/T_0. \tag{2.3}$$

Základní tón řeči má široké využití. Používá se např. pro analýzu řeči, detekce emočního stavu mluvčího, identifikaci mluvčího a detekci hlasové aktivity. Detekci základního tónu řeči provádíme v:

- Časové oblasti
- Kmitočtové oblasti
- Kepstru

Součástí pro určení základního tónu řeči je charakter segmentu řeči [4], protože právě z něj se základní tón počítá. Máme dvě možnosti. Buď narazíme na neznělý úsek řeči (obrázek 2.1), v kterém ovšem nemá smysl počítat základní tón, protože signál je neperiodický a výsledek by byl zkreslený. To znamená, že je potřeba se soustředit pouze na znělé úseky řeči (obrázek 2.2), z kterých můžeme vypočítat F_0.

Je prokazatelné, že znělý úsek má méně průchodů nulou a větší krátkodobou energii, zatímco u neznělého úseku je to naopak. Tyto dva příznaky můžeme vyjádřit pro i-tý segment jako (2.1) a (2.2).

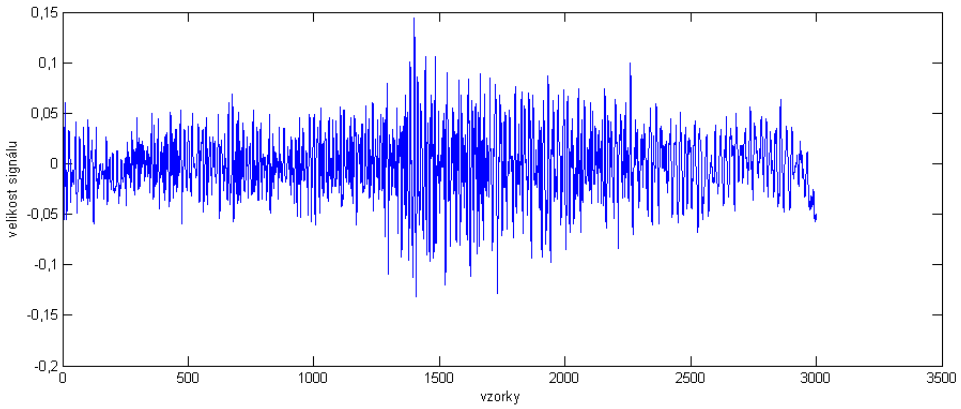
Detekce základního tónu řeči spočívá na výpočtu autokorelační funkce AKF, která určuje míru podobnosti v rámci jednoho signálu. Čím má AKF větší hodnotu, tím si je funkce v daném časovém posunu více podobná. To slouží také pro zjišťování periodicity signálu. S nadhledem lze říci, že v místech lokálních maxim AKF je začátek další periody. Z toho také vyplývá, že nejvyšší hodnotu bude mít AKF v nulovém časovém posunu. Vzhledem k tomu, že hledáme periodu signálu, nemůžeme použít pro výpočet neznělý úsek řeči. Ten se totiž chová jako šum (obrázek 2.1). Proto potřebujeme znělý úsek (obrázek 2.2) u kterého určíme první vrchol který následuje po maximální hodnotě AKF a opakující se vrcholy v okamžicích $\frac{f_{vz}}{F_0}$, $\frac{2f_{vz}}{F_0}$, $\frac{3f_{vz}}{F_0}$..., kde f_{vz} je vzorkovací kmitočet a F_0 je hledaný základní tón řeči.

2.3.1 Metoda centrálního klipování

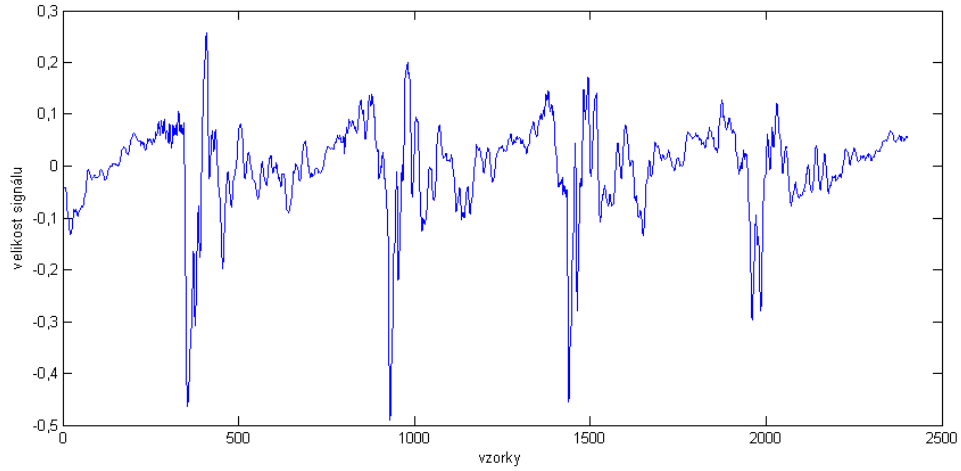
Tato metoda vychází z faktu, že k určení základního tónu řeči stačí znát pouze jednotlivé špičky v průběhu řeči.

Postup:

1. segmentace řečového signálu
2. vypočet prahu pro jednotlivé segmenty
3. normalizace signálu
4. AKF
5. výpočet kmitočtu základního tónu ze znělého úseku



Obr. 2.1: Neznělý úsek řečového signálu



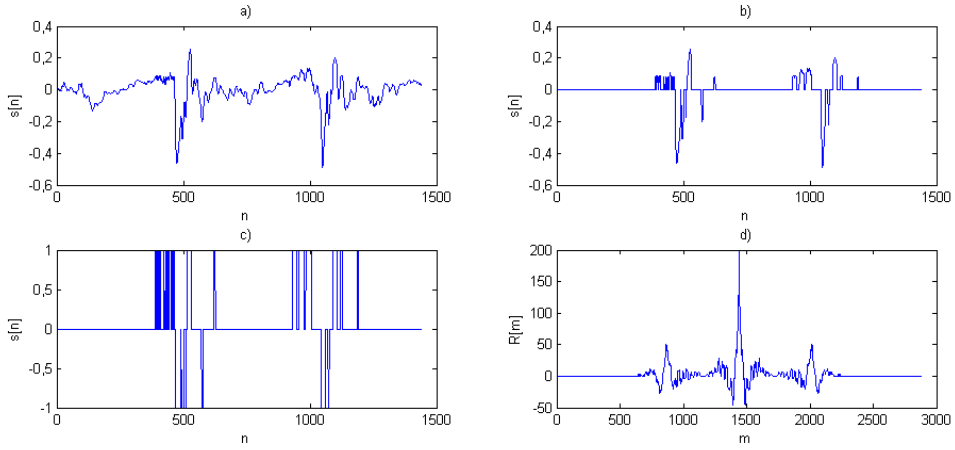
Obr. 2.2: Znělý úsek řečového signálu

Výpočet prahu pro jednotlivé segmenty provádíme z toho důvodu, že signál kolísá a nelze stanovit pevnou hodnotu prahu pro celý signál. Pro výpočet prahu i-tého segmentu P_i nejdříve určíme maxima v sousedních segmentech, Max_{i-1} a Max_{i+1} a P_i určíme ze vztahu

$$P_i = k \min(Max_{i-1}, Max_{i+1}), \tag{2.4}$$

kde k je redukční faktor, který má obvykle hodnotu 0,8. Normalizace se provádí na jednotkovou velikost 1, 0, -1.

Autokorelaci signálu provádíme pouze pro znělé úseky, ze kterých se základní tón počítá. Samotný základní tón hlasu se určí z autokorelační funkce, která je symetrická. Vycházíme z poloviny vektoru, kde je maximální shodnost signálu. Hledáme první maximum po průchodu nulou. Díky symetrii funkce si můžeme zvolit, kterou



Obr. 2.3: Postup výpočtu F0 metodou centrálního kliování, vstupní segment (a), vstupní segment po prahování (b), vstupní segment po klipování (c),oboustranná autokorelační funkce klipovaného signálu (d).

polovinu vektoru budeme počítat. Rozdíl nalezených maxim udává periodu signálu. Z toho poté vypočítáme základní tón hlasu jako podíl vzorkovacího kmitočtu a rozdílu maxim v korelační funkci. Postup výpočtu je také znázorněn na obrázku 2.3.

2.4 Příznaky ve frekvenční oblasti

Vhodným příznakem pro určování mluvčích jsou také spektrální nebo kepstrální koeficienty [10]. Jelikož v časové ose nelze provádět potřebné úpravy řečového signálu, převedeme si jej do frekvenční oblasti. Nejprve provedeme frekvenční analýzu signálu. Můžeme k tomu použít Diskrétní Fourierovu transformaci (DFT). Tu lze vypočítat ze vztahu:

$$X(k) = \sum_{n=0}^{N-1} x[n] e^{-j2\pi \frac{nk}{N}}, prok \in \langle 0, N-1 \rangle \tag{2.5}$$

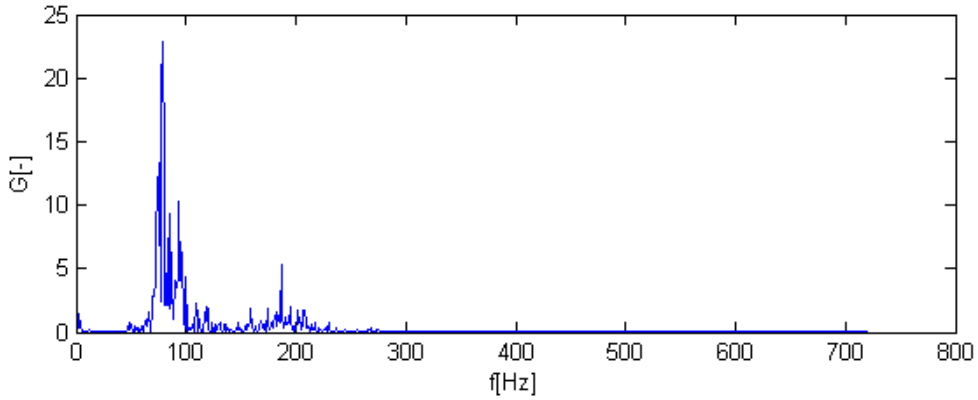
Vypočítané spektrum bude periodické podle vzorkovacího kmitočtu. Také se diskrétní s hodnotami kmitočtů vzdálených o velikosti $f_v = F_s/N$. Vzhledem k symetrii nám bude pro další výpočty stačit první polovina spektra.

Vzorků ve spektru je omezený počet. Pokud ale chceme nebo potřebujeme tento počet zvýšit, slouží k tomu metoda Zero padding. Nejedná se o nic jiného, než že do signálu přidáme nulové vzorky navíc a tím zvýšíme jeho hustotu a tímpádem i přesnost. Informace v signálu se nijak nezmění, pouze se zvýší počet vzorků.

Výkonová spektrální hustota nám slouží k analýze náhodného signálu a zjišťuje rozdělení výkonu ve frekvenční oblasti [10] (obrázek 2.4). Celý výpočet spočívá v tom,

že umocníme DFT koeficienty:

$$G_{DFT}(k\Delta f) = \frac{1}{N} |X[k]|^2 . \tag{2.6}$$



Obr. 2.4: Výkonová spektrální hustota

Obecně lze řečový signál rozdělit na buzení a modifikaci [10]. Buzení je dáno základním frekvencí hlasu mluvčího a modifikace souvisí s artikulačním traktem. U rozpoznávání je potřeba pouze modifikace, takže se musíme zbavit buzení. To však je obsaženo v celém spektru u vyšších harmonických. Řečový signál v čase je dán konvolucí buzení $q(t)$ a modifikace $h(t)$:

$$s(t) = q(t) * h(t) = \int_{-\infty}^{\infty} g(\tau) \cdot h(t-\tau) d\tau. \tag{2.7}$$

Po přenesení signálu pomocí DFT do frekvenční oblasti se z konvoluce stane součin:

$$S(f) = Q(f) \cdot H(f). \tag{2.8}$$

Ale ani ve frekvenční oblasti složky buzení a modifikace od sebe neoddělíme. Musíme tedy udělat spektrum spektra, tzv. kepstrum. V kepstru se z konvoluce stane součet a tudíž máme složky oddělené. U kepstrálních koeficientů se buzení vyskytuje na nižších a modifikace vyšších hodnotách. Ovšem vzhledem k lidskému uchu, které je ciltivé právě na nižší kmitočty a na vyšší už tolik nereaguje, musíme provést analýzu spektra právě na nižších kmitočtech.

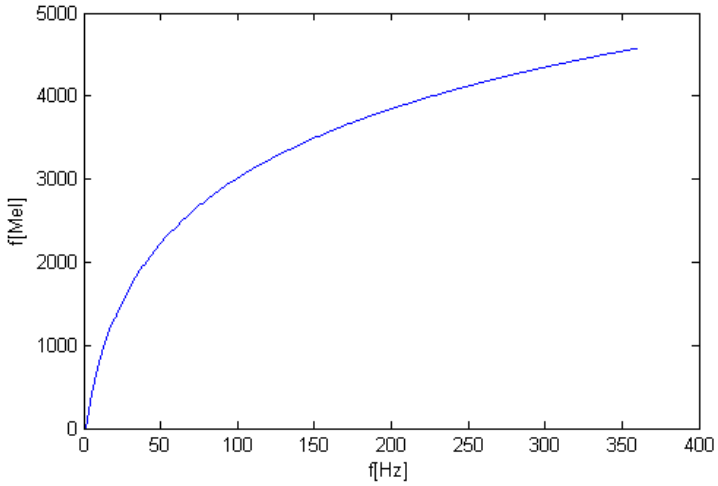
2.4.1 MFCC

Metoda MFCC právě uvažuje lidské slyšení na nižších kmitočtech [11]. Postup je takový, že na frekvenční osu rozmístíme nelineárně filtry a měříme energii na jejich

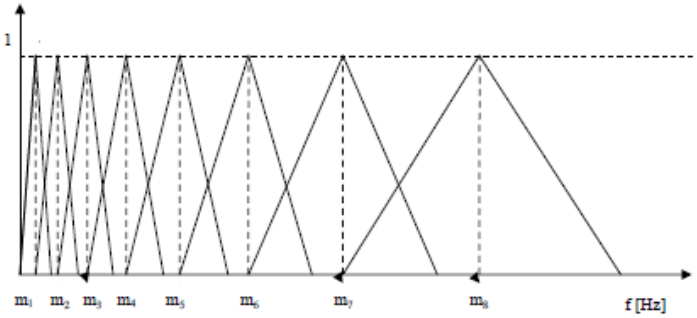
výstupu, kterou použijeme místo DFT pro výpočet kepstra. Frekvenční osu přepočítáme na melovskou. Jednotka Mel právě vychází ze slova melody, což charakterizuje právě slyšení lidského ucha. Použitá úprava pro převod Hertzů na Mely je:

$$F_{Mel} = 2959 \log_{10}(1 + \frac{F_{Hz}}{700}). \tag{2.9}$$

Filtry tvoří trojúhelníková okna s maximem v bodě jedna. To znamená, že spektrum je filtrem na krajích hodně potlačeno a uprostřed zůstává nezměněné. Počet filtrů je libovolný a minimálně je roven počtu koeficientů. Nejvíce informací o zvuku je obsaže v prvních několika koeficientech. Většinou se používá 10-14 koeficientů. Pokud na melovou osu rozmístíme filtry lineárně, tak na frekvenční ose budou rozmístěné nelineárně (obrázek 2.5). Rozmístění filtrů je znázorněno na obrázku 2.6.



Obr. 2.5: Závislost Hertzů na Melech



Obr. 2.6: Rozmístění Mel banky filtrů

Pro výpočet MFCC použijeme již vypočítanou výkonovou spektrální hustotu, kterou vynásobíme vytvořenou bankou filtrů, sestavenou z trojúhelníkových oken a sečteme. Poté provedeme zpětnou Cosinovou transformaci (IDCT), která nahrazuje zpětnou Fourierovu transformaci (IFFT):

$$c_{mf}(n) = \sum_{i=1}^{K} \log m_k \cos\left[n(k-0.5)\frac{\pi}{K}\right], \tag{2.10}$$

Kde K značí počet filtrů, e_k je energie na intervalu daného filtru a M je počet melovských kepstrálních koeficientů. Po této úpravě dostáváme výsledek v podobě Mel frekvenčních kepstrálních koeficientů (MFCC). Počet koeficientů je libovolný a maximálně je roven počtu filtrů. Nejvíce informací o zvuku je obsaže v prvních několika koeficientech. Většinou se jich používá 10-13. Také se často dopočítává ještě první koeficient, který je roven logaritmu krátkodobé energie přímo z řečového signálu.

2.4.2 Dynamické příznaky

V tomto případě to jsou příznaky, které vzniknou derivací kepstrálních příznaků v čase [12]. Kepstrální příznaky nazveme statické, tzn. dynamické příznaky vyjadřují změnu statických příznaků v čase. Často se používají statické příznaky a k nim se přidají dynamické příznaky prvního (delta příznaky) a druhého (akcelerační příznaky) řádu. Pokud jsou v příznacích použity koeficienty derivace řádu k, jsou také použity koeficienty derivace řádu $k-1$. Delta koeficienty vypočítáme podle vztahu:

$$\Delta c[m] = \frac{\sum_{i=1}^{k} i(c[m+i] - c[m-i])}{2\sum_{i=1}^{k} i^2}, \tag{2.11}$$

kde $2k+1$ je velikost regresního okna a $c[m]$ je m-tý MFCC koeficient. Typický systém pro rozpoznávání mluvčích obsahuje 13 statických koeficientů (v našem případě MFCC koeficientů), 13 delta koeficientů a 13 akceleračních koeficientů.

3 HLAVNÍ ALGORITMY

3.1 Detekce řečové aktivity

Detekce řečové aktivity (VAD) označuje řečové a neřečové segmenty v audio signálu [3]. VAD má velký vliv na kvalitu diarizace mluvčích a to z více důvodů.

První je přímo odvozen ze základu systému a to je chybovost systému (DER), která bere v úvahu počet falešně negativních a falešně pozitivních poplachů. Špatný výkon tedy vede ke zvýšení DER.

Druhý vyplývá ze skutečnosti, že neřečové segmenty mohou narušit proces diarizace mluvčích. Stěžuje to hlavně segmentaci signálu. Počáteční přístup k diarizaci mluvčích se snažili vyřešit VAD v reálném čase, tím že vytvořily neřečový cluster jako vedlejší produkt diarizace. Nicméně se ukázalo, že lepších výsledků se dosáhne vyhrazením řečového a neřečového detektoru ještě před zpracováním samotného signálu. Neřečové segmenty mohou obsahovat mlčení, ale také šustění papíru nebo jiný okolní hluk. V závislosti na těchto jevech se také mění hladina energie v neřečových signálech. Kromě toho musíme počítat s nastavením mikrofonu a různosti použitých místností, kde se liší SNR (odstup signál od šumu).

Proto se VAD zaměřuje na více příznaků (energie, průchody nulou [4] a viz. kapitola 2.1, rozdíl spektra řeči a hluku v pozadí a stupeň odhadu).

3.2 Klasifikace

Klasifikační metody slouží ke třídění dat do konečného počtu tříd pomocí funkcí, z předem vymezených funkčních systémů [15]. Tyto metody se mohlou rozdělit do více sklupin. Zde jsou uvedeny jen některé z nich:

- Klasifikační stromy
- k nejbližších sousedů (k-nn)
- Naive Bayess
- Support vector machines

Pro obsáhlost těchto metod se zaměříme pouze na jednu, u které pujde dobře pochopit princip klasifikace příznaků.

3.2.1 k-nn

U rozpoznávání příznaků je metoda k-nn založena na blízkém trénovacích vzorcích v příznakovém prostoru [15]. Tato metoda je nejjednodušší ze všech algoritmů strojového učení.

Princip metody spočívá v tom, že na začátku máme trénovací množinu Z o n prvcích, o kterých víme, do které třídy v patří. Třída je rovněž vektor o velikosti n. Pokud se v prostoru objeví další prvek nebo prvky, jejich přiřazení se provede podle nejmenší vzdálenosti od trénovací množiny. Pro zvolený bod X se vypočítá vzdálenost od všech prvků trénovací množiny a poté se vybere k nejbližších sousedů tohoto bodu. Bod X se přiřadí do té třídy, která má větší zastoupení v k nejbližších sousedů.

Citlivost tohoto algoritmu se reguluje nastavením k, tzn. počtu nejbližších sousedů. Obdobná je metoda 1-nn, kde se hledá pouze jeden nejbližší soused.

Pro měření vzdáleností se nejčastěji používá euklidovská metrika:

$$\mathrm{d(X,Z)} = \sqrt{\sum_{i=1}^{n}(X_i - Z_i)^2}, \tag{3.1}$$

kde X= $(X_1, X_2, ..., X_n)$ a Z= $(Z_1, Z_2, ..., Z_n)$ jsou dva body v $n-$rozměrném prostoru, mezi kterými vzdálenost počítáme.

3.3 Selekce příznaků

Cílem selekce příznaků je zredukování počtu příznaků. Výpočetní náročnost je úměrná počtu příznaků, které chceme redukovat. Tento krok není pro správnost systému důležitý, ale je vhodný, pokud máme vyšší počet příznaků.

Selekce nám odstraní nepotřebné příznaky, které nenesou informaci důležitou pro určení mluvčích a sníží tak dimenzi prostoru [13].

Rozlišujeme dva způsoby dimenze:

- extrakce (feature extraction - FE)
- selekce (feature selection - FS)

Při extrakci dochází ke generování nových příznaků, kdy každý nový příznak je kombinací původních příznaků. Většina extrakčních algoritmů provádí lineární transformaci založenou na vlastních číslech a vektorech. Mezi tyto algoritmy patří např. PCA - analýza hlavích komponent. Selekce vybírá nevhodnější příznaky z dané množiny. Nevypočítává žádné další příznaky navíc a tím je tento algoritmus snadnější. Používá se, pokud je potřeba zachovat význam příznaků [14].

Při selekci dochází k odstranění nerelevantních a redundantích příznaků.

Relevantní příznak je takový, když existuje dvojice instancí rozdílných tříd, které se liší pouze v tomto příznaku a které mohou být rozděleny do různých tříd pouze na základě tohoto příznaku.

Redundantní příznak je takový, který obsahuje nedbytečné informace. Pokud budeme uvažovat opět dvě instance různých tříd a dva příznaky, nenalezneme takovou dvojici příznaků, která by se v jednom příznaku lišila a v druhém shodovala.

Rozdíl algoritmů spočívá ve výběru jednoho ze dvou typů. Wrapper (obalovací) a Filter.

Algoritmy typu Filter vybírají příznaky na základě dat pomocí ověřovací funkce, např. výpočtu vzdálenosti mezi třídami. Model Wrapper je založen na prohledávání s klasifikátorem. Algoritmy tohoto typu v sobě mohou zahrnovat téměř jakýkoli známý klasifikátor.

3.3.1 Sekvenční dopředné hledání

Sekvenční dopředné hledání (SFS) je jeden z nejjednodušších. Používá se jako Wrapper nebo Filter.

Na ukázku ho použijeme jako Wrapper v kombinaci s klasifikátorem 1-NN (jeden nejbližší soused) [14]. SFS má na začátku prázdnou množinu ke které postupně přidává vhodné příznaky.

Mějme množinu příznaků X a soubor příznaků $X_m(x_1 - x_m)$ - podmnožinu X. Budiž x^+ takový příznak, pro který platí:

$$x^+ = argmax\, H^+(X_m, f), \tag{3.2}$$

kde $H^+(X_m, f)$ je oceňovací funkce, kterou použijeme k ohodnocení množiny X_m s přidaným příznakem f. Nazvěme přidáním ($ADD(X_m)$) operaci, při které přidáme x^+ k současnému souboru, tedy:

$$\mathrm{ADD}(X_m) \equiv X_m \cup x^+. \tag{3.3}$$

SFS vybere d nejlepších příznaků tak, že:

$$X_d = \mathrm{ADD}^d(\phi), \tag{3.4}$$

Kde ADD^d je iterace operace ADD. Číslo d si můžeme zvolit předem, nebo si ho necháme optimalizovat algoritmem. Jako funkci $H^+(X_m, f)$ považujeme vybrání takového příznaku, který zajistí největší redukci chyby u daného klasifikátoru.

3.4 Clustering

Shluková analýza (Cluster analysis) nebo také Clustering je postup formulovaný jako procedura, pomocí níž objektivně seskupujeme jedince do skupin na základě jejich podobnosti a odlišnosti [6]. Tato analýza se hodí zejména tam, kde objekty přirozeně

vykazují snahu se seskupovat. V našem případě budeme používat shlukovou analýzu na seskupování segmentů stejného mluvčího.

Clustering se zaměřuje na identifikaci a seskupování segmentů pro daného mluvčího, které mohou být lokalizovány kdekoli ve zvukovém toku [1]. V ideálním případě bude mít jeden mluvčí jeden cluster. Problém clusteringu je metrická vzdálenost.

Protože se jen zřídka jedná o alternativní přístupy s kombinací clusteringu a opakovaného nasegmentování, přispívá to k zavedení nedetekovaných změn mluvčích. Většina současných systémů tak provádí segmentaci a clustering současně a nebo rámcový clustering na základě clusteru.

Obecný přístup se týká Viterbiho přeskupení, kde je zvukový tok rozsegmentován v závislosti na aktuálním předpokladu clusteru předtím, než jsou modely předělány v nové segmentaci. Několik iterací je obvykle provedeno tak, aby bylo Viterbiho dekódování více stabilní. Za použití vyrovnávací paměti k vyhlazení průběhu se v clusteru nebo sekvenci mluvčího odstraní chybné detekce.

Alternativní přístup k clusteringu se týká většiny případů, kdy krátké rámce oken jsou zcela přiděleny nejbližšímu clusteru, tedy tomu, který přitahuje nejvíce snímků při dekódování. Tato technika vede k úsporám ve výpočtech, ale vhodnější je pro on-line nebo live aplikace diarizace mluvčích.

Shlukovací metody můžeme rozdělit podle cílů, k nimž směřují na hierarchické a nehierarchické [7].

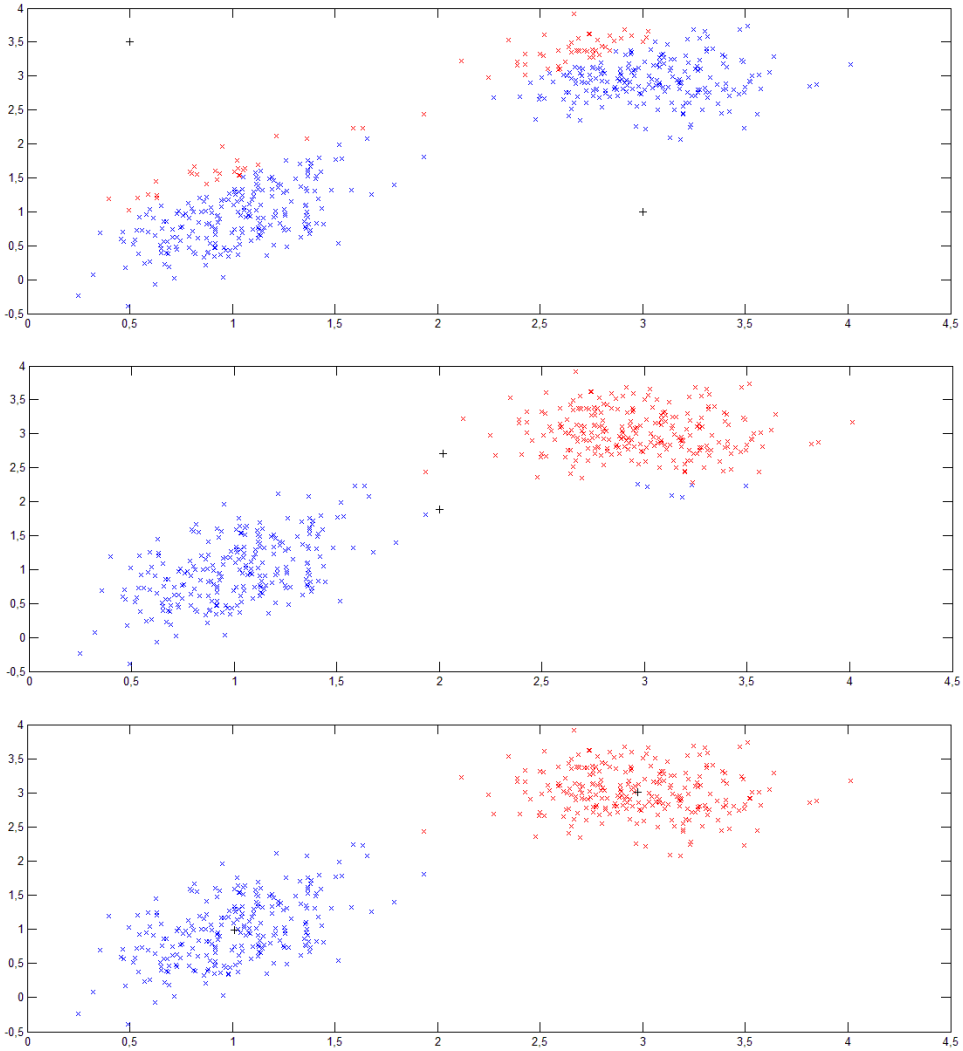
- Hierarchické shlukování je systém navzájem různých neprázdných podmnožin množiny, ve kterém průnikem každých dvou podmnožin je buď jedna z nich nebo prázdná množina a v němž existuje alespoň jedna dvojice podmnožin, jejichž průnikem je jedna z nich.
- Nehierarchické shlukování je systém navzájem různých neprázdných podmnožin množiny, v němž průnikem každých dvou podmnožin není žádná z nich.

My se opět zaměříme pouze na jednu metodu a to k-means, která patří do nehierarchického shlukování.

3.4.1 k-means

Jedná se o algoritmus s centroidním shlukováním [15]. Začneme tím, že uvažujeme problém identifikace skupin nebo shluků dat ve vícerozměrném prostoru. Předpokládejme, že máme množinu dat $x_1, ..., x_N$ obsahující N pozorování náhodné D-rozměrné Euklidovské proměnné x. Naším cílem je rozdělit množinu dat do K clusterů, kdy předpokládáme, že množina K je dána. Intuitivně můžeme smýšlet o clusteru jako skupině dat, jejichž vnitřní vzdálenosti jsou malé v porovnání se vzdálenostmi vnějších bodů mimo cluster. Můžeme to zapsat jako D rozměrný vektor μ_k, kde $k = 1, ..., K$, ve kterém μ_k je k-tý cluster. V podstatě lze říci, že μ_k obsahuje

středy shluků. Naším cílem je tedy přiřadit prvky do shluků, stejně jako soubor vektorů μ_k tak, aby součet čtverců vzdáleností každého prvku k jeho vektoru μ_k byl co nejmenší.



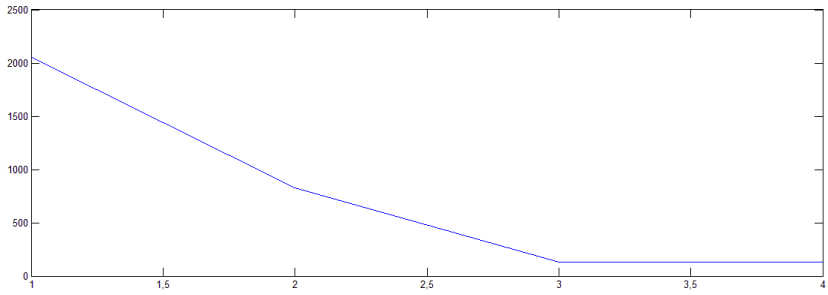
Obr. 3.1: Výpočet k-means

U každého prvku x_n jsme zavedli odpovídající sadu proměnných r_{nk}, kde $k = 1, ..., K$ popisuje, ke kterému z K clusterů prvek x_n náleží. Tedy pokud prvek x_n náleží clusteru K, potom $r_{nk} = 1$ a $r_{nj} = 0$ pro $j = k$. Toto je známo jako 1-of-K kódování. Pak můžeme definovat objektivní funkci, danou:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \mathrm{x}_n - \mu_k \right\|^2. \tag{3.5}$$

Tato funkce udává součet čtverců vzdáleností každého prvku k jeho přiřazenému vektoru μ_k. Naším cílem je najít takové hodnoty pro rnk a μ_k, aby funkce J byla

co nejmenší. Průběh výpočtu k-means můžeme vidět na obrázku 3.1 a objektivní funkci 3.2.



Obr. 3.2: Objektivní funkce J

3.5 Detekce změny mluvčího

Detekce změny mluvčího odpovídá umístění bodů ve zvukovém toku, kde došlo ke změně z jednoho mluvčího na druhého nebo z řeči na neřečový signál [2]. Je těžké označovat tyto změny a to i ručně. Vývojáři proto používají hrubší rozdělení času označených oken z 0,3 s na 2,4 s.

Délka označeného okna určuje nejkratší dobu homogenního segmentu mluvčího a proto ovlivňuje neřečovou detekci. Dokonce i mezi fonémy ve slově existují nepatrné mezery. Je zřejmé, že je lidé nevnímají jako ticho a diarizace mluvčích by to tak také neměla vnímat. Zde si představíme metody Energy-based, Model-based a Measure-based.

Nejjednodušší řešení detekce změny mluvčího je založeno na analýze akustické energie zvukového toku. Energy-based předpokládá, že všechny změny nastanou na tichých segmentech. Takže detekce tichých segmentů v toku odpovídá bodům případných změn mluvčích.

Systémy energy-based používají akustické funkce, které představují energii přes posuvné okna a detekují změny v oknech, kde tyto funkce nabývají lokálního minima. Pozice minim odpovídá tichu a potencionálním pozicím změny mluvčího. Posuvné okno vrací spoustu bodů potencionálních změn mluvčích a prahová hodnota udává, které z nich si má ponechat.

Model-based přístupy mají dobré výsledky, [2] a [16]. Spoléhají se na dvoutřídní

detektor, který obsahuje externí řečové a neřečové data. Řečové a neřečové modely se mohou libovolně přizpůsobovat specifickým splněním podmínek.

Hlavní nevýhodou model-based přístupů je jejich závislost na externích datech řečových a neřečových modelů, které je činí méně robustní pro změnu akustických podmínek.

Hybridní systémy byly navrženy jako řešení tohoto problému. Ve většině případů je nejprve aplikován přístup energy-based pro označení limitu řečových a neřečových dat, kde se klade velký důraz na klasifikaci. Ve druhém kroku používáme označené údaje ke zjišťování specifických projevů řečových a neřečových modelů, které jsou následně použity v přístupu model-based k získání výsledných řečových a neřečových segmentů. Nakonec spojíme model-based se 4 Hz modulací energy-based detektoru.

Measure-based je nejčastěji používaný algoritmus na detekci změny mluvčího [2].Tyto přístupy měří rozdíl mezi dvěma po sobě jdoucími segmenty zvukového toku, který je obvykle označován jako vzdálenost mezi oběma segmenty. Je-li tato vzdálenost větší než limit, změna mluvčího je detekována do dvou segmentů.

Vzdálenost mezi dvěma zvukovými segmenty lze měřit dvěma způsoby. Pomocí vyjádření uzavřeného tvaru a nebo porovnávání na základě pravděpodobnosti měření.

3.6 Segmentace mluvčích

V literatuře se pojem segmentace mluvčích používá jako segmentace a clustering dohromady [1]. Zatímco některé systémy řeší každý pojem zvlášť, mnoho současných nejmodernějších systémů je řeší současně. V tomto případě je segmentace a clustering na sobě nezávislý. Nicméně oba moduly jsou zásadní pro diarizaci mluvčích.

Segmentace mluvčích je jádrem diarizačního procesu a zaměřuje se na rozdělení audio proudu do jednotlivých homogenních segmentů, nebo alternativně pro detekci změny mluvčích.

Klasický přístup k segmentaci provádí testování hypotéz ve dvou posuvných a někdy překrývajících se, po sobě jdoucích oknech. U každého místa změny jsou dvě možnosti: První, že oba segmenty pochází od stejného mluvčího a že mohou být klidně zastoupeny pouze jedním modelem. A druhá, že existují dva různí mluvčí a tím je vhodnější použít dva modely.

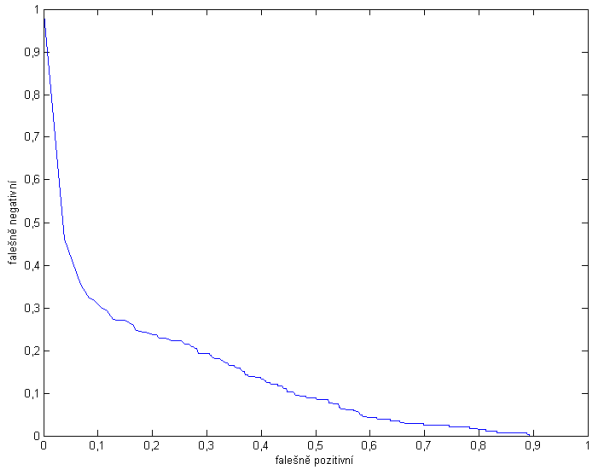
V praxi se modely odhadují v každém řečovém okně a některá kritéria se používají k určení, zda je lepší použít dva modely (dva oddělené mluvčí) a nebo jeden model (pouze jeden mluvčí). To se provádí v celém zvukovém toku a posloupnost mluvčích se extrahuje.

4 ROC KŘIVKA

V teorii signálů je ROC křivka grafické znázornění citlivosti, neboli falešně pozitivních (FP) a falešně negativních (FN) výsledků pro binárně klasifikovaný systém pro celou hodnotu prahu [5]. ROC křivka může být také udávána jako poměr pravdivě pozitivních a pravdivě negativních výsledků. To záleží na tom, jaké hodnoty se používají pro výpočet.

ROC křivka nám vykreslí hodnoty, po kterých se systém může pohybovat. Pro každý bod křivky je vypočítán poměr, např. falešně pozitivních a falešně negativních při nastaveném prahu, který je pro každý bod křivky jiný.

V našem případě používáme ROC křivku pro zobrazení účinnosti systému při použití příznaku energie signálu a počtu průchodů nulou. Na obrázku 4.1 vidíme možný příklad ROC křivky. Jak je vidět, body nabývají hodnot od 0 do 1, což značí procentuální vyjádření chybovosti.



Obr. 4.1: ROC křivka příznaku energie ze signálu 02.wav

5 POUŽITÝ SOFTWARE

5.1 Matlab a Octave

Systém pro diarizaci mluvčích je navržen v prostředí Matlab (matrix laboratory). Jedná se o interaktivní programové prostředí a skriptovací programovací jazyk čtvrté generace [8]. Programovací jazyk Matlab je určen pro vědeckotechnické účely, simulace, paralelní výpočty apod. Typické oblasti použití jsou:

- inženýrské výpočty
- tvorba algoritmů
- modelování a simulace
- analýza dat
- vědecká a inženýrská grafika
- tvorba aplikací (včetně grafického rozhraní)

Zatímco Matlab je komerční, Octave je svobodný software, navržený spíše pro práci v OS Linux. Rovněž slouží pro provádění číselných výpočtů a má nástroje pro lineární algebru, nelineární rovnice, integrování funkcí, práce s polynomy a diferenciální rovnice. Do určité míry je Octave kompatibilní s Matlabem.

Část programu byla odladěna i v prostředí Octave, pro porovnání výsledků, jestli se shodují s výsledky z Matlabu.

Více o programech na stránkách:

http://www.mathworks.com/ a http://www.octave.cz/

5.2 Praat

Praat je freeware program, který slouží pro analyzování řečových nahrávek. Jeho vlastností jsou také výstupní grafické zpracování průběhu signálu. Jeho nástroje jsou např.:

- spektrální analýza řeči
- analýza intenzity hlasu
- poslechové experimenty
- popisování a segmentace

V této práci jsme program Praat používali hlavně pro popisování a segmentaci signálu, kdy jsme si ručně označovaly segmenty jednotlivých mluvčích, znělé a neznělé úseky a nebo také speech a non-speech úseky signálu. Označené místa pak lze vyexportovat do txt souboru a následně importovat do Matlabu pro další zpracování.

Více o programu na stránkách:

http://www.fon.hum.uva.nl/praat/

5.3 Doxygen

Je volně dostupný program, který generuje dokumentaci ze zdrojového kódu, pro soubory s příponami c, cpp (C++), java, php a pro nás nejdůležitější m-file.

V dokumentaci jsou popsány jednotlivé funkce a proměnné, které se v programu používají. V podstatě nám Doxygen vytvoří jakýsi manuál pro zdrojový kód, aby se v něm šlo lépe orientovat a byl srozumitelný i pro ostatní uživatele. Výstupem je soubor formátu např. html, tex, pdf.

Více o programu na stránkách:

http://www.doxygen.org/

5.4 Git

Tento program má široké využití. Pro náš účel slouží jako archiv postupu práce. Git vytváří repozitář, do kterého si uživatel může ukládat naprogramované funkce, dokumenty či jiné soubory, které chce uchovat. Git má výhodu v tom, že pokud se programátor splete, udělá chybu v kódu a nebo něco omylem smaže, může se vrátit k poslední revizi, kde mu všechno fungovalo správně. Git se také používá na veřejných serverech, kde např. tým vývojářů má přístup ke stejnému kódu a mohou provádět změny nebo doplnění.

Program byl původně vyvinut pro OS Linux, ale později se začal používat i v ostatních OS. Jeho obsluha probíhá v příkazovém řádku, ale má také grafické rozhraní.

Více o programu na stránkách:

http://git-scm.com/

6 VÝPOČTY

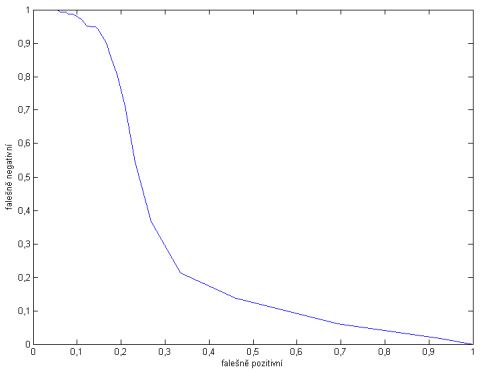
Pro analyování signálu je neprve nutné nějaký signál vytvořit a nebo použít již vytvořený. Aby jsme mohli výsledky porovnávat, použijeme databázi nahrávek rozhovorů. Dílčí výpočty budeme testovat pouze na jednom souboru, a to 01.wav. Jedná se o monofonní nahrávku se vzorkovacím kmitočtem 48 kHz o délce 3 min. 29 s., kde hovoří muž a žena. V nahrávce se nevyskytují žádné okolní rušivé zvuky, překrývání mluvčích ani změna dynamiky projevu, což je pro testování ideální.

6.1 Výpočet detekce řečové aktivity

Pro správné určení mluvčích je také třeba eliminovat non-speech části zvukového souboru. To lze provést několika způsoby. v našem případě použijeme způsob výpočtu hladiny energie v segmentu a počet průchodů nulou v segmentu.

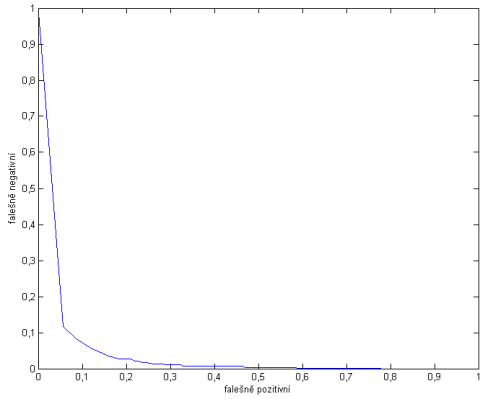
Za řečový segment lze považovat ten, který má velkou energii signálu a málo průchodů nulou. Nicméně tyto příznaky nelze aplikovat na všechny zvukové nahrávky a to z důvodu odlišnosti a kvality záznamů. Pokud by byl v signálu velký šum, energie by byla také velká, ale to by nám nezaručilo, že je segment řečový. A také, pokud je v segmentu hodně průchodů nulou, může to pouze znamenat, že obsahuje více neznělých hlásek. Aby tyto příznaky fungovaly správně, musíme použít také odpovídající zvukové nahrávky.

Pro dosažení co nejlepších výsledků musí být správně nastaven EER (Equal Eror Rate), což je práh, při kterém chybovost nebo správnost systému má stejnou nebo co nejbližší hodnotu falešně negativních a falešně pozitivních [9].

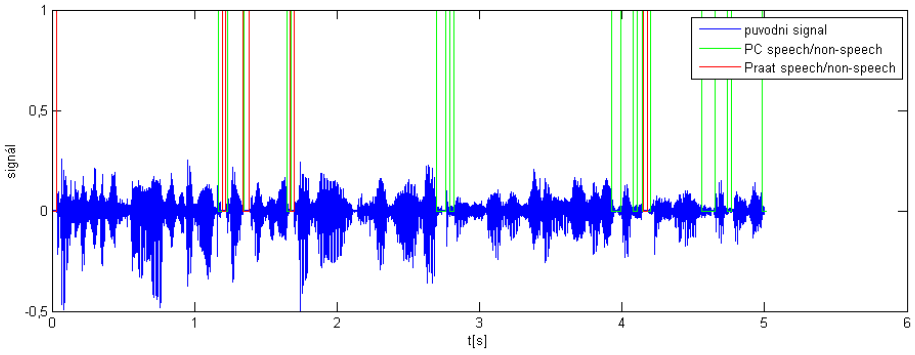


Obr. 6.1: ROC křivka příznaku počtu průchodů nulou z 01.wav

Abychom mohli změřené výsledky nějak porovnat, je potřeba si ručně označit speech a non-speech segmenty ručně. K tomu použijeme program Praat, který slouží



Obr. 6.2: ROC křivka příznaku energie z 01.wav

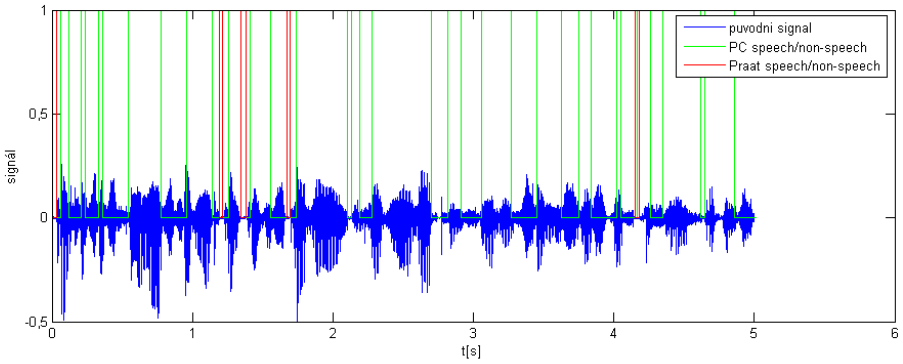


Obr. 6.3: VAD příznaku energie z části signálu 01.wav

k úpravě a zpracování zvukových souborů, zvláště řečových. V něm si označíme speech a non-speech části. Ruční označování se dá považovat za spolehlivé a tudíž bude rozhodně přesnější, než digitální zpracování na počítači. Takže výsledky z počítače budeme porovnávat ze soubory z Praatu. Pro zjištění přesnosti nám bude stačit, když ze zvukových souborů označíme segmenty pouze na 1 minutě.

U příznaku energie se nám podařilo dosáhnout velmi dobré účinnosti. Při nastavení EER energie na 0,16 u nahrávky 01.wav bylo falešně pozitivních 8,33 % a falešně negativních 8,25 % (obrázek 6.2).

S příznakem počtu průchodů nulou jsme nastavili u nahrávky 01.wav EER průchodů nulou na 70, při velikosti segmentu 30 ms (1440 vzorků). Výsledky už nebyly tak dobré jako u energie. Přesto jsme dosáhli chybovosti falešně pozitivních 30,22 % a falešně negativních 28,08 % (obrázek 6.1).



Obr. 6.4: VAD příznaku průchody nulou z části signálu 01.wav

Rozdílnost těchto dvou příznaků lze vidět i na obrázku (6.3) a (6.4).

6.2 Výpočet základního tónu hlasu

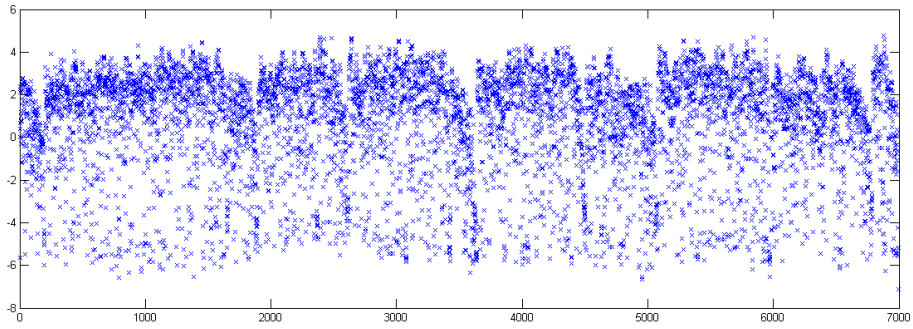
Pro výpočet základního tónu hlasu použijeme metodu centrálního klipování. Jedná se o univerzální metodu a k našemu případu se bude hodit nejlépe. Pro přesnost budeme dodržovat postup, jak je uvedeno výše v kapitole Výpočet základního tónu hlasu. Nejprve provedeme segmentaci řečového signálu. Pro jednotlivé rámce vypočítáme práh P a to z toho důvodu, že hodnota signálu kolísá a nelze určit pevnou hodnotu prahu pro celý signál (obrázek 2.3.a). Hodnoty prahu vypočítáme ze sousedních segmentů (obrázek 2.3.b). Po prahování signál znormalizujeme (obrázek 2.3.c) a vypočítáme autokorelační funkci (obrázek 2.3.d).

Abychom docílili dobré účinnosti a použitelné hlukové analýzy, musíme stanovit EER pro určení znělých segmentů. Při používání dvou různých příznaků můžeme také ovlivnit jejich váhu, tzn. jeden bude upřednostněn a nastavení jeho EER bude mít větší vliv na přesnost. Zatím ponecháme váhu obou příznaků stejnou a zkusíme nastavit EER co nejúčineji.

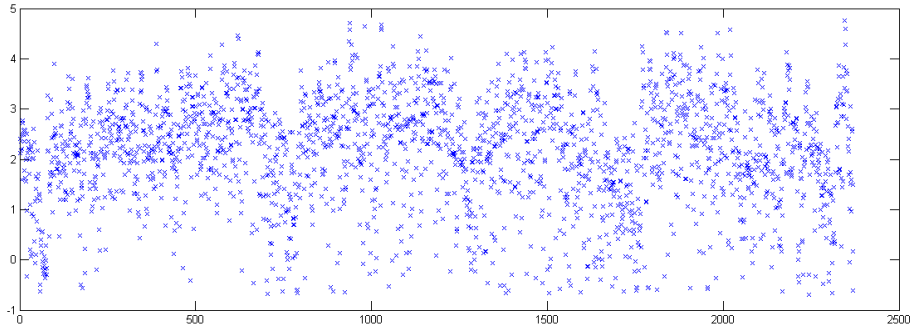
6.3 Výpočet MFCC koeficientů

Při dodržování výše uvedeného postupu jsme provedli výpočet MFCC koeficientů. V kódu jsme narazili pouze na menší problém a to sice, že musíme ošetřit logaritmování nulových hodnot. Ne všechny MFCC koeficienty jsou dobře odlišitelné a tak dále budeme používat jenom některé.

Také je pro clustering výhodné použít MFCC koeficienty jenom pro znělé úseky. Odlišitelnost je pak o něco lepší. Můžeme porovnat na obrázku 6.5 a 6.6.



Obr. 6.5: 1. MFCC koeficient ze souboru 01.wav



Obr. 6.6: 1. MFCC koeficient ze souboru 01.wav pouze u znělých úseků

6.4 Výpočet delta koeficientů

Výpočet delta koeficientů nedosáhnul očekávaných výsledků a proto ho v clusteringu nepoužíváme. Výpočet byl prováděn podle vzorce z kapitoly Dynamické koeficienty.

6.5 Výpočet k-nn

Klasifikační metodu k-nn jsme nakonec z časových důvodů nevyužili, protože neprovádíme klasifikaci příznaků. Vyzkoušeli jsme si její funkčnost a případné použití pro klasifikaci na testovací množině příznaků. Její výpočet je rovněž uveden ve zdrojovém kódu.

6.6 Výpočet k-means

Metoda k-means nám slouží pro clustering. Vzhledem k tomu, že nepoužíváme selekci příznaků, opět z časových důvodů, vybrali jsme si nejvhodnější příznaky ručně.

Aby byl výpočet o něco přesnější, vybrali jsme pouze znělé úseky (pomocí výpočtu základního tónu, kde tuto funkci používáme). Pro výpočet jsme použili velikost energie, základní tón hlasu, pokud mluvčí budou dobře odlišitelní, jinak vynecháme a MFCC koeficienty, ze kterých jsme vybrali 1., 2. a 14. To nám vytvořilo shluky v 5ti rozměrném prostoru.

U výpočtu je vždy nutné zadat souřadnice středových bodů. Např. při 2 mluvčích budeme mít 2 středové body a za použití 5 příznaků bude mít každý středový bod 5 hodnot, pro výpočet v prostoru. Tyto body mohou být libovolné, výsledek se tím neovlivní. Akorát se změní počet iterací.

6.7 Segmentace mluvčích

Pro segmentaci mluvčích jsme použili výsledky metody k-means, která ji v podstatě udělala za nás. Vzhledem k určité nepřesnosti bylo potřeba rozdělení do tříd upravit mediánovým filtrem. Aby byla segmentace provedena po celé délce signálu, museli jsme ještě dopočítat segmenty, které jsme pro jejich neznělost vyřadili. Jednalo se pouze o doplnění třídy, kdy neznámá třída na pozici k se přiřadila ke třídě $k - 1$.

6.8 Výsledky

V tabulce 6.1 uvádíme přehled výsledků systému pro diarizaci mluvčích naměřených na databázi nahrávek.

Z výsledků testů vyplývá, že tento systém není úplně ideální a je náchylný k chybným vyhodnocením. To je způsobeno několika faktory.

Jako první a zároveň největší faktor chyby je zvolená metoda k-means. Tato metoda vytváří v n-rozměrném prostoru shluk o počtu n-příznaků. Tato metoda má nehierarchické shlukování, tzn. společné rysy z celého signálu, budou v jednom shluku. Pokud ovšem příznaky jednotlivých mluvčích nejsou dobře odlišitelné, tvoří jeden velký shluk. Je zde také problém, pokud daný mluvčí vystupuje v nahrávce jen krátkou dobu v poměru s ostatními mluvčími. Takovýto shluk, i když dobře odlišitelný, muže dost snadno pohltit sousední shluk, který bude tvořen z mnohonásobně vyššího počtu segmentů.

Druhý významný faktor je volba příznaků. Ne všechny se pro clustering hodí, ovšem pokud nepoužíváme klasifikační metodu pro vhodný výběr příznaků, musíme

Tab. 6.1: Přehled naměřených výsledků systému pro diarizaci mluvčích

soubor	mluvčí		shoda mluvčích	změny mluvčích		poznámky
	muž	žena		FP	FN	
wav01	1	1	97,55%	0%	0%	-
wav02	1	1	96,7%	9,09%	0%	-
wav03	3	1	68,9%	20%	40%	shluk A,B a C
wav04	2	0	69,63%	75%	75%	-
wav05	2	0	63,6%	66%	50%	-
wav06	3	0	71,65%	57,14%	0%	shluk B a C
wav07	3	0	56,83%	75%	25%	shluk B a C
wav08	3	0	81,22%	0%	50%	shluk A a B
wav09	3	0	95,5%	0%	50%	shluk B a C
wav10	3	0	80,1%	0%	25%	shluk A a C

tyto příznaky vybrat ručně na základě jejich vlastností a všeobecných znalostí, jak by se takový příznak měl chovat.

Třetí faktor spočívá ve kvalitě nahrávky. Soubor wav01 má vzorkovací kmitočet 48 kHz. Soubor wav02 44,1 kHz, ovšem další soubory pouze 16 kHz. To nám významně zmenšuje množství informací o daném signálu. Další nevýhodou souborů wav03 - wav10 je obsah cizích zvuků, překřikování mluvčích, delší pauzy v řeči, zasekávání se mluvčích apod. Navíc se v těchto nahrávkách dost často objevují mluvčí stejného pohlaví, někdy s hodně podobnou frekvencí základního tónu hlasu, který je pak dobré v takových případech jako příznak vynechat. A jak už vychází z volby clusteringu, pokud je mluvčí v nahrávce slyšet jen krátce, pohltí ho ostatní shluky mluvčích.

Při testech bylo použito maximálně 5 příznaků a to velikost energie, základní tón hlasu a z MFCC koeficientů byly vybrány příznaky 1,2 a 14. Je pravděpodobné, že při kombinaci jiných příznaků, např. přidání nebo záměna některých z MFCC koeficientů a delta koeficientů, bychom dosáhli lepších výsledků.

7 ZÁVĚR

Cílem této práce bylo vytvořit funkční systém pro diarizaci mluvčích a otestovat ho na databázi nahrávek. Při zvoleném postupu jsme nedosáhli ideálních výsledků, ale i přesto lze systém považovat za funkční, pokud zvolíme vhodný typ nahrávek.

Z velké části může výsledky ovlivnit uživatel, jelikož je potřeba předem nastavit spoustu parametrů, které ovlivňují výpočty. Systém klade vyšší nároky pro uživatele, jelikož musí vědět, jak parametry nastavit, aby byly výsledky co nejlepší. Měření jsme opakovali vícekrát, abychom dosáhli co nejlepších výsledků, ovšem nutno dodat, že takhle by to fungovat nemělo.

Po konečném otestování jsme došli k závěru, že pro vyšší kvalitu systému tento postup není dostačující. Lepší výsledky mohou vést přes pravděpodobnostní a statistické metody, jako Bayesovo informační kritérium (BIC), Hidden Markov Model (HMM) nebo Gaussian Mixture Model (GMM).

Jelikož je toto téma velmi rozmanité a způsobů řešení mnoho, nebylo účelem projít všechny, ale pokusit se o jedno možné řešení. Z časových důvodů je zde uveden pouze jeden způsob navržení systému pro diarizaci mluvčích. Pro další pokračování je potřeba mít větší všeobecné znalosti této problematiky, která je v podstatě stále nová a vyvíjí se dopředu každou chvíli.

LITERATURA

[1] ANGUERA, X. et al. *Speaker diarization : A review of recent research.* Accepted for publication in "IEEE Transactions On Acoustics Speech and Language Processing" (TASLP), special issue on "New Frontiers in Rich Transcription", 2011.

[2] NOULAS, A. *Audiovisual fusion for speaker diarization.* Faculteit der Natuurwetenschappen, Wiskunde en Informatica : FNWI: Informatics Institute (II), 2010. 167 s. ISBN 90-75691-06-8.

[3] WEI, J. at al. A new algorithm for voice activity detection. In *Circuits and Systems, 2003. ISCAS '03. Proceedings of the 2003 International Symposium on,* vol.2, pp. II-588- II-591, 25-28 May, 2003 Bangkok, Thailand. ISBN 0-7803-7761-3.

[4] ATASSI, H. *Metody detekce základního tónu řeči.* Ústav telekomunikací, FEKT VUT, Brno, 2008. ISSN 1213-1539

[5] Zweig, MH, Campbell, G. *Receiver operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine.* Clin Chem 1993. vol. 39 no. 4 561-577.

[6] HANUŠ, J. *Shluková analýza a její aplikace.* Plzeň, 2009. 42 s. Bakalářská práce. Fakulta aplikovaných věd.

[7] Kelbel, J., Šilhán, D. *Shluková analýza.* Praha : s.n., 2002. Dostupné z WWW: http://staff.utia.cas.cz/nagy/skola/Projekty/Classification/ShlukovaAnalyza.pdf

[8] *Matlab* URL: http://www.mathworks.com/products/matlab [cit. 2011-11-28]

[9] *Equal Eror Rate* URL: http://www.answers.com/topic/equal-error-rate [cit. 2011-11-29]

[10] Psutka, J.; Müller, L.; Matoušek, J.; Radová, V.; , *Mluvíme s počítačem česky.* Academia, Prague, 2006, ISBN 80-200-1309-1.

[11] ČERNOCKÝ, J. *Zpracování řečových signálů.* 2006. Dostupné z WWW: http://www.fit.vutbr.cz/study/courses/ZRE/public/opora/zre_opora.pdf.

[12] Jinjin Ye, B.S. *Speech recognition using time domain features from phase space reconstructions.* Dostupné z WWW: http://speechlab.eece.mu.edu/papers/Ye_thesis.pdf Marquette University. Milwaukee, Wisconsin. May 2004.

[13] Saeys, Y., Inza, I. a Larrańaga, P. *A review of feature selection techniques in bioinformatics.* Bioinformatics. 2007, Sv. 23, 19.

[14] Somol, P., Novovičová, J. a Pudil, P. *Efficient Feature Subset Selection an Subset Size Optimization.* Survey. Pattern Recognition Recent Advances, 2010.

[15] Bishop, Ch. *Pattern Recognition and Machine Learning.* New York, USA, 2006. ISBN-10: 0-387-31073-8.

[16] ANGUERA, X. et al. Hybrid Speech/non-speech detector applied to Speaker Diarization of Meetings. In *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The,* pp.1-6, 28-30 June, 2006 San Juan, Puerto Rico. ISBN 1-424400471-1.

A SEZNAM SYMBOLŮ A ZKRATEK

1-nn ..... Jeden nejbližší soused
AKF ..... Autokorelační funkce
BIC ..... Bayesovo informační kritérium
DER ..... Chybovost systému
DFT ..... Diskrétní Fourierova Transformace
EER ..... Práh chybovosti
F0 ..... Základní kmitočet
FE ..... Extrakce příznaků
FP ..... Falešně pozitivní
FS ..... Selekce příznaků
GMM ..... Gaussian Mixture Model
HMM ..... Hidden Markov Models
Hz ..... Hertz
IDCT ..... Zpětná Cosinova transformace
IFFT ..... Zpětná Fourierova transformace
kHz ..... Kilo Hertz
k-nn ..... K nejbližších sousedů
Mel ..... Melody
MFCC ..... Mel frekvenční kepstrální koeficienty
ms ..... Mili sekunda
NIST ..... National Institute of Standards and Technology
PCA ..... Analýza hlavích komponent
PN ..... Falešně negativní
ROC ..... Křivka chybovosti
s ..... Sekunda
SFS ..... Sekvenční dopředné hledání
SNR ..... Odstup signál šum
VAD ..... Detekce řečové aktivity