UNIVERZITA PARDUBICE
Fakulta elektrotechniky a informatiky

Určování druhu zdroje signálu v zarušeném prostředí

Bc. David Cvrk

Diplomová práce
2014

Univerzita Pardubice
Fakulta elektrotechniky a informatiky
Akademický rok: **2012/2013**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. David Cvrk**
Osobní číslo:	**I11354**
Studijní program:	**N2612 Elektrotechnika a informatika**
Studijní obor:	**Komunikační a řídicí technologie**
Název tématu:	**Určování druhu zdroje signálu v zarušeném prostředí**
Zadávající katedra:	**Katedra elektrotechniky**

Zásady pro vypracování:

V teoretické části popište možnosti záznamu akustického signálu, který bude obsažen v databázi signálů. Pro různé zdroje signálu popište charakteristiky signálů. Jako zdroj signálu bude uvažován zvuk s velkou dynamikou, ultrazvuk nebo slovo.
V praktické části navrhněte postup záznamu různých typů signálu v zarušeném prostředí směrovým mikrofonem a všesměrovým senzorovým polem. Zaznamenejte vybrané typy signálů v různých prostředích a proveďte analýzu možnosti jejich rozpoznání. Proveďte záznam dynamického signálu pomocí mikrofonního pole a vzdálenějšího mikrofonu, záznam bude časově synchronizován.

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

BENESTY, Jacob, J CHEN a Yiteng HUANG. Microphone array signal processing. Berlin: Springer, c2008, x, 240 p. ISBN 9783540786122.
KOČÍ, Miloslav 2010. Rozpoznání slov diskrétního diktátu. Pardubice. 2010. 75s. Diplomová práce. Univerzita Pardubice.
MANDLÍK, Michal 2010. Využití mikrofonního pole pro určení směru příchodu zvuku. Pardubice. 2010. 83s. Diplomová práce. Univerzita Pardubice.

Vedoucí diplomové práce: **Ing. Zdeněk Němec, Ph.D.**
Katedra elektrotechniky

Datum zadání diplomové práce: **31. října 2013**
Termín odevzdání diplomové práce: **22. srpna 2014**

Univerzita Pardubice Česká republika

L.S.

prof. Ing. Simeon Karamazov, Dr.
děkan

Ing. Zdeněk Němec, Ph.D.
vedoucí katedry

V Pardubicích dne 15. listopadu 2013

Prohlášení autora

Prohlašuji, že jsem tuto práci vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně.

V Pardubicích dne 14. 7. 2014 Bc. David Cvrk

Poděkování

Chtěl bych poděkovat svému vedoucímu práce Ing. Zdeňku Němcovi, PhD. za cenné rady, které mi pomohly k vytvoření této diplomové práce. Také bych chtěl poděkovat rodině a přítelkyni, za podporu v průběhu vytváření práce.

Anotace

Tato diplomová práce popisuje charakteristiky různých typů akustického signálu a způsobu jeho zaznamenání. Dále popisuje, jak lze zaznamenaný akustický signál zpracovat, aby ho bylo možné použít pro rozpoznávání slov. Následně popisuje metodu borcení času, která slouží pro rozpoznávání slov. Podle těchto informací byl vytvořen program v prostředí Matlab, pomocí kterého lze rozpoznávat slova zaznamenaná od slov referenčních při působení různých typů rušení a šumu.

Klíčová slova

Slovo; mel kepstrální koeficienty, metoda borcení časové osy, rušení, odstup signál od šumu

Title

Sound source recognition in noisy enviroments

Annotation

This thesis describes the characteristics of different types of acoustic signal and the way it was recorded. It also describes the method of acoustic signal processing and DTW method for word detection and recognition with dynamic time wrapping. In practical part of thesis the program for acoustic signal processing and word recognition. The program in Matlab recognizes words recorded by test speakers with reference words, when exposed to variable type of interference and noise.

Keywords

Word; mel cepstral coeficients, time warping method, interference, signal to noise ratio

Obsah

Seznam zkratek ... **8**

Seznam obrázků ... **9**

Seznam tabulek ... **10**

Seznam grafů ... **11**

Úvod ... **12**

1 Zvuk ... **13**

1.1 Ultrazvuk ... 15

1.2 Slovo ... 17

1.3 Zvuk s velkou dynamikou ... 18

2 Záznam akustického signálu ... **20**

2.1 Mikrofon ... 21

2.2 Elektrodynamický mikrofon ... 21

2.3 Kondenzátorový mikrofon ... 21

2.4 Použité mikrofony ... 22

2.4.1 Dexon MC 500 ... 23

2.4.2 Behringer C-3 ... 24

2.5 Mikrofonní předzesilovač ... 25

2.6 A/D převodník ... 26

2.6.1 A/D NI USB-6251 ... 27

3 Zpracování zaznamenaného signálu ... **28**

3.1 Filtrace signálu ... 28

3.2 Rozdělení signálu do rámců ... 29

3.3 Odstranění ss složky ... 30

3.4 Střední krátkodobá energie a intenzita ... 30

3.5 Počet průchodů nulou ... 32

3.6 Detekce slova ... 33

3.7 Kepstrální koeficienty ... 34

4 Rozpoznávání slov ... **36**

4.1 Metoda borcení časové osy DTW 36
4.1.1 Lokální omezení DTW cesty 38
4.1.2 Výpočet vzdáleností 39
4.1.3 Postup vypočítání optimální cesty funkce DTW 39
4.1.4 Normalizační faktor 40
5 Praktické využití rozpoznávání slov 41
5.1 SpeechTech ASR 41
5.2 Dragon Dictation 42
5.3 Dragon Search 42
6 Praktická část 44
6.1 Vytvoření databáze slov 44
6.2 Přidání rušení k nahrávkám 45
6.3 Parametry DTW 49
6.4 Vyhodnocení 51
6.4.1 Slova nahraná směrovým mikrofonem s přidaným rušením 51
6.4.2 Slova nahraná všesměrovým mikrofonem s přidaným rušením 55
6.4.3 Slova s přidaným rušením a AWGN 59
Závěr 63
Literatura 65
Příloha A – Tabulka výsledků rozpoznávání slova "Vloupání" s úrovní rušení SNR=3dB pro 2.typ lokálního omezení DTW 66
Příloha B – Tabulka výsledků rozpoznávání slova "Techniky" s úrovní rušení SNR=3dB pro 1. typ lokálního omezení DTW 67

Seznam zkratek

FIR	Finite Impulse Respons
DTW	Dynamic Time Warping
UTC	Unit Tap Constraint
PC	Personal Computer
USB	Universal Serial Bus
DFT	Discrete Fourier Transform
DCT	Discrete Cosine Transform
ASR	Automatic Speech Recognition
LVCSR	Large Vocabulary Continous Speech Recognition
SNR	Signal to Noise Ratio
AWGN	Additive white Gaussian Noise

Seznam obrázků

Obrázek 1 - Záznam ultrazvuku 16
Obrázek 2 - Záznam slova bomba 18
Obrázek 3 – Záznam zvuku s velkou dynamikou 19
Obrázek 4- Nahrávací místnost 20
Obrázek 5 - Schéma záznamu akustického signálu 20
Obrázek 6 - Oddělení napájecího a signálového napětí 22
Obrázek 7 - Směrový mikrofon dexon MC 500 23
Obrázek 8- Směrová charakteristika mikrofonu 24
Obrázek 9 - Všesměrový mikrofon Behringer C-3 24
Obrázek 10 – Směrová charakteristika mikrofonu 25
Obrázek 11 - Mikrofonní předzesilovač Behringer 26
Obrázek 12 - A/D převodník NI USB- 6251 27
Obrázek 13- Blokové schéma zpracování signálu 28
Obrázek 14- Impulsní odezva FIR filtru 28
Obrázek 15 - Vykreslení 10tého rámce 29
Obrázek 16 - Střední krátkodobá intenzita 31
Obrázek 17 - Počet průchodů nulou v jednotlivých rámcích 32
Obrázek 18 - Detekce slova bomba 33
Obrázek 19 - Rozložení filtrů ve frekvenční oblasti 34
Obrázek 20 – Rozložení filtrů 35
Obrázek 21 - Matice vzdáleností DTW 37
Obrázek 22 - Ukázka aplikace proměnné k na funkci DTW 38
Obrázek 23 - Záznam celé věty 45
Obrázek 24 – Záznam rušení 46
Obrázek 25 - Nezarušené slovo "Rostoucí" 47
Obrázek 26 - Zarušené slovo “Rostoucí“ 48
Obrázek 27 - Zarušené slovo "Rostoucí" s přidaným bílým šumem 49

Seznam tabulek

Tabulka 1 - Příklad hladin intenzity zvuku pro některé známé zvuky (SVOBODA, a další, 2006) 15
Tabulka 2 - Průměrná úspěšnost detekce všech slov 52
Tabulka 3 - Úspěšnost detekce slova "Bezpečnostní" 53
Tabulka 4 - Úspěšnost detekce slova "Vloupání" 53
Tabulka 5 - Úspěšnost detekce slova "Jedná" 54
Tabulka 6 – Průměrná úspěšnost detekce všech slov 55
Tabulka 7 – Úspěšnost detekce slova "Bezpečnostní" 56
Tabulka 8 – Úspěšnost detekce slova “Vloupání“ 57
Tabulka 9- Úspěšnost detekce slova "Jedná" 58
Tabulka 10 - Průměrná úspěšnos detekce všech slov 59
Tabulka 11 - Úspěšnost detekce slova "Bezpečnostní" 60
Tabulka 12 - Úspěšnost detekce slova "Vloupání" 61
Tabulka 13 - Úspěšnost detekce slova "Jedná" 62

Seznam grafů

Graf 1 - Průměrná úspěšnost detekce všech slov 51
Graf 2 - Úspěšnost detekce slova "Bezpečnostní" 52
Graf 3 - Úspěšnost detekce slova "Vloupání" 53
Graf 4 - Úspěšnost detekce slova "Jedná" 54
Graf 5 - Průměrná úspěšnost detekce všech slov 55
Graf 6 - Úspěšnost detekce slova "Bezpečnostní" 56
Graf 7- Úspěšnost detekce slova “Vloupání“ 57
Graf 8 – Úspěšnost detekce slova “Jedná“ 58
Graf 9 - Průměrná úspěšnost detekce všech slov 59
Graf 10 - Úspěšnost detekce slova "Bezpečnostní" 60
Graf 11 - Úspěšnost detekce slova "Vloupání" 61
Graf 12 - Úspěšnost detekce slova "Jedná" 62

Úvod

Snaha naučit počítače rozumět lidské řeči je téměř stejně stará jako počítače samotné. Ale teprve v posledních několika letech došlo k výraznějšímu kroku kupředu a na trhu se objevují různé aplikace, které pracují s lidskou řečí. Snad každý z nás má v mobilním telefonu funkci hlasového vytáčení, která uživateli umožňuje pomocí hlasu vytočit požadované telefonní číslo. Dalším příkladem můžou být aplikace, které občas nazýváme tzv. „Sekretářky“. Ty přepisují mluvenou řeč do psané podoby. Některé, třeba jako mobilní aplikace Siri, která je součástí Apple iOS, dokážou dokonce na dotazy položené uživatelem i odpovídat. Budoucností je aplikace, která by u živého televizního vysílání bez chyby přepisovala komentář komentátora a zobrazovala ho ve formě titulků, což by neslyšícím velice zpříjemnilo sledování.

Jedním z největších problémů pro systémy pracující s lidskou řečí je rušení způsobené zvukem v pozadí nebo přítomností dalších mluvčích v oblasti nahrávání. Lidské ucho je adaptibilní a dokáže se v místnosti, kde mluví najednou více řečníků, zaměřit jen na jednoho z nich. To stroj bohužel nedokáže a vnímá zvuk celistvě. Zlepšit tuto nepříjemnou vlastnost je možno použitím směrového mikrofonu, který je natočen směrem k mluvčímu, kterého chceme detekovat. Tímto krokem se alespoň částečně potlačí nežádoucí zvuky způsobené ostatními mluvčími. Další možností je použití různých filtrů, které dokážou potlačit zvuk na určitých kmitočtech.

Cílem této práce je ukázat, jakým způsobem závisí úspěšnost detekce na úrovni rušení. V první kapitole jsou popsány elementární charakteristiky zvuku a několik jeho příkladů. Druhá kapitola se zabývá tím, jakým způsobem je možné pořídit nahrávku od mluvčího, aby ji mohl dále zpracovávat program v počítači. Jsou tam detailně popsány funkce jednotlivých bloků nahrávacího řetězce a také konkrétní zařízení, které jsme použili pro pořízení nahrávek v diplomové práci. Třetí kapitola se věnuje způsobu, jakým se zpracovává zaznamenaný signál. Jsou tu detailně popsány jednotlivé kroky vedoucí k vypočtení mel-kepstrálních koeficientů, které jsou vstupem pro rozpoznávání slov pomocí funkce borcení časové osy. Ve čtvrté kapitole nalezneme informace o metodě rozpoznávání slov pomocí borcení časové osy. V páté kapitole jsou uvedeny příklady existujících aplikací, které využívají rozpoznávání slov. V poslední šesté kapitole je popsáno, jakým způsobem se postupovalo v praktické části, a nalezneme tam i výsledky úspěšnosti detekce slov zatížených různými typy rušení.

1 Zvuk

Zvuk můžeme obecně definovat jako mechanické vlnění. Zvuk, který je člověk schopen vnímat, kmitá s frekvencí přibližně 16 Hz do 20 kHz a jeho intenzita se pohybuje mezi prahem slyšitelnosti, což je asi 10^{-12} W/m^2, a prahem bolesti 10^2 W/m^2. Toto slyšitelné frekvenční pásmo je ale hodně individuální a hlavně horní hranice se s rostoucím věkem snižuje. Zvuky, které leží mimo toto pásmo, sice neslyšíme, ale vnímáme je a můžou nám způsobit i zdravotní komplikace. Zvukům, které leží pro člověka nad slyšitelnou hranicí a do 50 kHz, říkáme ultrazvuk a těm, které leží pod slyšitelnou hranicí v pásmu 0,7 Hz až 16 Hz, říkáme infrazvuk (SMETANA, 1998).

Zdroj zvukového vlnění nazýváme zdroj zvuku a látkové prostředí. Ze zdroje se zvuk šíří jen pružným látkovým prostředím libovolného skupenství. Nejčastěji to je vzduch, v němž se zvuk šíří jako podélné postupné vlnění. Nejdůležitější charakteristikou prostředí z hlediska šíření zvuku je rychlost zvuku v daném prostředí.

Rychlost zvuku ve vzduchu závisí na složení vzduchu (nečistoty, vlhkost), ale nejvíce na jeho teplotě. Ve vzduchu o teplotě t v Celsiových stupních má zvuk rychlost

$$v = (331{,}81 + 0{,}61T)\ [m.s^{-1}], \tag{1.1}$$

kde v je rychlost šíření zvuku

T je teplota vzduchu

Při výpočtech budeme používat přibližnou teplotu pro běžné teploty vzduchu 340 m.s^{-1}. Rychlost zvuku není ovlivněna tlakem vzduchu a je stejná pro zvuková vlnění všech frekvencí. V kapalinách a pevných látkách je rychlost zvuku větší než ve vzduchu (popř. v jiných plynech).

Šíření zvuku je ovlivněno i překážkami, na které vlnění dopadá, a projevuje se odraz i ohyb zvukového vlnění. Zvláštním případem odrazu zvuku od rozlehlé překážky (skalní stěna, velká budova) je ozvěna. Je v podstatě důsledkem vlastnosti sluchu, kterým rozlišujeme dva po sobě následující zvuky, jestliže mezi nimi uplyne doba alespoň 0,1s. To je přibližně doba, kterou potřebujeme k vyslovení jedné slabiky, a zvuk urazí celkovou vzdálenost 34 m (tzn. 17 m od pozorovatele k přepážce a zpět).

Jestliže je překážka blíže než 17 m, zvuky již neodlišíme, částečně se překrývají a odražený zvuk splývá se zvukem původním. To se projevuje jako prodloužení trvání zvuku, které nazýváme dozvuk. S dozvukem je třeba počítat při projektování velkých místností, koncertních síních apod. Působí rušivě a snižuje srozumitelnost řeči nebo

zkresluje hudbu. Proto se akustické vlastnosti sálů zlepšují rozčleňováním ploch stěn, závěsy, použitím materiálů, které pohlcují zvuk, apod. (SVOBODA, a další, 2006).

Zvuky můžeme rozdělit na hudební (tóny) a nehudební. Nehudebním zvukem je každé nepravidelné vlnění vodiče zvuku, jehož příčinami jsou nepravidelné rozruchy. Příkladem nehudebního zvuku může být například výstřel nebo zvuk způsobený přeskočením elektrické jiskry. Hudební zvuky vznikají pravidelným, periodicky probíhajícím kmitáním. Zdrojem hudebního zvuku můžou být třeba hlasivky nebo hudební nástroje. Každý zvuk hudební nebo nehudební se vyznačuje intenzitou zvuku. Ta je definována jako zvuková energie, dopadající na jednotku plochy za jednotku času a vyjádřena rovnicí:

$$I = \frac{E}{S.t}\ [W.m^{-2}] \tag{1.2}$$

kde I je intenzita zvuku

E je zvuková energie

S je plocha

t je čas

A veličina hladina intenzity zvuku udává intenzitu zvuku v jednotkách decibel:

$$L = 10log\frac{I}{I_0}\ [dB] \tag{1.3}$$

kde L je hladina intenzity zvuku

I je intenzita zvuku

I_0 je vztažená hodnota intenzity, $I_0=10^{-12}\ Wm^{-2}$

Tabulka 1 - Příklad hladin intenzity zvuku pro některé známé zvuky (SVOBODA, a další, 2006)

Zdroj zvuku	Vzdálenost [m]	Hladina akustického výkonu [dB]
Práh slyšení		0
Tichý šepot	1	10
Šepot	2	20
Šum ve sborovně		30
Tichý rozhovor	1	40
Normální rozhovor	1	65
Motor automobilu	5	70
Symfonický orchestr	3 až 5	80
Hluk motorových vozidel	10	90
Hudba na diskotéce		100
Startující letadlo	10	110
Práh bolesti		120

Hudební zvuky se kromě intenzity zvuku vyznačují výškou a zabarvením. Výška tónu se udává absolutně nebo relativně. Absolutní výška tónu je určena frekvencí. Absolutní výška tzv. komorního A je 440 Hz. Relativní výška dvou hudebních zvuků se rovná podílu jejich frekvencí nebo absolutních výšek (SMETANA, 1998).

1.1 Ultrazvuk

Uvedli jsme už, že mechanické vlny, které se vyznačují frekvencemi nad 20 000 Hz, nazýváme ultrazvukem. V přírodě vzniká ultrazvuk např. při elektrických jiskřivých výbojích. Víme též, že netopýři vysílají ultrazvukové impulzy a odraz těchto impulzů na překážkách využívají k navigačním účelům. Technický a vědecký význam získal ultrazvuk tehdy, když se našly způsoby na jeho laboratorní výrobu. Zařízení, které se na tento účel používají, jsou ultrazvukové generátory. Jejich podstatou je využití tzv. piezoelektrického jevu, resp. jevu magnetostrikce.

Pro technické a vědecké využití ultrazvuku je důležité, že ultrazvukový generátor umožňuje vyrobit ultrazvukové vlny s podstatně vyššími energiemi, než jakými se vyznačují zvukové vlny. Přenos energie jednotkou plochy může být při ultrazvukových vlnách až 10^{10} krát větší než při běžných zvukových vlnách. Vlny s takovým vysokým obsahem přenášené energie umožňují dosahovat velmi intenzivní mechanické účinky. Využití této vlastnosti ultrazvuku je velmi mnohostranné. Uveďme alespoň několik příkladů. Pomocí ultrazvuku mohou kapaliny, které se navzájem nemíchají (například rtuť a voda), celkem promíchat a utvořit tak emulzi. U fotografických emulzí ultrazvuk umožňuje získat nepatrnou zrnitost. Ultrazvuk účinkuje i na větší molekuly a podporuje

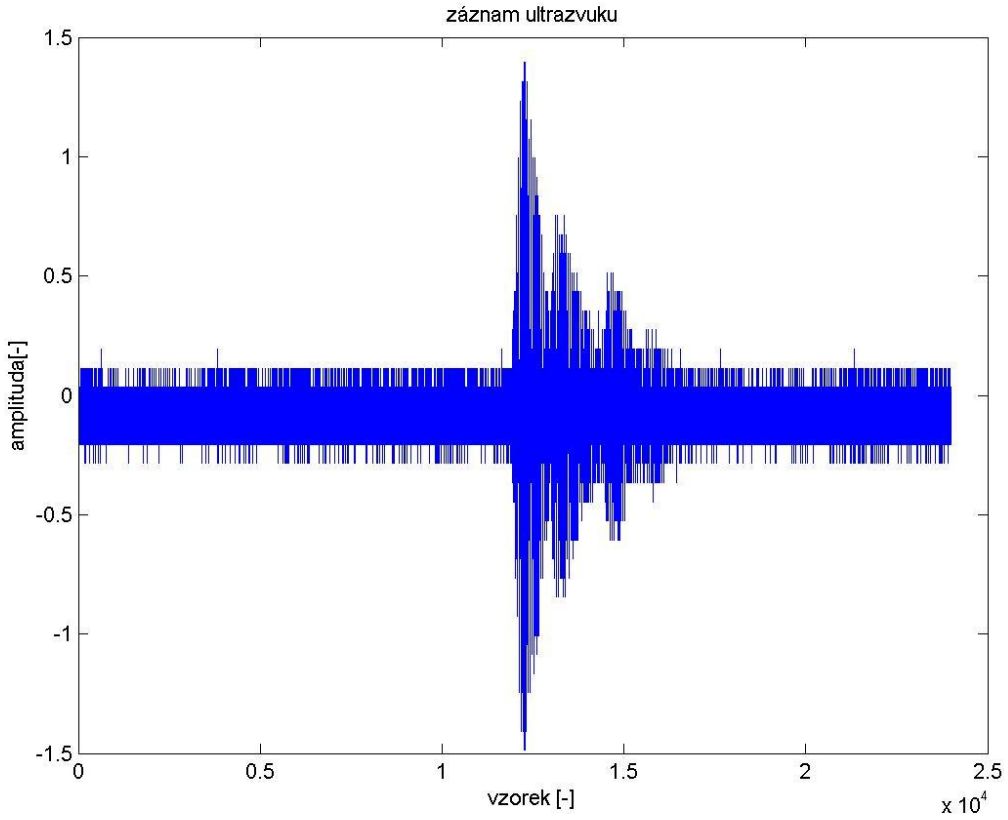
jejich chemické reakce. Jinou význačnou vlastností ultrazvuku je, že na rozdíl od obyčejného zvuku ho vzduch i jiné plyny značně pohlcují, a to tím víc, čím je jeho vlnová délka větší. V kapalinách, například ve vodě, se však může ultrazvukové vlnění rozšířit i do velkých vzdáleností. Malá absorpce ultrazvuku ve vodě umožňuje jeho praktické využití např. na rychlé a velmi pohodlné měření hloubek moří tzv. metodou ozvěny ultrazvuku. Zdroj ultrazvuku upevněný na lodi pod vodní hladinou vysílá velmi krátké ultrazvukové impulzy, které se po odraze na dně moře vrací a jsou zachycené detektorem ultrazvuku. Hloubku moře můžeme určit podle vztahu:

$$h = v * \frac{\Delta t}{2} \tag{1.4}$$

kde h je hloubka

v je rychlost zvuku ve vodě

Δt je časový rozdíl mezi vyslaným a zachyceným signálem



Obrázek 1 - Záznam ultrazvuku

Ultrazvuk se ve velké míře využívá i v tzv. ultrazvukové defektoskopii na hledání kazů v kovových výrobcích. Jakmile jsou v kovu trhliny, ultrazvukové vlny se na těchto místech odrazí, anebo absorbují a z takto vzniklého „akustického stínu“ nebo ze směru zpětných ultrazvukových paprsků můžeme odhadnout polohu a velikost trhliny nebo jiného kazu. Ultrazvuk se v širokém měřítku užívá i v lékařské diagnostice. (HAJKO, a další, 1973)

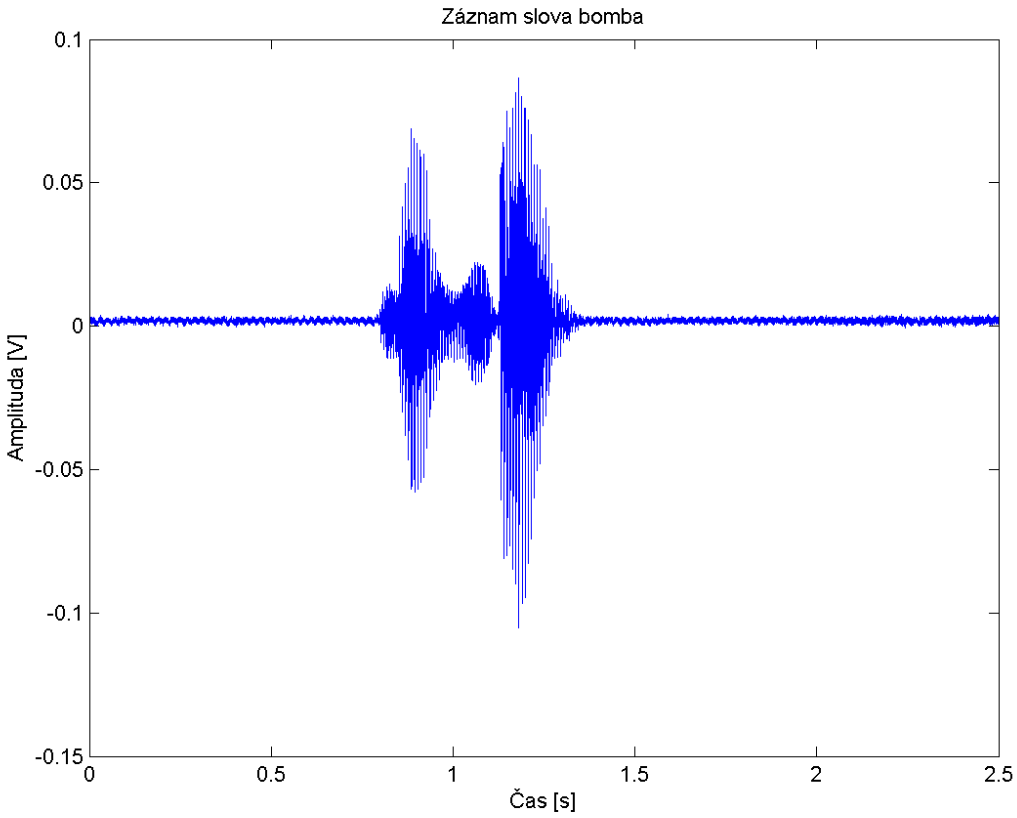
1.2 Slovo

Za nejmenší jednotku řeči, která může rozlišovat jednotlivá slova, lze považovat foném. Fonémy lze od sebe odlišit například podle způsobu a místa tvoření, podle artikulujícího orgánu nebo podle sluchového dojmu. (PSUTKA, 1995)

Počet fonémů v existujících světových jazycích se pohybuje od 12 do 60. V českém jazyce je jich 36, v anglickém 42, v ruském 40 apod. Fonémy se spojují do posloupnosti spojených celků, v nichž lze nalézt další stavební jednotku – slabiku (tu lze již přesně srovnávat s psanou formou). Libovolné promluvy jsou vlastně pravidelným opakováním různých posloupností slabik. Slovo je určitou kombinací slabik, přičemž jejich počet tvoří vždy celé číslo. Slovanské jazyky používají kolem 2500-3500 slabik a 45 000-50 000 slov.

Při běžném rozhovoru vysloví člověk asi $80-130$ slov za minutu, což představuje frekvenci výskytu asi 10 fonémů za sekundu. Jestliže uvážíme průměrné množství informace na jeden foném $H=3-4$ bit, dostaneme pro mluvenou řeč rychlost přenosu informace asi $30-40$ bit/s. Tento výsledek tedy charakterizuje informační obsah řeči objevující se v její fonetické struktuře. Z psychoakustických testů bylo zjištěno, že člověk je schopen zpracovat informaci o rychlosti maximálně 50 bit/s

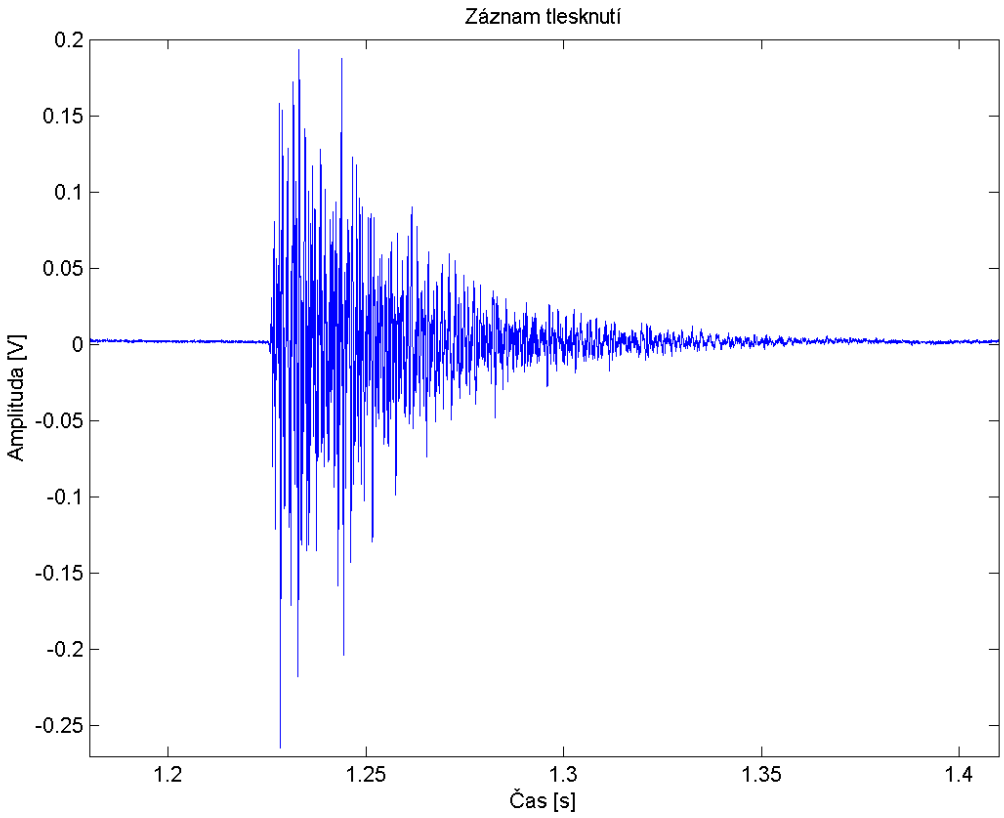
V českém jazyce používáme 36 různých fonémů, a jestliže jich několik spojíme, vznikne slabika. Ty už lze převést na psanou formu. A z jednotlivých slabik se skládají slova (PSUTKA, 1995).



Obrázek 2 - Záznam slova bomba

1.3 Zvuk s velkou dynamikou

Jestliže má zvuk velký rozdíl mezi maximální a minimální hlasitostí, můžeme o něm říct, že se jedná o zvuk s velkou dynamikou. Typickými příklady může být výstřel z pistole nebo třeba tlesknutí.

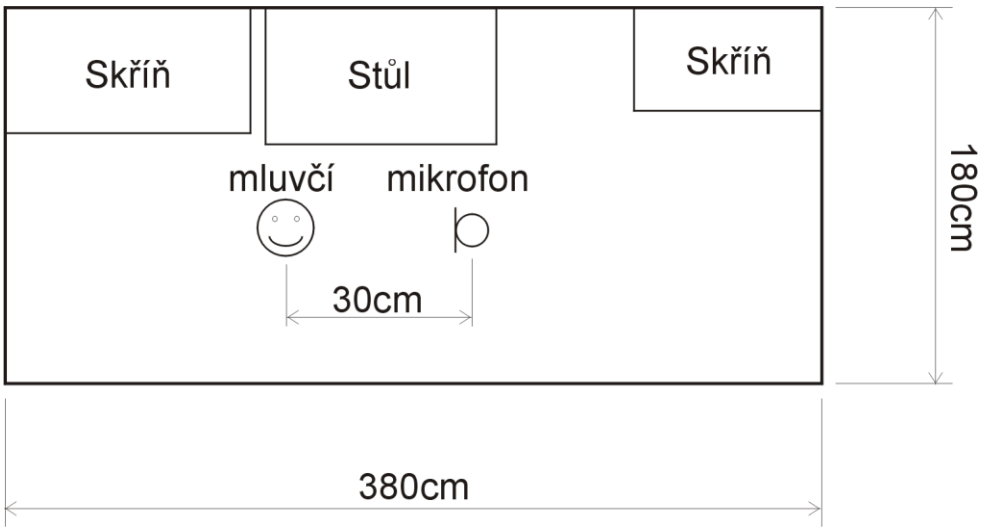


Obrázek 3 – Záznam zvuku s velkou dynamikou

Na obrázku vidíme, jak vypadá záznam tlesknutí. Pro tlesknutí je charakteristická ostrá vstupní hrana a pozvolné klesání amplitudy až k nulové hodnotě. Díky této ostré hraně je poměrně snadné přesně detekovat čas příchodu signálu na mikrofony, pomocí kterého lze dále vypočítat jeho polohu vzhledem k mikrofonům.

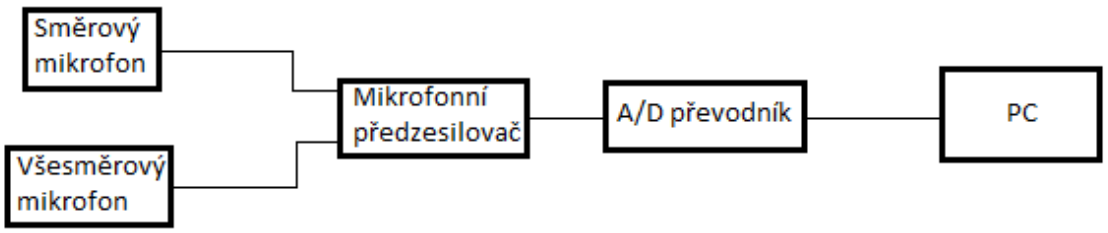
2 Záznam akustického signálu

Nahrávky slov, které jsou použity v diplomové práci, byly pořízeny v místnosti o rozměrech 380 x 180 cm. Mluvčí stáli ve vzdálenosti 30 cm od mikrofonů a mluvili přímo do mikrofonu.



Obrázek 4- Nahrávací místnost

Záznam akustického signálu provádíme pomocí dvojice kondenzátorových mikrofonů. Výstupní signál z mikrofonů je ale velice slabý, takže ho musíme zesílit pomocí mikrofonních zesilovačů. Zesílený signál, který vychází z předzesilovačů je ale analogový, a abychom byli schopni s ním pracovat, musíme ho převést na digitální signál. K tomu použijeme analogově digitální převodník, neboli tzv. A/D převodník. Tento digitální signál už můžeme v počítači zpracovávat pomocí programu Matlab.



Obrázek 5 - Schéma záznamu akustického signálu

Schéma, jak se může provádět záznam akustického signálu je znázorněno na obrázku 5. A následně si rozebereme principy funkce jednotlivých bloků.

2.1 Mikrofon

Mikrofon je základním prvkem pro nahrávání akustického signálu. Jeho funkcí je převést akustické vlnění na vlnění elektrické.

Princip mikrofonu spočívá v tom, že obsahuje membránu, která se rozechvěje, jestliže na ní přijde akustická vlna a toto kmitání je převedeno elektroakustickým měničem na elektrické kmitání. Nejpoužívanějšími druhy mikrofonů jsou elektrodynamické a kondenzátorové.

Nejdůležitějšími vlastnostmi mikrofonů je citlivost, směrová charakteristika a amplitudová frekvenční charakteristika. Citlivost vyjadřuje poměr mezi napětím na svorkách k akustickému tlaku, který napětí způsobil, a udává se v jednotkách $[\frac{mV}{Pa}]$ většinou na kmitočtu 1 kHz. Amplitudová frekvenční charakteristika je závislost výstupního napětí mikrofonu na kmitočtu při konstantním akustickém tlaku. Směrová charakteristika vyjadřuje závislost citlivosti na směr, ze kterého se zvuková vlna šíří na mikrofon. Směrovost se udává v polárních souřadnicích a závisí na druhu mikrofonu (HORÁK).

2.2 Elektrodynamický mikrofon

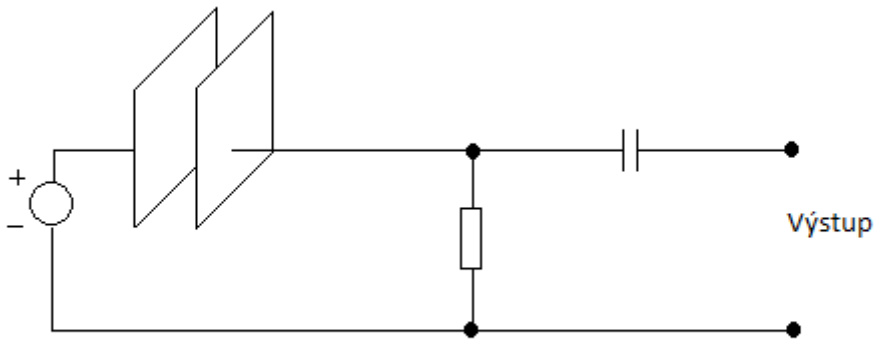
Elektrodynamické mikrofony jsou pravděpodobně nejpoužívanějšími mikrofony ze všech. Často se používají pro zpracování hlasitých zpěvů a různých hudebních nástrojů. Většinou jsou poměrně odolné proti mechanickému poškození, ale jsou méně citlivé než kondenzátorové mikrofony.

Principem elektrodynamického mikrofonu je membrána, která je spojená s cívkou volně posazenou k ornamentnímu magnetu. Jestliže přijde na membránu nějaký zvuk, dojde k jejímu pohybu a protože je s ní spojená cívka, dojde i k pohybu cívky. Jelikož se cívka bude pohybovat magnetickém poli, které vytváří permanentní magnet, bude se na cívce indukovat střídavé napětí, které bude mít stejný průběh jako akustická vlna.

2.3 Kondenzátorový mikrofon

Kondenzátorové mikrofony používají odlišný způsob snímání zvuku než elektrodynamické mikrofony. Kondenzátor se skládá ze dvou kovových desek, které jsou velmi blízko sebe, ale přitom se nedotýkají. Jestliže na tyto elektrody přivedeme stejnosměrné napětí, vznikne mezi deskami elektrické pole, ve kterém se i po odpojení udrží elektrický náboj. Obecně platí, že když zvětšíme vzdálenost mezi deskami, tak se

sníží jeho kapacita, ale zvýší jeho napětí z důvodu zachování elektrického náboje. Naopak při zmenšení vzdáleností mezi deskami dojde ke snížení napětí vlivem zvýšení kapacity. Tento jev využívá kondenzátorový mikrofon tak, že jedna deska je pevná a druhá deska slouží jako membrána. Pokud tedy na membránu dopadne akustický signál, bude se na takto vytvořeném kondenzátoru měnit napětí v závislosti na zvuku.



Obrázek 6 - Oddělení napájecího a signálového napětí

Nevýhodou je, že kondenzátorový mikrofon musí být napájený a zároveň musíme oddělit napájecí a signálové napětí. Oddělení provedeme přiřazením dalšího kondenzátoru, ten propustí požadovaný střídavý signál a napájecí stejnosměrné napětí nepropustí.

Kondenzátorové mikrofony jsou velice kvalitní a často se používají pro měřící účely nebo pro pořízení profesionálního záznamu. (HORÁK)

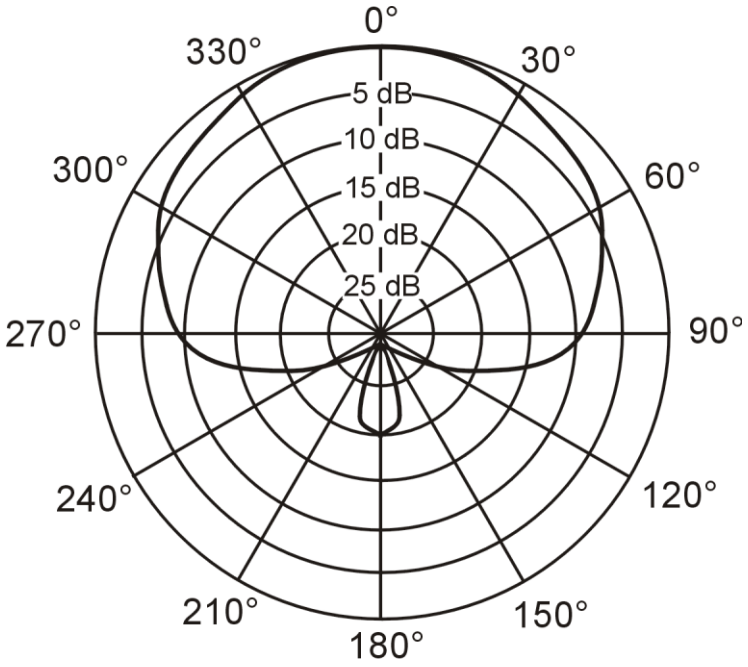
2.4 Použité mikrofony

Pro diplomovou práci jsou použity dva druhy kondenzátorových mikrofonů, přičemž má každý z nich svoji specifickou funkci. Prvním je směrový mikrofon MC 500 značky Dexon. Ten bude snímat mluvčího, jehož řeč budeme dále analyzovat. Druhým mikrofonem je všesměrový Behringer C-3, jehož funkcí je snímání zvuku z celé místnosti.

2.4.1 Dexon MC 500

Obrázek 7 - Směrový mikrofon dexon MC 500

Dexon MC 500 je kondenzátorový mikrofon, u kterého lze přepínat mezi širokou kardoidní a úzkou hyperkardiodní směrovou charakteristikou. Snímá zvuky ve frekvenčním pásmu od 60 Hz do 14 kHz. Jeho citlivost je v závislosti na nastavené směrové charakteristice -45 dB, respektive -30 dB s tolerancí ±3 dB. Výstupní impedance tohoto mikrofonu je 500 Ω, respektive 1600 Ω. Jeho výhodou je fakt, že pro napájení stačí použít vestavěnou baterii AA 1,5 V.



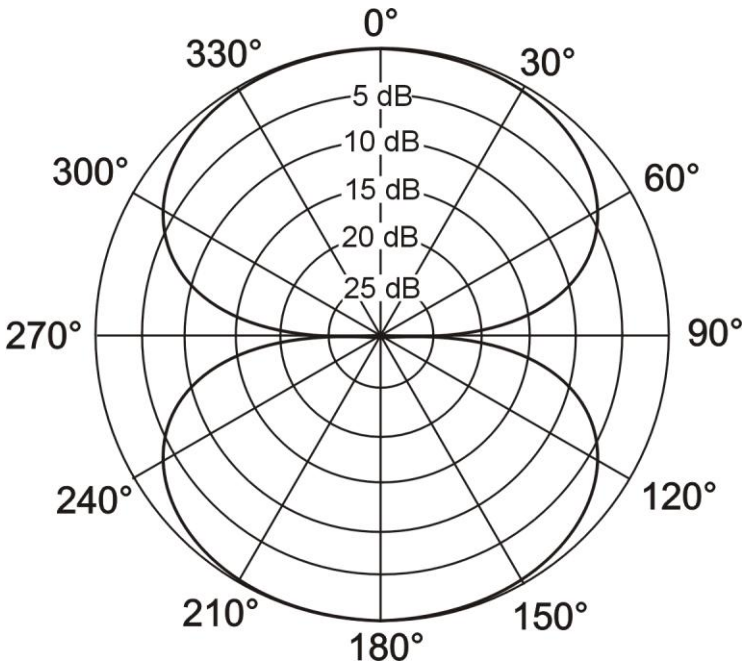
Obrázek 8- Směrová charakteristika mikrofonu

Jak je vidět na směrové charakteristice, mikrofon snímá zvuk pouze ze směru před mikrofonem. Zadní lalok je jen nepatrný a zvuk z tohoto směru je dostatečně potlačen.

2.4.2 Behringer C-3

Obrázek 9 - Všesměrový mikrofon Behringer C-3

Behringer C-3 je kondenzátorový mikrofon, u kterého lze nastavit podle potřeby jeho směrovou charakteristiku na kruhovou, kardoidní nebo osmičkovou charakteristiku. Citlivost tohoto mikrofonu dosahuje -40 dBV/pa a dokáže snímat zvuky ve frekvenčním rozsahu od 40 Hz do 18 kHz. Jeho výstupní odpor je 350 Ω a je napájen fantomovým napětím 48 V.



Obrázek 10 – Směrová charakteristika mikrofonu

Jak bylo zmíněno výše, mikrofon behringer C3 má osmičkovou směrovou charakteristiku. To znamená, že snímá zvuk se stejným ziskem z oblasti před i za mikrofonem.

2.5 Mikrofonní předzesilovač

Elektrický signál, který generuje mikrofon se pohybuje v řádech mV a je tudíž příliš slabý na to, abychom z něj byli schopni získat užitečnou informaci. Tento problém ale lze vyřešit použitím mikrofonního předzesilovače, který zesílí signál z mikrofonu. Při nahrávání zvuku pro účely praktické části diplomové práce byly použity mikrofonní předzesilovače typu Tube ultragain mic 100 od firmy Behringer.

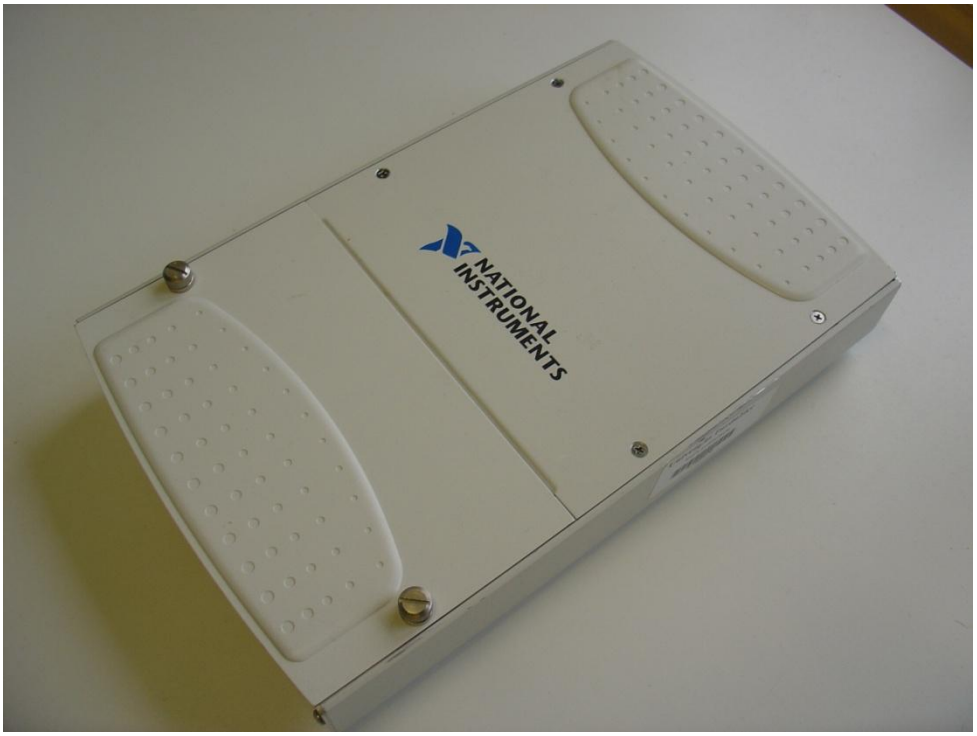
Obrázek 11 - Mikrofonní předzesilovač Behringer

Jedná se o elektronkový mikrofonní předzesilovač s limiterem, který využívá technologii UTC pro snížení šumu. Jelikož je vybaven fantomovým napájením +48 V, je vhodný pro aplikaci s kondenzátorovými mikrofony. Dále je tento zesilovač vybaven útlumovým článkem 20 dB, tlačítkem pro otočení fáze a symetrickými vstupy a výstupy.

2.6 A/D převodník

Abychom mohli zesílený signál z mikrofonů zpracovávat v PC, musíme ho převést z analogového na digitální signál pomocí A/D převodníku. Při nahrávání zvuku pro účely této diplomové práce byl použití A/D převodník značky National Instrument.

2.6.1 A/D NI USB-6251

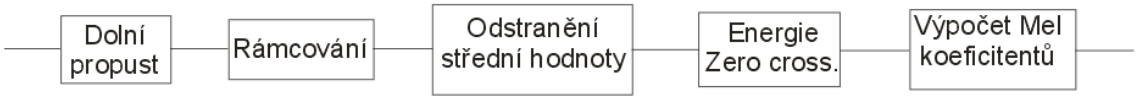


Obrázek 12 - A/D převodník NI USB- 6251

A/D převodník National Instrument USB – 6251 je 16-ti bitový převodník, který dokáže analogový signál vzorkovat s maximální rychlostí 1,25 Ms/s. Obsahuje 8 diferenciálních a 16 analogových vstupů.

3 Zpracování zaznamenaného signálu

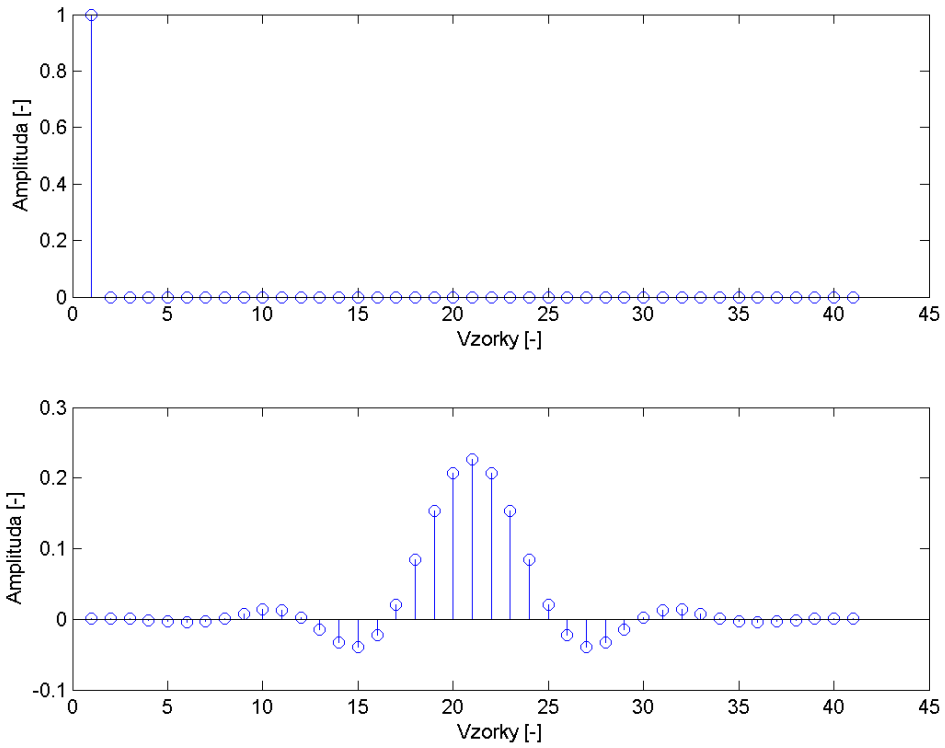
Zaznamenanou nahrávku upravíme pomocí programu Matlab různými operacemi, které nám umožní detekovat slova v nahrávce. V této kapitole si jednotlivé operace popíšeme. Postupovat budeme podle blokového schématu na obrázku 13



Obrázek 13- Blokové schéma zpracování signálu

3.1 Filtrace signálu

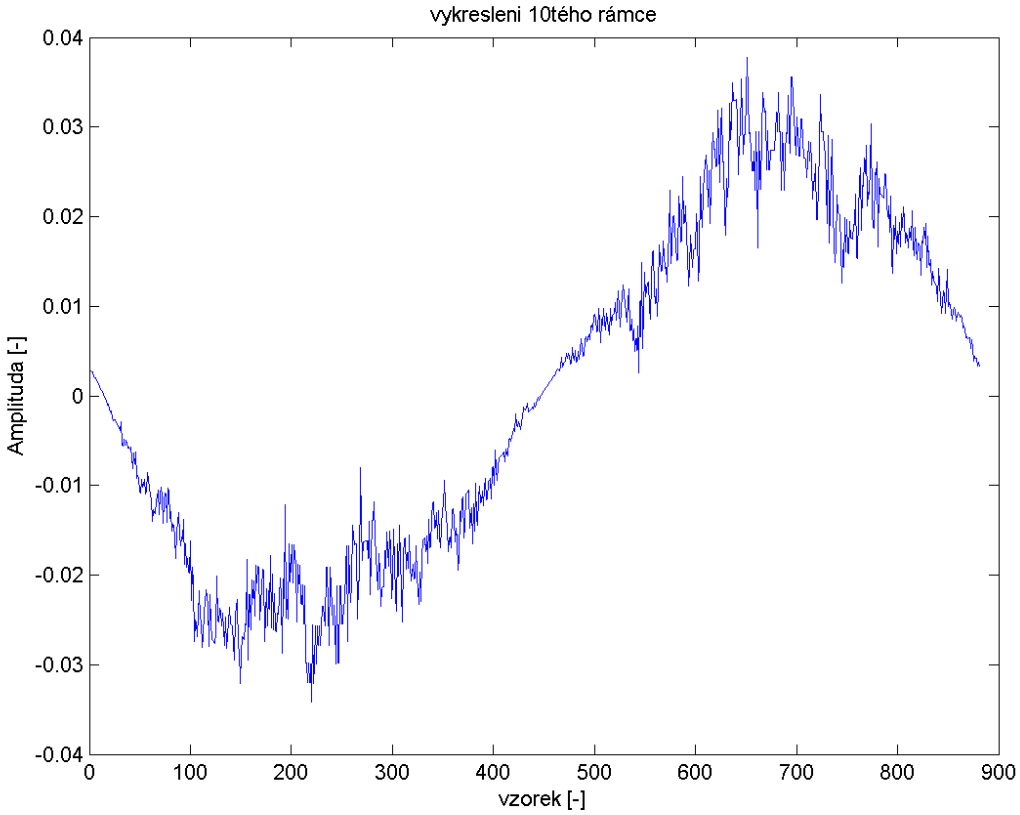
První operací, kterou budeme se signálem provádět, je filtrace signálu. Ta potlačí šum, který nahrávka obsahuje, takže se zlepší její vlastnosti a usnadní se tím detekce hledaných slov. K filtraci použijeme filtr s konečnou impulsní odezvou, takzvaný FIR filtr typu dolní propust s mezní frekvencí 7500 Hz a řád filtru zvolíme 40.



Obrázek 14- Impulsní odezva FIR filtru

3.2 Rozdělení signálu do rámců

Dalším krokem je to, že si celý vyfiltrovaný signál rozdělíme do mnoha kratších úseků, takzvaných rámců. Důvodem pro tento krok je, že celý signál považujeme za náhodný a nestacionární a metody, pomocí kterých chceme získat ze signálu hledaná slova, pracují dobře pouze se stacionárními signály. Takže musíme zvolit délku rámců dostatečně malou, abychom mohli signál považovat za stacionární, ale na druhou stranu musí být délka rámce dostatečně dlouhá, aby bylo odhadnutí parametrů přesné. Těmto podmínkám nejlépe odpovídá délka jednoho rámce 20 ms. Pro lepší zpracování signálu se jednotlivé rámce mezi sebou navzájem překrývají o 10 ms, takže je každý vzorek signálu obsažen ve 2 rámcích.



Obrázek 15 - Vykreslení 10tého rámce

Na obrázku č. 15 vidíme, jak vypadá 10. rámec slova bomba. Jak už bylo řečeno, rámec je dlouhý 20 ms, nahrávka byla pořízena se vzorkovací frekvencí 44000 Hz, z čehož vyplývá, že jeden rámec obsahuje 880 vzorků.

3.3 Odstranění ss složky

Po rozdělení signálu do kratších rámců, odstraníme stejnosměrnou složku signálu. Můžeme si to dovolit kvůli tomu, že stejnosměrná složka nenese žádnou užitečnou informaci o signálu, naopak může způsobit špatné určení parametrů, jako je krátkodobá energie signálu. Odstranění stejnosměrné složky je poměrně jednoduchá operace a provede se odečtením střední hodnoty od původního signálu.

$$s'[n] = s[n] - \mu_s \tag{3.1}$$

kde $s'[n]$ je n-tý vzorek signálu bez ss složky

$s[n]$ je n-tý prvek signálu

μ_s je střední hodnota signálu

Jestliže máme k dispozici celý signál, tak můžeme určit jako střední hodnotu jeho průměrnou hodnotu:

$$s = \frac{1}{N}\sum_{n=1}^{N} s[n] \tag{3.2}$$

kde N je celkový počet vzorků signálu

s[n] je n-tý vzorek signálu

3.4 Střední krátkodobá energie a intenzita

Důležitou charakteristikou slova je střední krátkodobá energie, pomocí které lze detekovat začátek a konec slova. Také podle ní můžeme určit, jestli slovo obsahuje znělé nebo neznělé souhlásky. Pro znělé souhlásky (b, u, d, ď, g, h, m, n, j, l, r, ř atd.) platí, že mají vysokou energii a neznělé souhlásky (p, t, k, ch, s, c atd.) mají energii nízkou. Střední krátkodobá energie se vypočítá podle vztahu:

$$E = \frac{1}{L_{ram}}\sum_{n=0}^{Lram} x^2[n] \tag{3.3}$$

kde E je střední krátkodobá energie

L_{ram} je počet vzorků v rámci.

x[n] je n-tý vzorek signálu

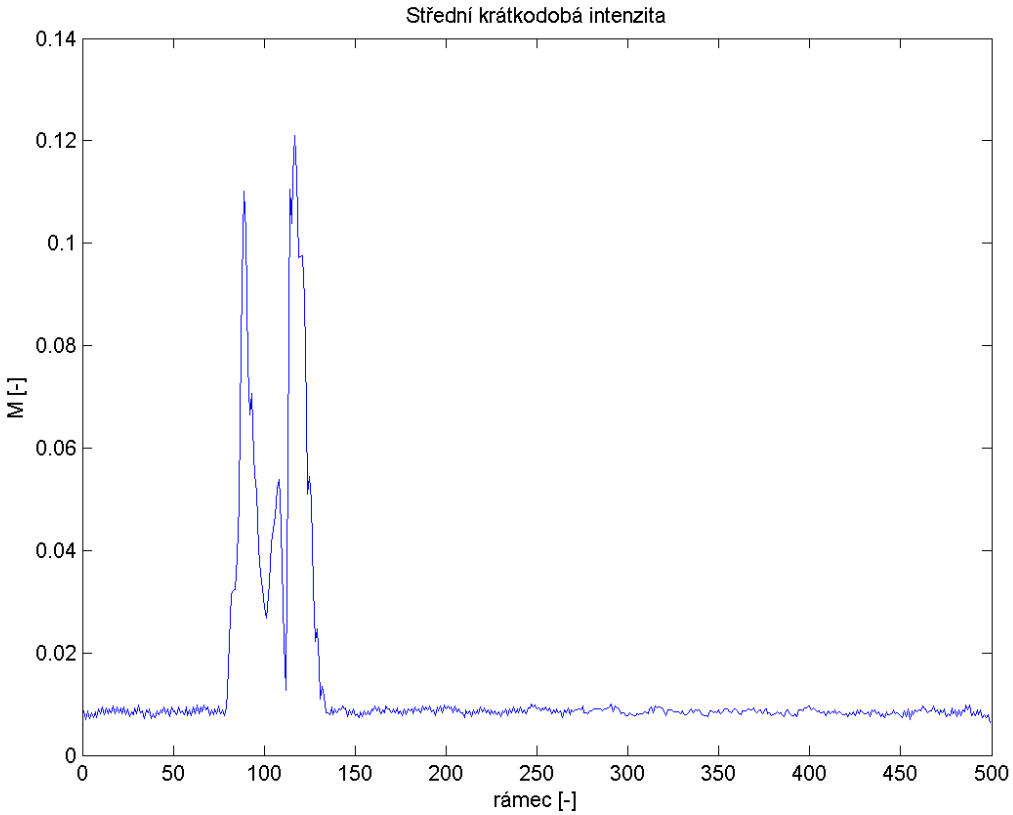
Kvůli zmenšení dynamického rozsahu se místo energie pracuje se střední krátkodobou intenzitou:

$$M = \frac{1}{L_{ram}} \sum_{n=0}^{L_{ram}-1} |x[n]| \tag{3.4}$$

kde M je střední krátkodobá intenzita

L_{ram} je počet vzorků v rámci

x[n] je n-tý vzorek signálu



Obrázek 16 - Střední krátkodobá intenzita

3.5 Počet průchodů nulou

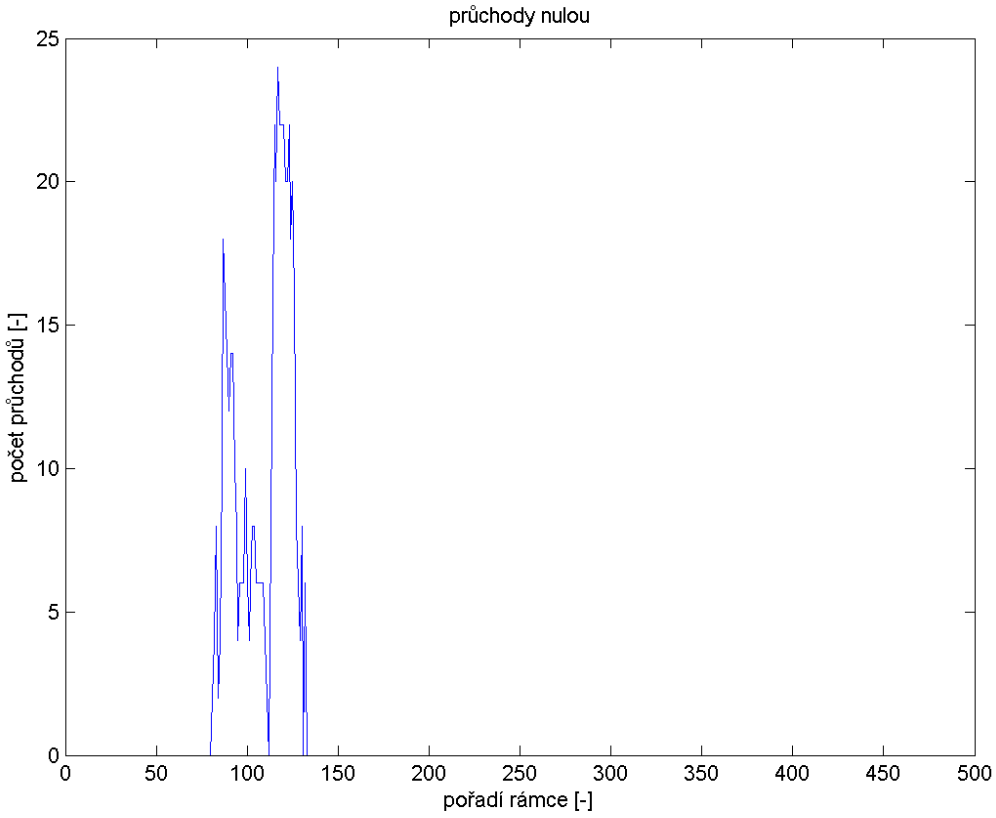
Další důležitou charakteristikou je počet průchodů nulou v jednotlivých rámcích, pomocí které můžeme určit znělé a neznělé souhlásky ve slově. Pro znělé souhlásky platí, že mají nízkou hodnotu průchodů nulou a neznělé mají vysokou hodnotu průchodů nulou v jednotlivých rámcích. Počet průchodů nulou vypočítáme podle vztahu:

$$Z = \frac{1}{2}\sum_{n=1}^{L_{ram}} |sign\ x[n] - sign\ x[n-1]| \tag{3.5}$$

kde $\begin{Bmatrix} sign\ x[n] = 1 & pro\ x[n] > 0 \\ sign\ x[n] = -1 & pro\ x[n] < 0 \\ sign\ x[n] = 0 & pro\ x[n] = 0 \end{Bmatrix}$

Z je počet průchodů nulou

x[n] je n-tý vzorek signálu

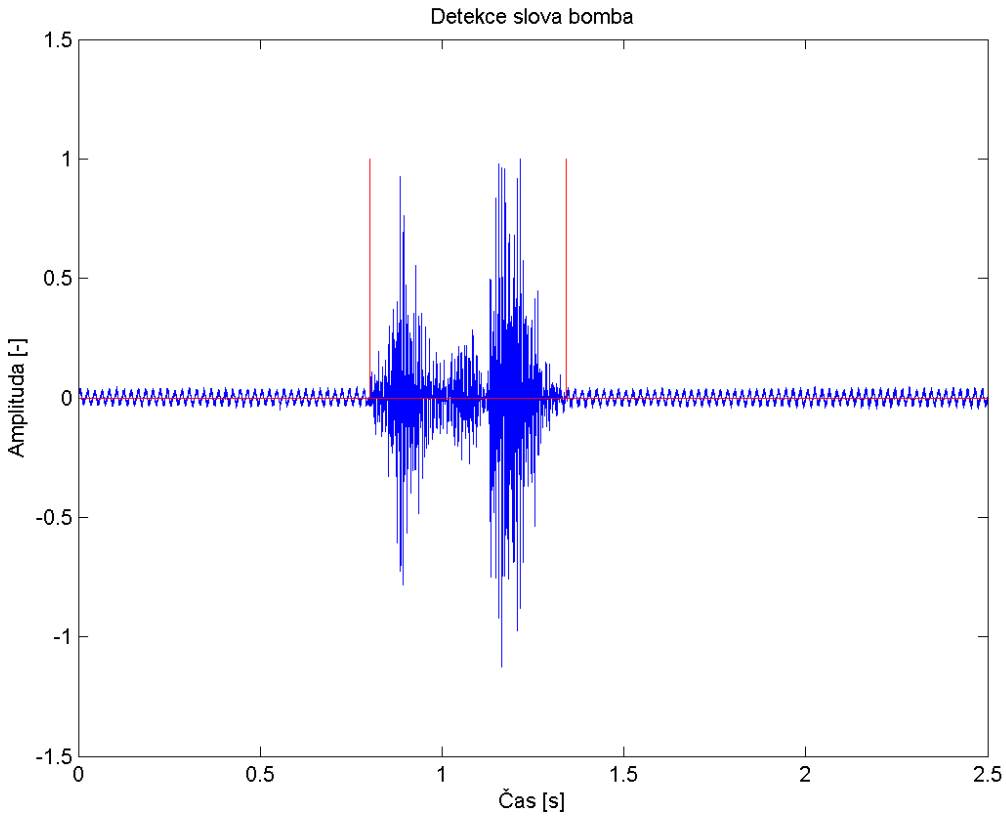


Obrázek 17 - Počet průchodů nulou v jednotlivých rámcích

3.6 Detekce slova

Detekce začátku a konce slova je v ideálním případě, kdy je zabezpečen vysoký odstup signálu od šumu, snadná na základě zvýšené intenzity signálu. Jenže obvykle pracujeme v prostředí, kde se v pozadí vyskytuje značný šum. Kvůli tomu je vhodnější použít kombinaci průběhů krátkodobé intenzity a počtu průchodu nulou v jednotlivých rámcích.

Pro detekci začátku a konce slova je velice podstatné, jestli slovo začíná nebo končí znělou nebo neznělou souhlásku. Pro znělé souhlásky platí, že mají vysokou energii. Pro neznělé souhlásky zase platí, že mají vysoký počet průchodů nulou. Takže si stanovíme horní a dolní mez krátkodobé intenzity a průchodů nulou a jestliže signál v několika rámcích za sebou překročí horní mez, můžeme předpokládat začátek slova. Konec slova nastane, jestliže krátkodobá intenzita nebo počet průchodů nulou klesne pod dolní mez v několika rámcích za sebou.



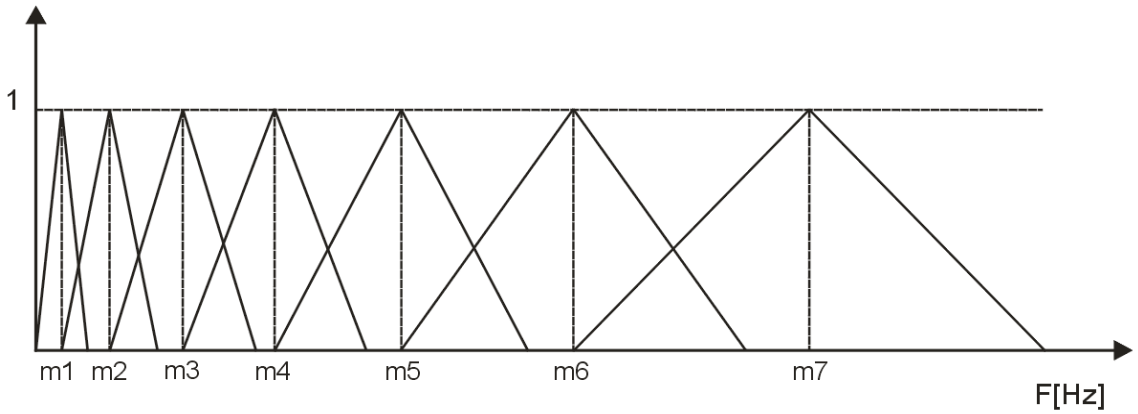
Obrázek 18 - Detekce slova bomba

Na obrázku č. 18 je zobrazen záznam slova bomba se zvýrazněnými body začátku a konce promluvy určenými pomocí průběhů krátkodobé intenzity a průchodů nulou.

3.7 Kepstrální koeficienty

Když už máme v signálu detekovaná jednotlivá slova a chceme je mezi sebou porovnávat, musíme převést vzorky v jednotlivých rámcích na kepstrální koeficienty, které budou reprezentovat jednotlivá slova. Konkrétně budeme vytvářet Mel-kepstrální koeficienty, přestože je to metoda poměrně jednoduchá, vede k dobrým výsledkům.

Metoda mel-kepstrálních koeficientů klade větší důraz na nižší kmitočty a je to metoda s nelineárně rozloženou bankou filtrů ve spektru. Rozložení filtrů je znázorněno na obrázku.



Obrázek 19 - Rozložení filtrů ve frekvenční oblasti

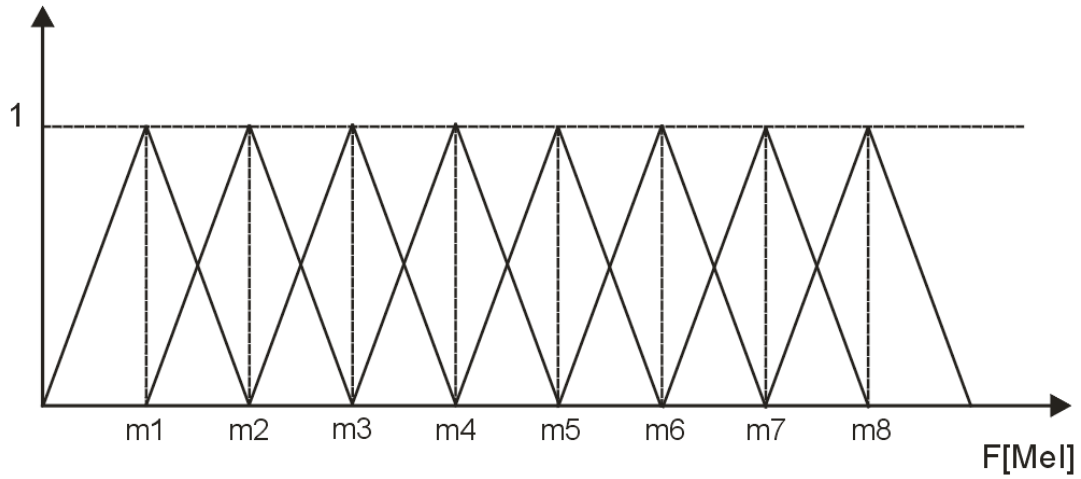
Toto rozložení filtrů se volí z toho důvodu, že citlivost ucha na všechny frekvence není konstantní a s rostoucí frekvencí klesá. Při návrhu filtrů nejprve převedeme frekvenční osu z Hertzů na Mely podle vztahu:

$$F_{mel} = 2959 log_{10}\left(1 + \frac{F_{Hz}}{700}\right) \tag{3.6}$$

kde F_{mel} je frekvence v Melech

F_{Hz} je frekvence v Hertzích

Jak vypadá banka filtrů po převedení frekvenční osy z Hertzů na Mely je znázorněno na obrázku 20.



Obrázek 20 – Rozložení filtrů

Při výpočtů mel koeficientů postupujeme tak, že nejdříve signál převedeme do spektra za pomocí DFT, které následně umocníme a poté vynásobíme trojúhelníkovým filtrem. To způsobí, že hodnoty spektra na okrajích filtrů jsou značně potlačeny a ve středu jsou téměř nezměněny. Pak sečteme všechny hodnoty v každém okně a umocníme je na druhou a získáme energie e_k. Tyto energie pak zlogaritmujeme a za pomocí DCT získáme koeficienty MFCC.

$$c_m(n) = \sum_{k=1}^{K} log\ e_k\ cos\left[n(k-\in 0{,}5)\frac{\pi}{K}\right] \text{ pro n} \in \langle 0, M\rangle \tag{3.7}$$

kde K je počet filtrů

e_k je energie na intervalu daného filtru

M je počet mel kepstrálních koeficientů

Maximální počet koeficientů je stejný, jako je počet filtrů, ale protože je hlavní informace obsažena v několika prvních koeficientech, používá se obvykle jen 10 až 13 koeficientů. Velmi často se přidává na začátek ještě jeden koeficient, který je roven logaritmu krátkodobé energie přímo z řečového signálu rámce (PSUTKA, a další, 2006).

$$c_m(0) = log \sum_{n=0}^{L_{ram}-1} x^2[n] \tag{3.8}$$

4 Rozpoznávání slov

O počítačové rozpoznání řeči se odborníci zajímají již 50 let. Přesto se rozpoznat libovolné slovo z mluvy neznámého řečníka stále dokonale nedaří. Důvody nezdarů se skrývají v obrovské variabilitě mluvčího, v prostředí, kde se záznam provádí, ale také v obtížnosti řešené úlohy. Každý člověk má originální hlasové ústrojí a odlišný způsob artikulace, to se projevuje rozdílnou barvou hlasu, přízvukem, rychlostí řeči atd. I hlas jednoho řečníka je variabilní a závislý na mnoha aspektech (otázka, příkaz, nálada, nemoc atd.). To se projevuje v délce jednotlivých úseků řeči i v intenzitě řečového signálu. Ve skutečnosti je vlastně nemožné, aby bylo slovo řečeno dvakrát naprosto stejně. Rozpoznávače řeči můžeme podle složitosti rozdělit do tří skupin (PSUTKA, a další, 2006):

- **Rozpoznávání izolovaných slov** (malý slovník, např. číslovky, povely).
- **Rozpoznávání diskrétního diktátu** (rozsáhlejší slovník, slova jsou vyslovována izolovaně s krátkou mezislovní pauzou).
- **Rozpoznávání souvislé řeči** (slovník na desítky tisíc slov).

4.1 Metoda borcení časové osy DTW

Tato metoda slouží především k rozpoznávání izolovaných slov nebo krátkých úseků promluvy. Vzory pro rozpoznávání se většinou realizují referenční nahrávkou. Tato nahrávka je parametrizována a uložena do databáze modelů pro určitý rozpoznávací úkol. Jak z předchozího vyplývá, trénování modelů spočívá v nahrání požadovaných slov a jejich parametrizace. Nevýhody této metody rozpoznávání řeči jsou v závislosti rozpoznávacího systému na mluvčím a dále je rozpoznávání ve většině případů omezeno na celá izolovaná slova. Výhodami tohoto systému jsou jednoduchost a nenáročnost (trénovací fáze se redukuje na náhradní potřebných vzorů) (VOPIČKA, 2002).

Vstupem pro tuto metodu jsou slova, která jsou reprezentována sekvencí vektorů. Říkáme jim obrazy slova, jejich délka je dána délkou slova a velikost počtem kepstrálních koeficientů. Metoda DTW je založena na porovnávání neznámého testovaného a referenčního obrazu slova:

$$O = [o(1), o(i), \dots, o(T)]$$

kde o(i) jsou vektory kepstrálních koeficientů testovaného obrazu

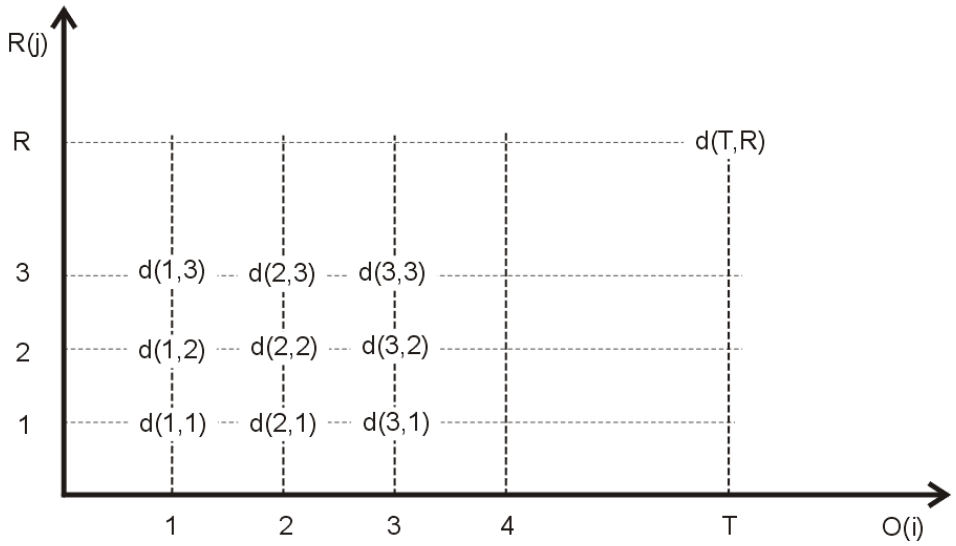
$$R = [r(1), r(j), \dots, r(R)]$$

kde r(j) jsou vektory kepstrálních koeficientů referenčního obrazu

Testovaný obraz se přiřadí k tomu referenčnímu obrazu, od kterého má nejmenší vzdálenost (součet odchylek jednotlivých vektorů slova). Největší odlišnosti mezi jednotlivými slovy nejsou ve spektrální oblasti, ale jsou způsobeny odlišnými délkami slov nebo vnitřních částí slova. Metoda DTW minimalizuje rozdíly mezi obrazy borcením časové osy jednoho z nich. Pomocí lokálních vzdáleností mezi body obrazů v rovině (T, R) určíme transformační funkci, která optimálně přizpůsobí referenční slovo testovanému. Při výpočtu funkce DTW většinou uvažujeme testované příznaky podél horizontální osy a referenční podél osy vertikální. Lokální vzdálenosti se můžou jednoduše určit:

$$d(o,r) = d(d\big(o(i)r(j)\big) = \sqrt{\sum_{n=1}^{M} |o_n(i) - r_n(j)|^2} \tag{4.1}$$

Z těchto vzdáleností vytvoříme matici o rozměrech TxR.

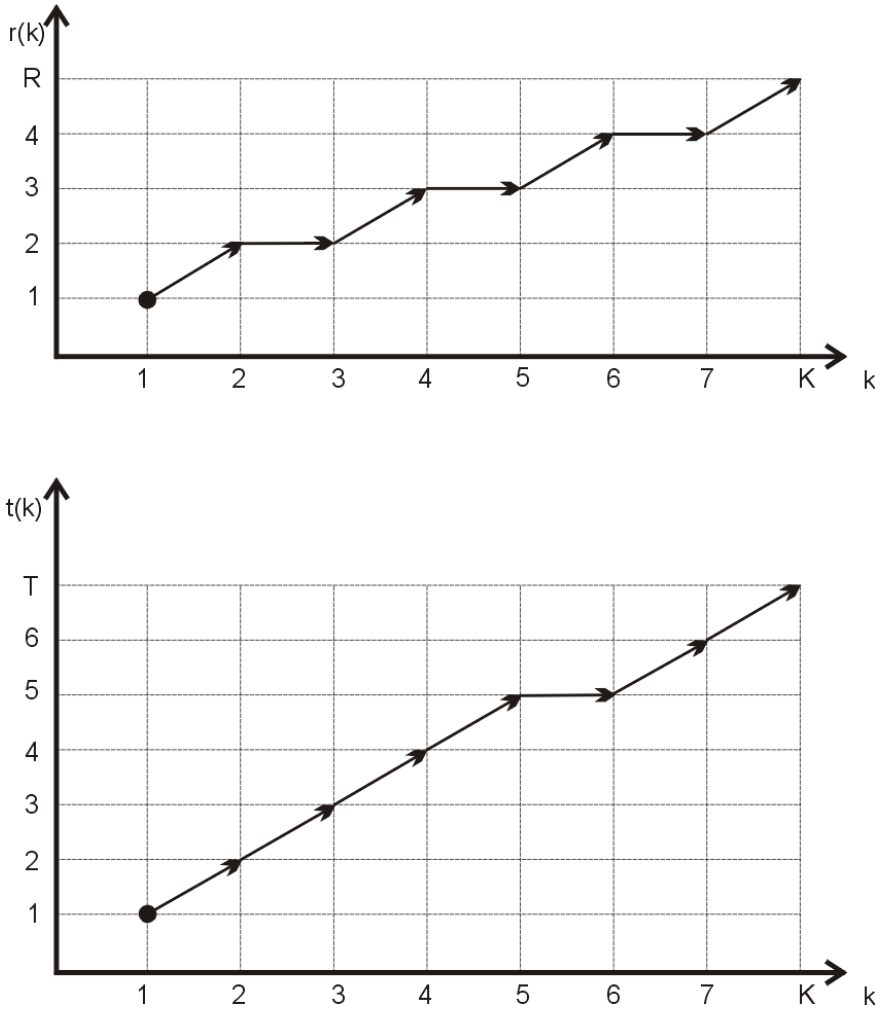


Obrázek 21 - Matice vzdáleností DTW

Po vytvoření této matice si spočítáme cesty uprostřed matice vedoucí z počátku d(1,1) do konce d(T, R). Dále zavedeme časovou proměnnou k=1,2,3,...,K, která odpovídá délce porovnávací cesty a na ni závislé transformační funkce:

- referenční sekvenci cesty r(k)
- testovanou sekvenci cesty t(k)

Abychom mohli porovnávat referenční a testovaný obraz, musí být obě cesty stejně dlouhé, i když jsou obrazy rozdílné velikosti.



Obrázek 22 - Ukázka aplikace proměnné k na funkci DTW

Hledáme cestu s celkovou minimální vzdáleností. Pro její nalezení je ale potřeba prozkoumat všechny možné cesty. (KOČÍ, 2010)

4.1.1 Lokální omezení DTW cesty

Aby nedocházelo k nadměrné kompresi nebo expanzi časového měřítka u porovnávaných obrazů, použijeme na funkci DTW omezení monotónnosti a souvislosti.

$$0 \leq r(k) - r(k-1) \leq R^x$$

$$0 \leq t(k) - t(k-1 \leq T^x$$

kde T^x,R^x jsou kladná celá čísla

Jedná se o omezení délky kroku funkce DTW. Funkce nesmí v dané délce vynechat žádné segmenty, jestliže bude jedna z hodnot T^x ,R^x rovna jedné. Většinou se ve funkci nechávají vynechat maximálně dva segmenty za sebou. Dále se zavádí omezení strmosti, které zabrání tomu, aby byla funkce příliš strmá. To funguje tak, že když funkce postupuje ve směru jedné osy n krát za sebou, není jí dovoleno v tomto směru pokračovat, dokud nepostoupí m krát ve směru jiném (KOČÍ, 2010).

4.1.2 Výpočet vzdáleností

Každá cesta je jednoznačně dána svou délkou K_c, a průběhem funkcí $t_c(k)$ a $r_c(k)$. Pro tuto cestu se vzdálenosti mezi obrazy O a R vypočítají takto:

$$D_c(O,R) = \frac{1}{N_c(\widehat{W})} \sum_{k=1}^{K_c} d[o(t_c(k)), r(r_c(k))] \widehat{W_c(k)} \tag{4.2}$$

kde d[o(...),r(...)] je vzdálenost dvou vektorů

$\widehat{W_c}(k)$ je váha odpovídající k-tému kroku cesty,

$N_c(\widehat{W})$ je normalizační faktor závislý na vahách

Vzdálenost mezi obrazy O a R je dána jako minimální vzdálenost ze všech možných cest:

$$D(O,R) = \min_{\{c\}} D_c(O,R) \tag{4.3}$$

Druhy váhových funkcí $\widehat{W_c}(k)$ a normalizačních faktorů volíme podle typu lokálního omezení (KOČÍ, 2010).

4.1.3 Postup vypočítání optimální cesty funkce DTW

Při výpočtu optimální cesty postupujeme tak, že hledáme minimální cestu už od začátku a ne až na konci. Pomocí znalosti, z jakých předchozích bodů se do současného můžeme dostat (lokální omezení), vybíráme od začátku pouze ty nejlepší varianty pro jednotlivé body. Na konci obrazců v bodě (T, R) je pak k dispozici velikost cesty s nejmenší vzdáleností. Postup této metody je následující (ČERNOCKÝ, 2006).

Nejdříve vytvoříme matici d s rozměry T x R a do ní vložíme vzdálenosti jednotlivých referenčních a testovaných vektorů.

Pak vytvoříme matici g, která má oproti matici d navíc nultý řádek a nultý sloupec, a vložíme do ní částečně kumulované vzdálenosti. Nultý sloupek a řádek definujeme takto:

$$g(0,0) = 0$$

$$g(0, m \neq 0) = g(n \neq 0,0) = \infty$$

Částečně kumulovanou vzdálenost, kterou vložíme do matice g, vypočítáme takto:

$$g(m,n) = \min_{\forall \text{předchůdci}} [g(\text{předchůdce}) + d(m,n)\widehat{W}(k)] \quad (4.4)$$

- Možní předchůdci určíme podle tabulky lokálního omezení.
- Váha $\widehat{W}(k)$ odpovídá pohybu z předchůdce do bodu [m,n].

Konečnou minimální vzdálenost mezi referenčním a testovaným obrazem určíme podle vztahu:

$$D(O,R) = \frac{1}{N(\widehat{W})} g(T,R) \quad (4.5)$$

4.1.4 Normalizační faktor

Tento faktor je zaveden, aby kompenzoval délku cesty funkce DTW. Je závislý na váhové funkci a můžeme ho spočítat jako (ČERNOCKÝ, 2006):

$$N(\widehat{W}) = \sum_{k=1}^{K} \widehat{W}(k). \quad (4.6)$$

5 Praktické využití rozpoznávání slov

Na trhu existuje velké množství aplikací, které využívají rozpoznávání lidské řeči. V této kapitole jsou uvedeny některé z nich.

5.1 SpeechTech ASR

SpeechTech ASR je modul určený k použití v dalších systémových celcích, kde je požadováno využití technologie rozpoznávání řeči. A to jak souvislých promluv, tak i izolovaných frází a slov. Tento modul je interně využit i v řadě dalších vlastních produktů jako jsou SpeechTech Spojovatelka, SpeechTech MegaWord, SpeechTech IVR a další.

Jedná se o na řečníkovi nezávislý (speaker independent) systém rozpoznávání řeči. Umí rozpoznávat buď fráze popsané gramatikami nebo "volné" promluvy souvislé řeči s využitím velkých slovníků (Large Vocabulary Continous Speech Recognition - LVCSR). Systém umožňuje rozpoznávat souvislou řeč v reálném čase se slovníky až o několika stech tisících slovech.

Podle rozsahu slovníku existují dvě verze:

- ASR standard – zhruba do 1000 slov
- ASR large - rozpoznávání více jak 1000 slov

V současnosti je podporována hlavně čeština. Podpora dalších jazyků se připravuje.

SpeechTech ASR je k dispozici, jak ve verzi pro telefonní aplikace určené k integraci se systémy pracujícími v prostředí telefonních nebo IP sítí, tak i ve verzi pro mikrofonní aplikace, jako jsou například různé diktovací nebo jiné desktopové aplikace.

Telefonní verze je určena pro rozpoznávání řeči v běžném telefonním hovoru v běžném relativně tichém prostředí. Kromě běžného sluchátka ASR rovněž rozpoznává hovory z hlasitého handsfree či z bezdrátových a bluetoothových headsetů. Mikrofonní verze ASR je vhodná pro použití rozpoznávání záznamů do diktafonů nebo rozpoznávání s kvalitním headsetem připojeným k PC, buď do integrované, nebo externí zvukové karty (SpeechTech).

SpeechTech ASR jsou k dispozici s různými rozhraními jako například:

- s DLL rozhraním pro integraci například s vlastními dektopovými aplikacemi zákazníka
- se SpeechTech MRCP rozhraním pro snadnou integraci s IVR telefonními aplikacemi renomovaných výrobců

- se síťovým rozhraním

SpeechTech ASR podporuje v současnosti tyto platformy:

- Intel, Windows (2K, XP, 2003, Vista, 7), 32 bitů
- Intel, Windows (2003, Vista, 7), 64 bitů - jako 32 bitové knihovny
- Linux, 32 bitů
- Linux, 64 bitů

5.2 Dragon Dictation

Dragon Dictation je aplikací od společnosti Nuance Communications, která slouží jako náhrada sekretářky, dokáže přepsat řeč do textu. Funguje tak, že se na telefonu namluví požadovaná slova a zvuková stopa bude posléze odeslána na server. Server zajistí převod řeči na text a odešle ho zpět do telefonu. Nejedná se tedy o aplikaci, která by nějakým chytrým algoritmem překládala řeč na text přímo v telefonu, pro její použití je nutné mobilní datové připojení nebo Wi-fi síť.

Rozhraní aplikace je velmi jednoduché, tvoří ho víceméně jen dvě obrazovky. Na té první je stříbrno červené tlačítko, pomocí kterého zahájíte nahrávání. Konec řeči je možné nechat detekovat automaticky, nebo ho ukončit stiskem tlačítka.

Výsledný rozpoznaný text je zobrazen v okně na další obrazovce. Odtud je možné ho zkopírovat do systémové schránky a následně ho vložit do jiné aplikace. Je zde i možnost ho odeslat e-mailem, SMS či ho publikovat na Facebooku nebo Twitteru.

Chce-li uživatel diktovat hodně dlouhý text, nemusí ho odříkat najednou. Při každém dalším rozpoznání se text doplní za ten stávající (ŠIMON, a další, 2012).

5.3 Dragon Search

Dragon Search je taktéž aplikace od společnosti Nuance Communications, je zaměřená vyloženě jen na vyhledávání. Jejím základem je opět rozpoznávací hlasový modul, který přepíše řeč na text, ten pak vloží do vyhledávače. Uživateli se na displeji rovnou zobrazí výsledky jeho dotazu.

V nastavení aplikace je možnost nastavení, který z vyhledávačů bude použitý jako výchozí. Na výběr je Google a Yahoo, tím ale možnosti aplikace nekončí. Po rozpoznání textu si můžeme mimo těchto dvou vybrat ještě hledání na Wikipedii, Twitteru, YouTube a iTunes.

Princip fungování je totožný, jako u Dragon Dictation, jen s tím rozdílem, že výsledný text je rovnou vložen do vyhledávacího políčka. Vedle něj je tlačítko symbolizující nahrávání, kterým můžeme rychle spustit další rozpoznávání a nemusíme se vracet na výchozí obrazovku (ŠIMON, a další, 2012).

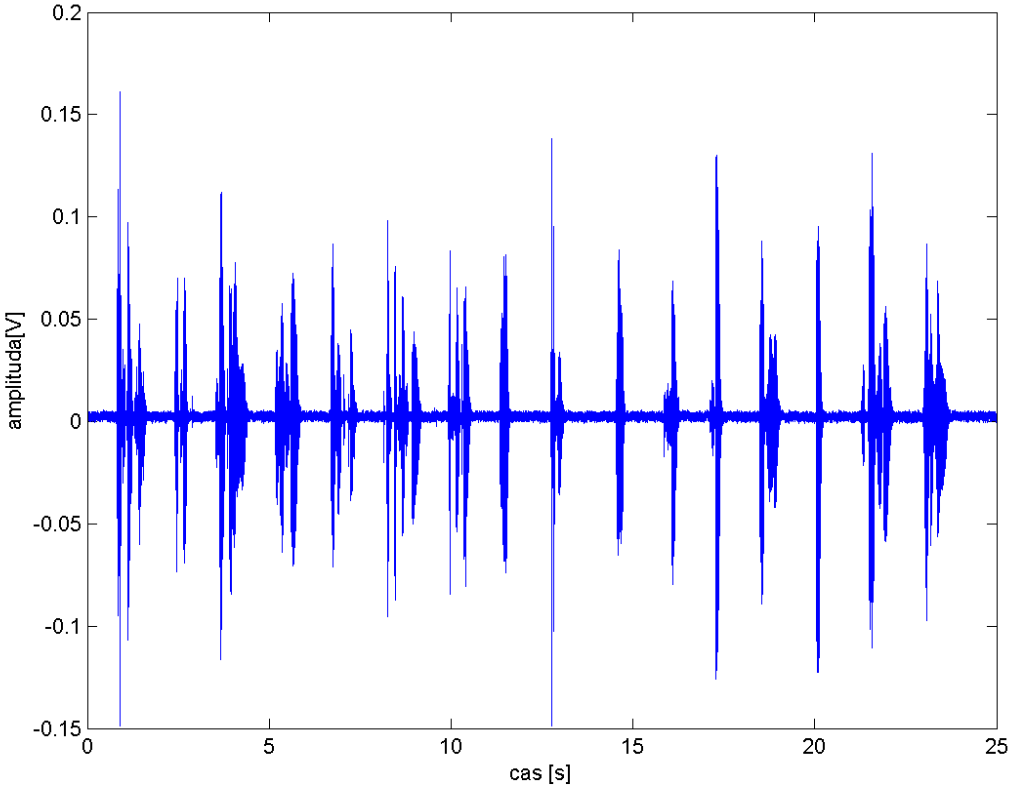
6 Praktická část

V praktické části budeme analyzovat, jak bude úspěšné rozpoznávání slov s klesajícím odstupem signálu od šumu. Testovat budeme souvětí: "Rostoucí počet vloupání zvyšuje nabídku bezpečnostní techniky na trhu. Ne vždy se jedná o spolehlivé zařízení." Věta byla záměrně vybrána kvůli tomu, že neobsahuje žádné spřažené spojky, které by mohli ovlivnit detekování jednotlivých slov a jejich následné rozpoznávání.

6.1 Vytvoření databáze slov

Pro záznam akustického signálu použijeme postup a vybavení popsané v kapitole 2. Nahrávku pořídíme od 6 mluvčích, z nichž jsou 3 ženy a 3 jsou muži. Slova zaznamenaná od těchto mluvčích budeme pak porovnávat pomocí funkce dynamického borcení času DTW se slovy namluvenými mužským referenčním mluvčím a budeme hledat, které jsou si nejvíce podobné.

Nahrávky bude pořizovat v prostředí s co nejmenším šumem. Samotné nahrávání bude probíhat tak, že každý mluvčí řekne formou diskrétního diktátu (tzn. že mezi jednotlivými slovy udělá krátkou pauzu) celou větu najednou. To usnadní přesnější detekování začátku a konce jednotlivých slov, takže i následné rozpoznávání slov bude přesnější.

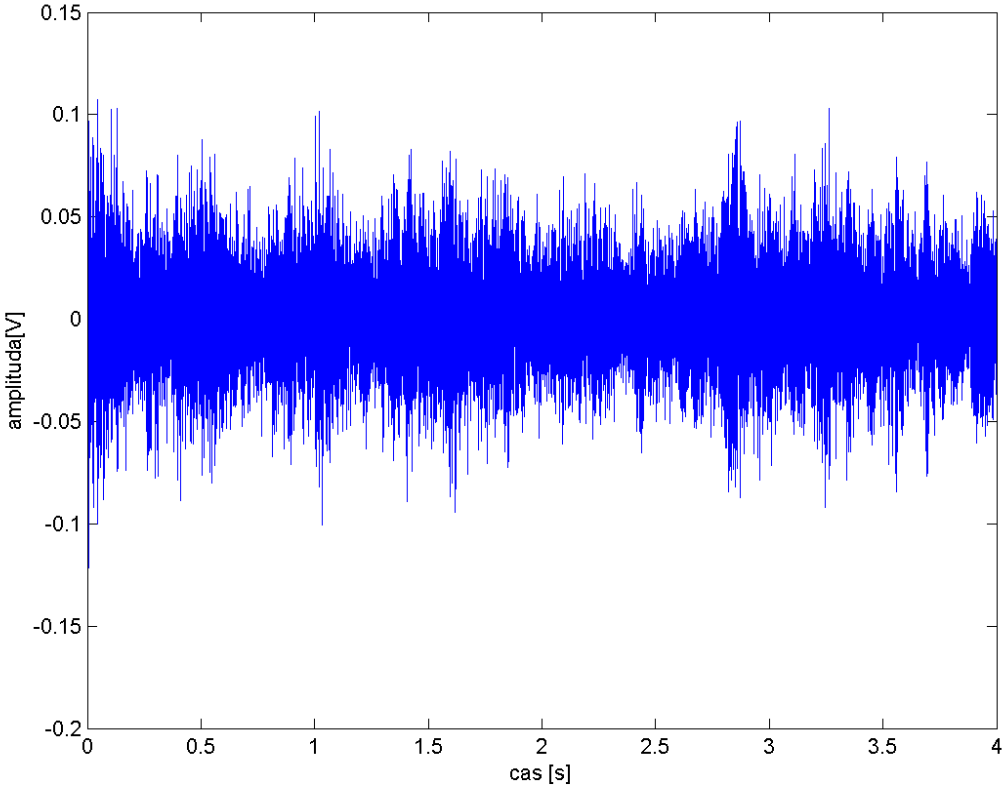


Obrázek 23 - Záznam celé věty

Tuto celou větu pak pomocí postupů popsaných ve 3. kapitole rozdělíme na jednotlivá slova a pro každé z těchto slov vypočítáme MFCC koeficienty. Prostřednictvím těchto koeficientů budeme následně mezi sebou porovnávat slova nahraná testovacími mluvčími s jednotlivými slovy od referenčního mluvčího pomocí funkce DTW a určovat, mezi kterými je největší shoda.

6.2 Přidání rušení k nahrávkám

Abychom mohli otestovat, zda a o kolik se bude snižovat rozpoznávací schopnost softwarového programu s klesajícím odstupem signálu od šumu, budeme k nahrávkám pořízenými od testovaných mluvčích postupně přidávat rušení s rostoucí intenzitou. Tím se bude snižovat odstup signálu od šumu.



Obrázek 24 – Záznam rušení

Rušení vytvoříme zaznamenáním 15 mluvčích, kteří budou najednou mluvit v uzavřené místnosti. Pro pořízení záznamu použijeme všesměrový mikrofon, který díky své směrové charakteristice snímá celý prostor místnosti rovnoměrně. Intenzitu rušení budeme nastavovat tak, že si spočítáme jeho celkový výkon a pak ho budeme v různých poměrech k výkonu čistého slova smíchávat. Pomocí tohoto poměru můžeme nastavit požadovaný odstup signálu od šumu. Vzorec pro jeho výpočet je následující:

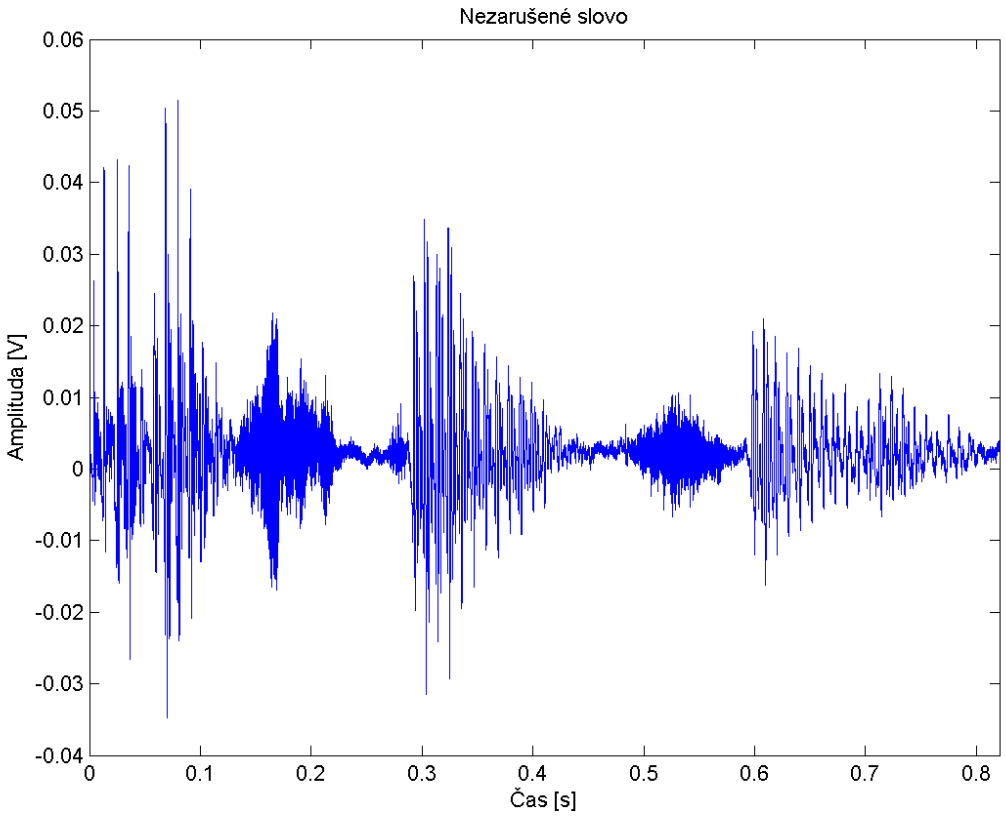
$$SNR = 10.log_{10}\left(\frac{P_s}{P_š}\right)\ [dB] \tag{6.1}$$

kde SNR je odstup signálu od šumu

Ps je výkon signálu

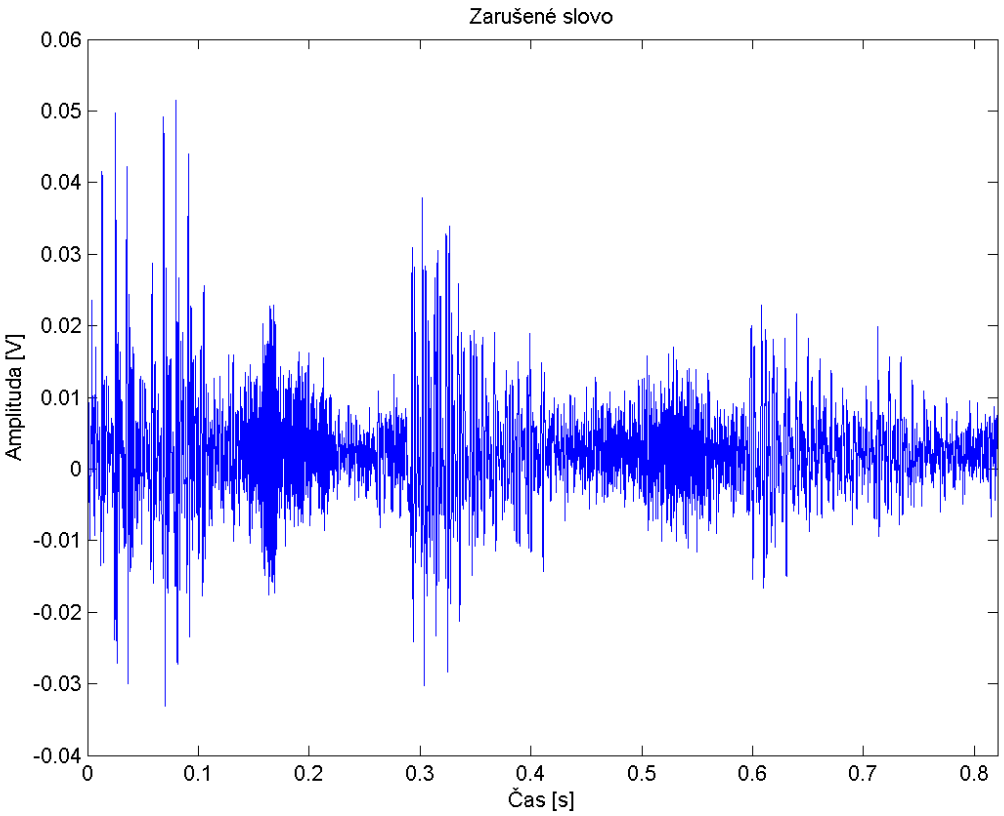
Pš je výkon šumu

V našem případě to bude v rozmezí od SNR=0 dB, což znamená, že se výkon čistého slova bude rovnat výkonu šumu, do 20 dB, kdy bude výkon šumu 100 krát menší než výkon slova.



Obrázek 25 - Nezarušené slovo "Rostoucí"

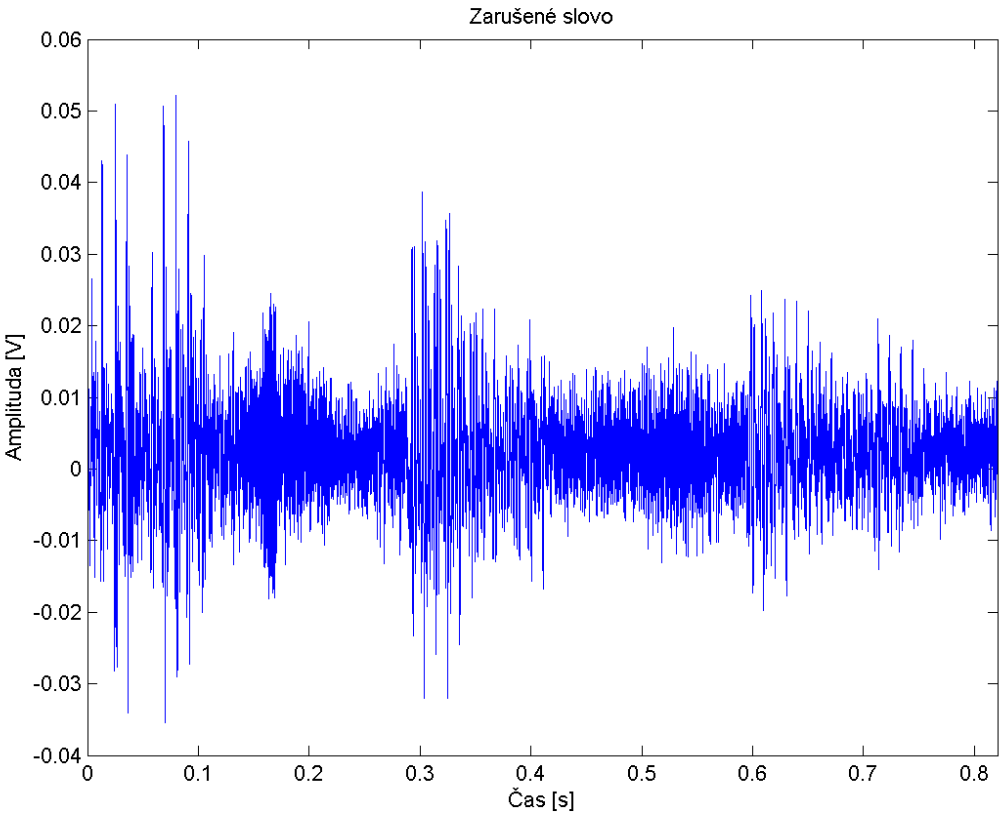
Na obrázku je znázorněno, jak vypadá nezarušené slovo “Rostoucí“, namluvené druhým mluvčím.



Obrázek 26 - Zarušené slovo "Rostoucí"

Na obrázku vidíme, jak se změní průběh nahrávky slova "Rostoucí", když přidáme rušení s odstupem SNR=3dB. To znamená, že výkon rušení se rovná polovině výkonu slova, takže zkreslení slova v důsledku rušení je značné.

Dalším krokem je, že k zarušeným nahrávkám přidáme bílý šum s odstupem SNR=10dB. Základní vlastností bílého šumu je skutečnost, že je to náhodný signál, který má rovnoměrnou spektrální hustotu v nekonečném frekvenčním rozsahu. Jedná se pouze o teoretický signál, který vytvoříme v programu matlab pomocí funkce awgn. Parametr této funkce použijeme 'measured', který zajistí, že se dodrží požadovaný odstup signálu od šumu.



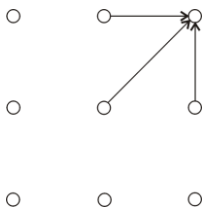
Obrázek 27 - Zarušené slovo "Rostoucí" s přidaným bílým šumem

Na obrázku je znázorněno, jak vypadá slovo “Rostoucí“ se zarušením s odstupem SNR=3 dB a s přidaným bílým šumem s odstupem SNR 10dB. Vidíme, že výsledná směs těchto 3 signálu se značně liší od původního čistého slova.

6.3 Parametry DTW

Pro rozpoznávání slov jsou použity 3 různé typy lokálního omezení funkce DTW. Matematické zápisy jednotlivých lokálních omezení vypadají takto:

- **První typ lokálního omezení**

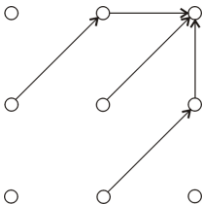


$$g(m,n) = min\begin{Bmatrix} g(n,m-1) + d(n,m) \\ g(n-1,m-1) + 2d(n,m) \\ g(n-1,m) + d(n,m) \end{Bmatrix} \tag{6.2}$$

Pro tento typ lokální funkce je vhodné použít symetrickou váhovou funkci:

$$\hat{w}(k) = [t(k) - t(k-1)] + [r(k) - r(k-1)] \tag{6.3}$$

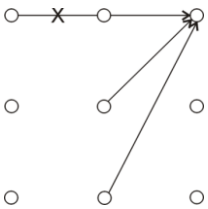
- **Druhý typ lokálního omezení**



$$g(m,n) = min\begin{Bmatrix} g(n-1,m-2) + 2d(n,m-1) + d(n,m) \\ g(n-1,m-1) + 2d(n,m) \\ g(n-2,m-1) + 2d(n-1,m) + d(n,m) \end{Bmatrix} \tag{6.4}$$

Stejně jako pro první typ lokálního omezení použijeme symetrickou váhovou funkci viz. rovnice (6.3).

- **Třetí typ lokálního omezení**



$$g = (m,n)min\begin{Bmatrix} g(n-1,m) + kd(n,m) \\ g(n-1,m-1) + d(n,m) \\ g(n-1,m-2) + d(n,m) \end{Bmatrix} \tag{6.5}$$

$$k = 1\ pro\ j(k-1) \neq j(k-2)$$

$$k = \infty\ pro\ j(k-1) = j(k-2)$$

Pro tento typ lokálního omezení použijeme asymetrickou váhovou funkci:

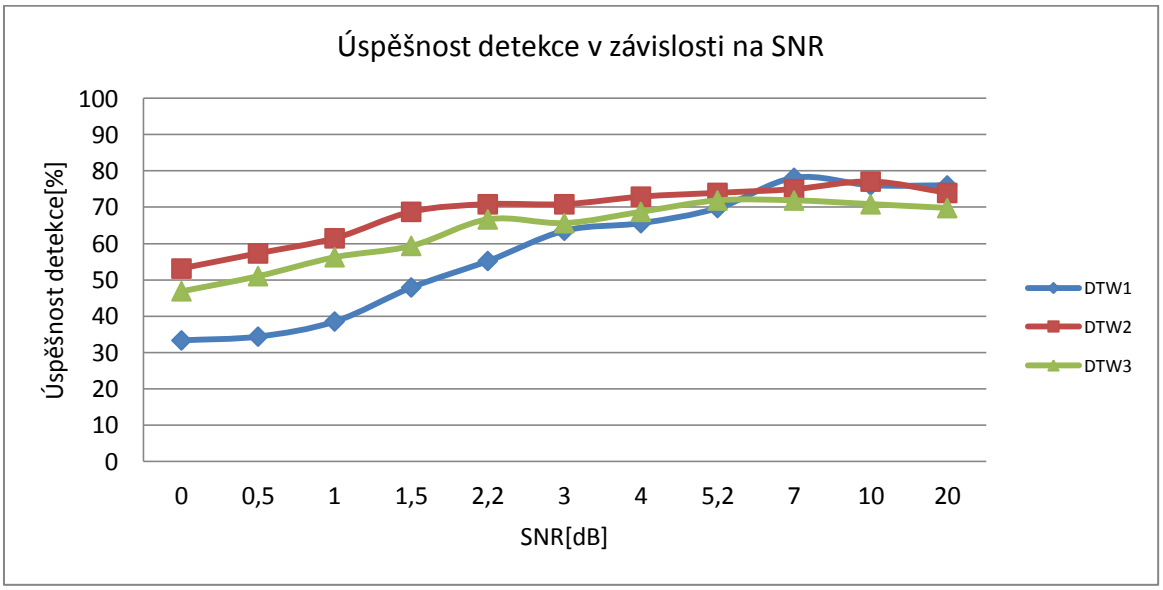
$$\hat{w}(k) = t(k) - t(k-1) \tag{6.6}$$

6.4 Vyhodnocení

Jak už bylo řečeno, vyhodnocovat budeme úspěšnost detekce slov "Rostoucí počet vloupání zvyšuje nabídku bezpečnostní techniky na trhu ne vždy se jedná o spolehlivé zařízení". Tyto slova pořídíme od referenčního mluvčího a budeme je porovnávat se stejnými slovy namluvenými 6 testovacími mluvčími a vypočítáme procentuální úspěšnost jejich detekce.

Výsledky rozpoznávání vyhodnotíme pomocí grafů a tabulek, kde bude znázorněno, jakým způsobem ovlivňuje úspěšnost rozpoznávání referenčních slov rušení, které přidáme k testovacím slovům. Tento postup provedeme pro testovací slova nahraná směrovým i všesměrovým mikrofonem. V dalším pokusu budeme postupovat obdobně, ale použijeme data jen ze směrového mikrofonu a k testovacím slovům přidáme kromě rušení ještě bílý šum s odstupem SNR=10dB. V grafech bude vykreslena procentuální úspěšnost rozpoznávání referenčních slov pro všechny 3 výše popsané typy lokálního omezení DTW. Grafy vytvoříme pro několik vybraných slov i pro průměrnou úspěšnost rozpoznávání všech testovaných slov.

6.4.1 Slova nahraná směrovým mikrofonem s přidaným rušením

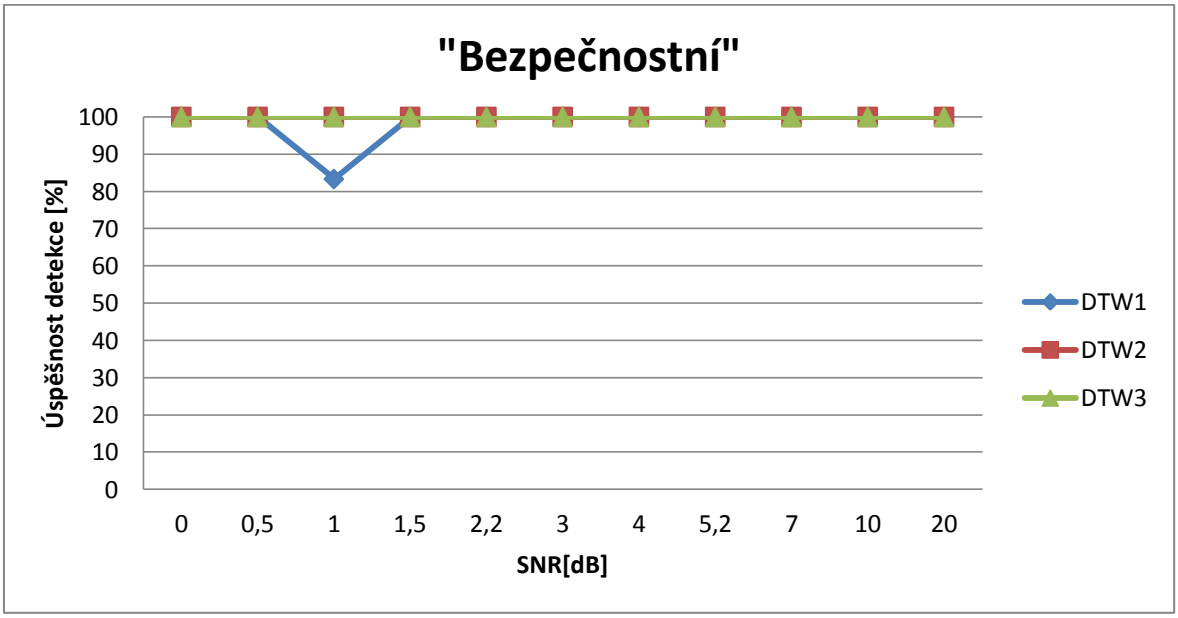


Graf 1 - Průměrná úspěšnost detekce všech slov

Tabulka 2 - Průměrná úspěšnost detekce všech slov

SNR[dB]	0	0,5	1	1,5	2,2	3	4	5,2	7	10	20
Úspěšnost DTW1 [%]	33,3	34,4	38,5	47,9	55,2	63,5	65,6	69,8	78,1	76	76
Úspěšnost DTW2 [%]	53,1	57,3	61,5	68,8	70,8	70,8	72,9	74	75	77,1	73,9
Úspěšnost DTW3 [%]	46,9	51	56,3	59,4	66,7	65,6	68,8	71,9	71,9	70,8	69,8

V grafu a tabulce je znázorněna průměrná úspěšnost rozpoznávání všech testovaných slov nahraných pomocí směrového mikrofonu v závislosti na odstupu SNR. Nejlepší úspěšnost rozpoznávání prokazuje 2. typ DTW a v průměru vychází na 68,7%. Dokonce i při SNR=0, což znamená, že je úroveň signálu stejná jako úroveň rušení, je úspěšnost detekce nad 50% hranicí. 3. DTW měl úspěšnost rozpoznání 63,5% a nejhorší úspěšnost měl 1. typ DTW, u kterého vycházela průměrná úspěšnost 58% a hlavně pro velmi zarušená slova byla úspěšnost detekce velmi nízká.

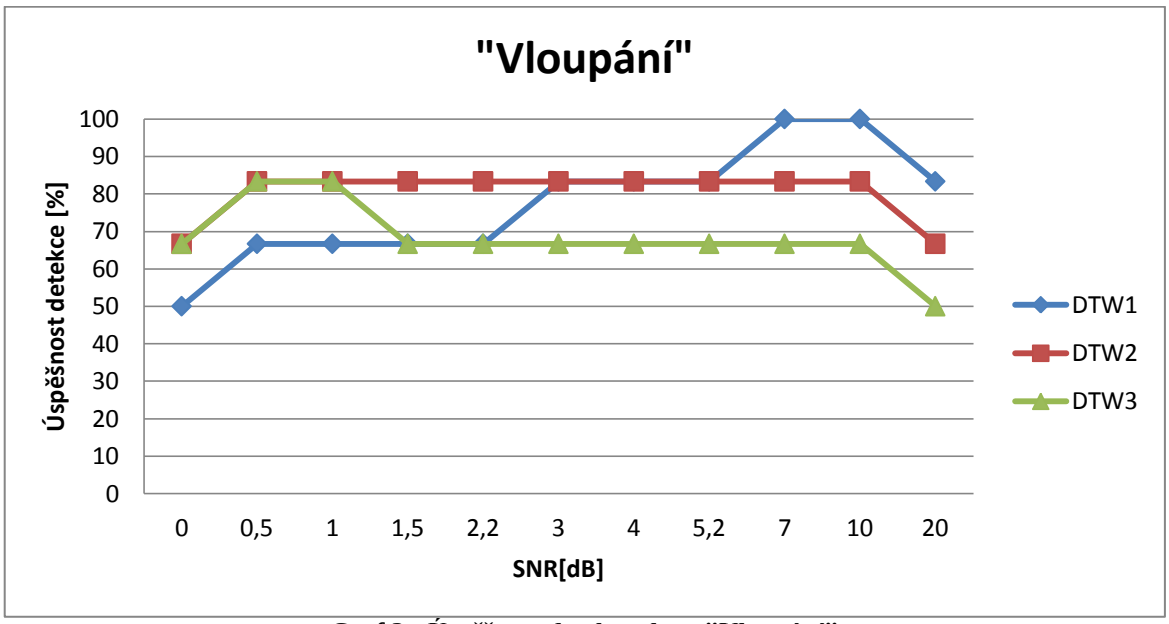


Graf 2 - Úspěšnost detekce slova "Bezpečnostní"

Tabulka 3 - Úspěšnost detekce slova "Bezpečnostní"

SNR[dB]	0	0,5	1	1,5	2,2	3	4	5,2	7	10	20
Úspěšnost DTW1 [%]	100	100	83,3	100	100	100	100	100	100	100	100
Úspěšnost DTW2 [%]	100	100	100	100	100	100	100	100	100	100	100
Úspěšnost DTW3 [%]	100	100	100	100	100	100	100	100	100	100	100

Nejlepší úspěšnost rozpoznávání ze všech slov mělo slovo “Bezpečnostní“. Vidíme, že u 2. a 3. typu omezení DTW proběhlo rozpoznávání úplně bez chyby. U 1. typu lokálního omezení DTW došlo pouze k jedné chybné detekci a to při odstupu signálu od šumu SNR= 1 dB. K záměně došlo se slovem “Zvyšuje“.



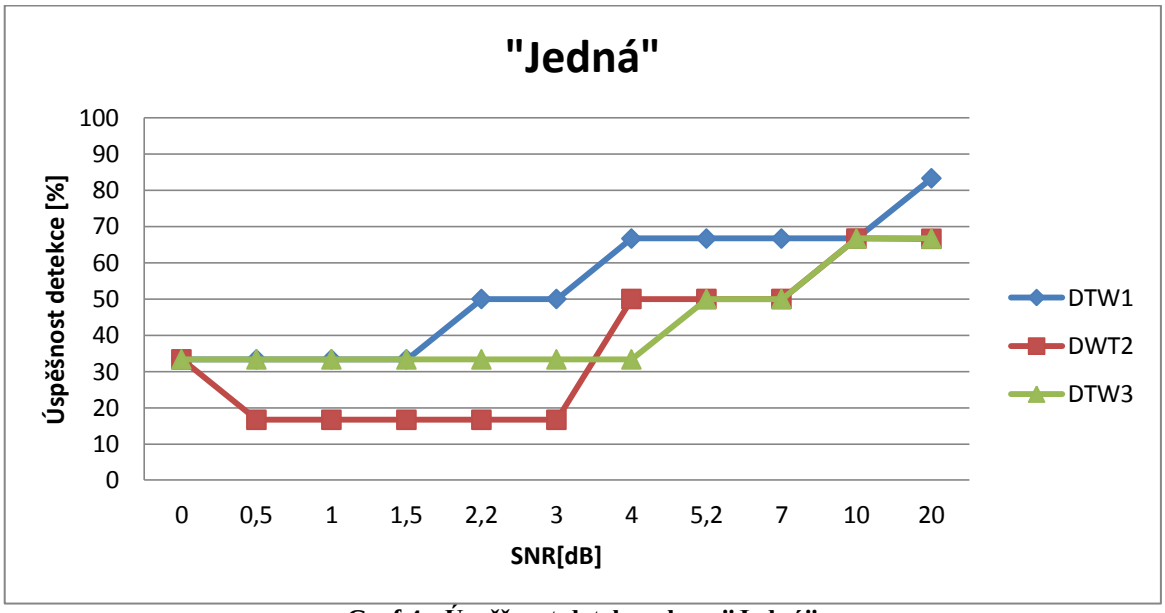
Graf 3 - Úspěšnost detekce slova "Vloupání"

Tabulka 4 - Úspěšnost detekce slova "Vloupání"

SNR[dB]	0	0,5	1	1,5	2,2	3	4	5,2	7	10	20
Úspěšnost DTW1 [%]	50	66,7	66,7	66,7	66,7	83,3	83,3	83,3	100	100	83,3
Úspěšnost DTW2 [%]	66,7	83,3	83,3	83,3	83,3	83,3	83,3	83,3	83,3	83,3	66,7
Úspěšnost DTW3 [%]	66,7	83,3	83,3	66,7	66,7	66,7	66,7	66,7	66,7	66,7	50

Průměrná úspěšnost detekce slova “Vloupání“ všech 3 typů DTW je 75,3% a ani u jednoho typu DTW neklesla úspěšnost detekce pod 50%. Nejvyšší úspěšnost detekce měl 2. typ DTW s průměrnou úspěšností 80,3%. K záměně docházelo se slovy “Nabídku“ a “Bezpečnostní“. 1. Typ DTW měl průměrnou úspěšnost detekce 77,3%. Tento typ lokálního omezení chybně detekoval slova “Nabídku“ a „Jedná“. Nejhorší úspěšnost

detekce měl 3. typ DTW s průměrnou úspěšností 68,2%. Stejně jako u obou předchozích typů lokálního omezení nastala chybná detekce se slovem “Nabídku“. Navíc bylo špatně detekováno slovo “Na“.



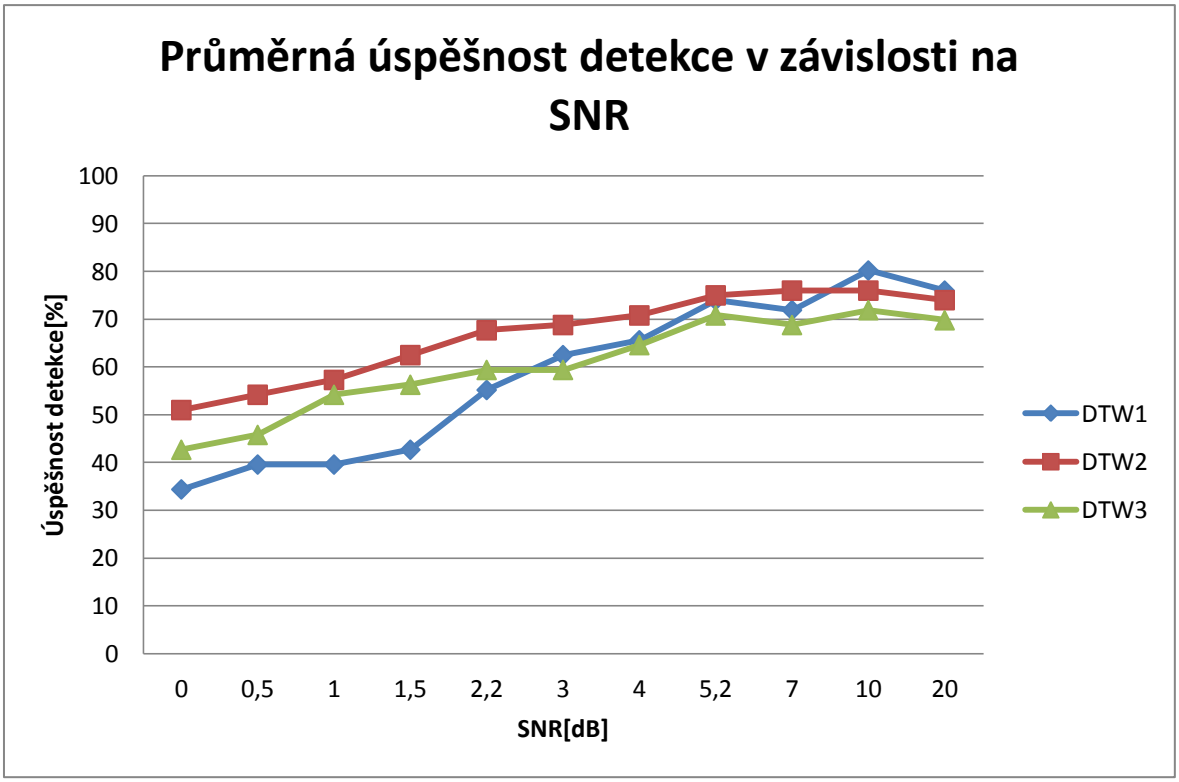
Graf 4 - Úspěšnost detekce slova "Jedná"

Tabulka 5 - Úspěšnost detekce slova "Jedná"

SNR[dB]	0	0,5	1	1,5	2,2	3	4	5,2	7	10	20
Úspěšnost DTW1 [%]	33,3	33,3	33,3	33,3	50	50	66,7	66,7	66,7	66,7	83,3
Úspěšnost DTW2 [%]	33,3	16,7	16,7	16,7	16,7	16,7	50	50	50	66,7	66,6
Úspěšnost DTW3 [%]	33,3	33,3	33,3	33,3	33,3	33,3	33,3	50	50	66,7	66,7

Nejhorší úspěšnost detekce ze všech zarušených testovaných slov nahraných směrovým mikrofonem mělo slovo “Jedná“, které mělo průměrnou úspěšnost detekce 43,9%. Nejvyšší úspěšnost detekce měl 1. typ lokálního omezení DTW s průměrnou úspěšností 53%. K záměně docházelo se slovy “Vloupání“, “Techniky“, “Na“ a “Zařízení“. 3. typ DTW měl průměrnou úspěšnost detekce 42,4%. Tento typ lokálního omezení chybně detekoval slova “ Vloupání“, “Nabídku“ a “Trhu“. Nejhorší úspěšnost detekce měl 2. typ DTW s průměrnou úspěšností 36,4%. Chybně byla detekována slova “Nabídku“, “Trhu“ a “Ne“.

6.4.2 Slova nahraná všesměrovým mikrofonem s přidaným rušením

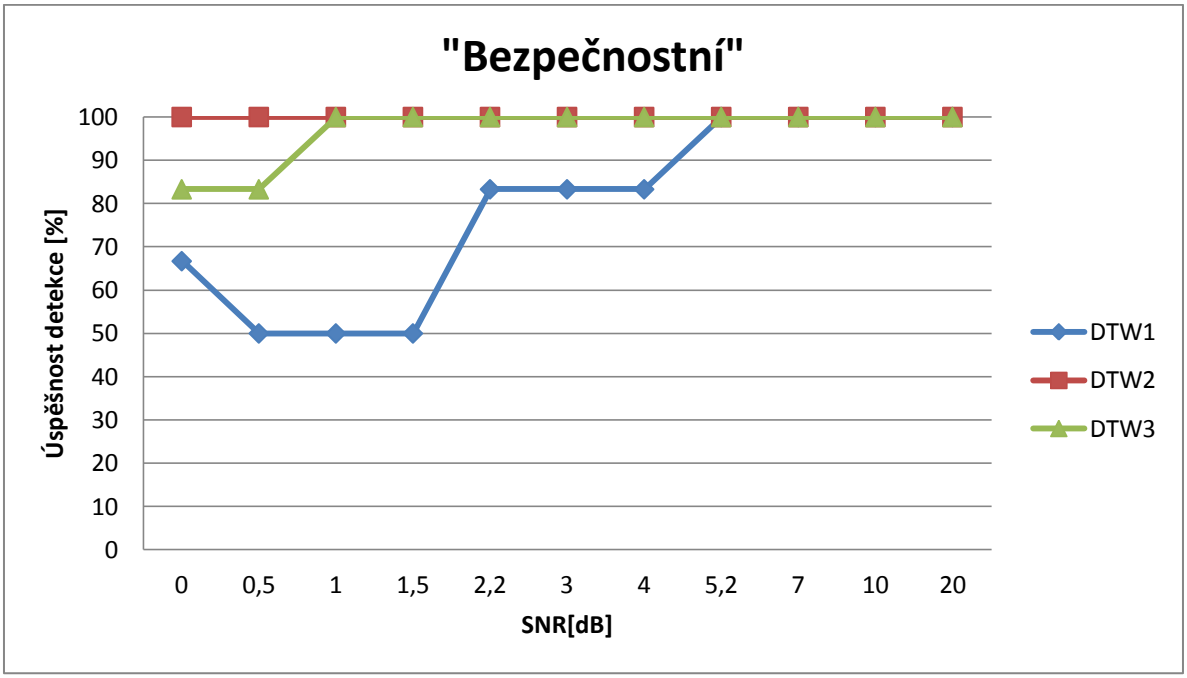


Graf 5 - Průměrná úspěšnost detekce všech slov

Tabulka 6 – Průměrná úspěšnost detekce všech slov

SNR[dB]	0	0,5	1	1,5	2,2	3	4	5,2	7	10	20
Úspěšnost DTW1 [%]	34,4	39,6	39,6	42,7	55,2	62,5	65,6	74	71,9	80,2	76
Úspěšnost DTW2 [%]	51	54,2	57,3	62,5	67,7	68,8	70,8	75	76	76	74
Úspěšnost DTW3 [%]	42,7	45,8	54,2	56,3	59,4	59,4	64,6	70,8	68,8	71,9	69,8

V grafu a tabulce je znázorněna průměrná úspěšnost rozpoznávání všech testovaných slov nahraných všesměrovým mikrofonem v závislosti na odstupu SNR. Nejlepší úspěšnost rozpoznávání prokazuje 2. typ DTW a v průměru vychází na 66,7%. 3. DTW měl úspěšnost rozpoznání 60,3% a nejhorší úspěšnost měl 1. typ DTW, u kterého vycházela průměrná úspěšnost 58,3%. V grafu můžeme pozorovat, že pro velký odstup SNR (10dB a 20dB) má nejlepší výsledky DTW1, ale pro zarušená slova (SNR <2,2dB) má nejhorší úspěšnost. Pro malý odstup SNR je nejvhodnější použít DTW2, které má i pro SNR=0 úspěšnost rozpoznávání nad 50%.

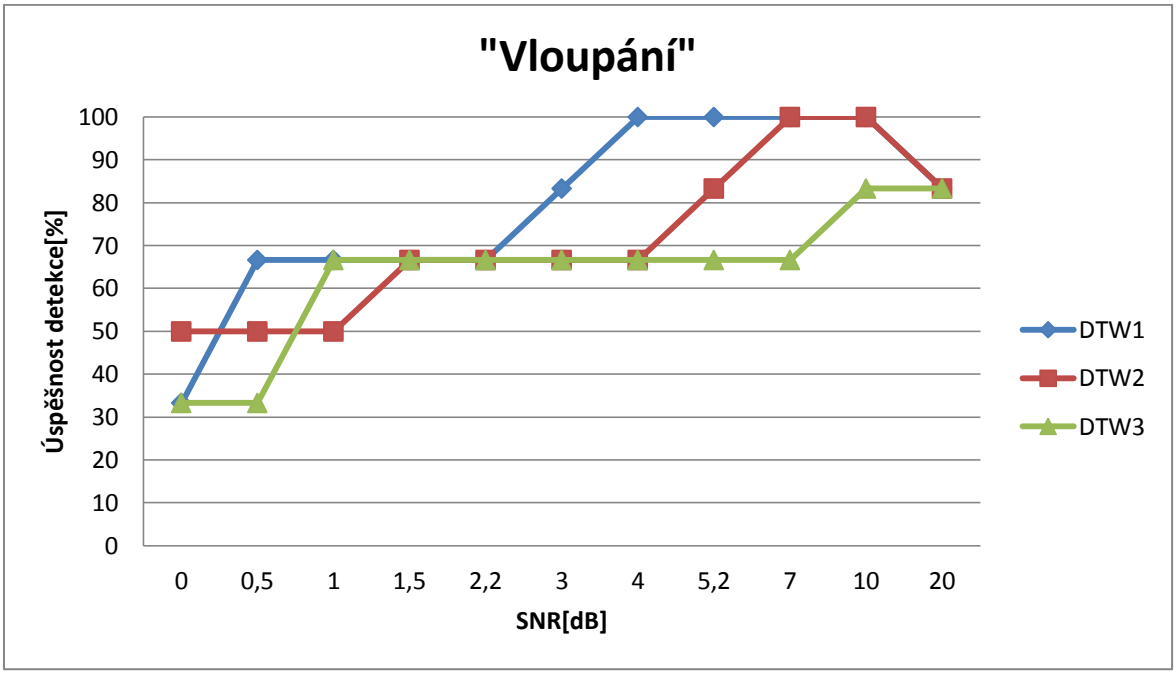


Graf 6 - Úspěšnost detekce slova "Bezpečnostní"

Tabulka 7 – Úspěšnost detekce slova "Bezpečnostní"

SNR[dB]	0	0,5	1	1,5	2,2	3	4	5,2	7	10	20
Úspěšnost DTW1 [%]	66,7	50	50	50	83,3	83,3	83,3	100	100	100	100
Úspěšnost DTW2 [%]	100	100	100	100	100	100	100	100	100	100	100
Úspěšnost DTW3 [%]	83,3	83,3	100	100	100	100	100	100	100	100	100

Nejlepší úspěšnost rozpoznávání ze všech slov mělo slovo “Bezpečnostní“. Vidíme, že u 2. typu DTW proběhlo rozpoznávání úplně bez chyby. I 3. typ DTW dosahuje výborných výsledků, pouze u 2 nejmenších odstupů SNR = 0 a 0,5 bylo chybně detekováno slovo “Rostoucí“. Nejhorší úspěšnost rozpoznávání měl 1. typ DTW s průměrnou úspěšností 78,8%. U tohoto typu lokálního omezení docházelo k záměnám se slovy “Rostoucí “ a “Zařízení“

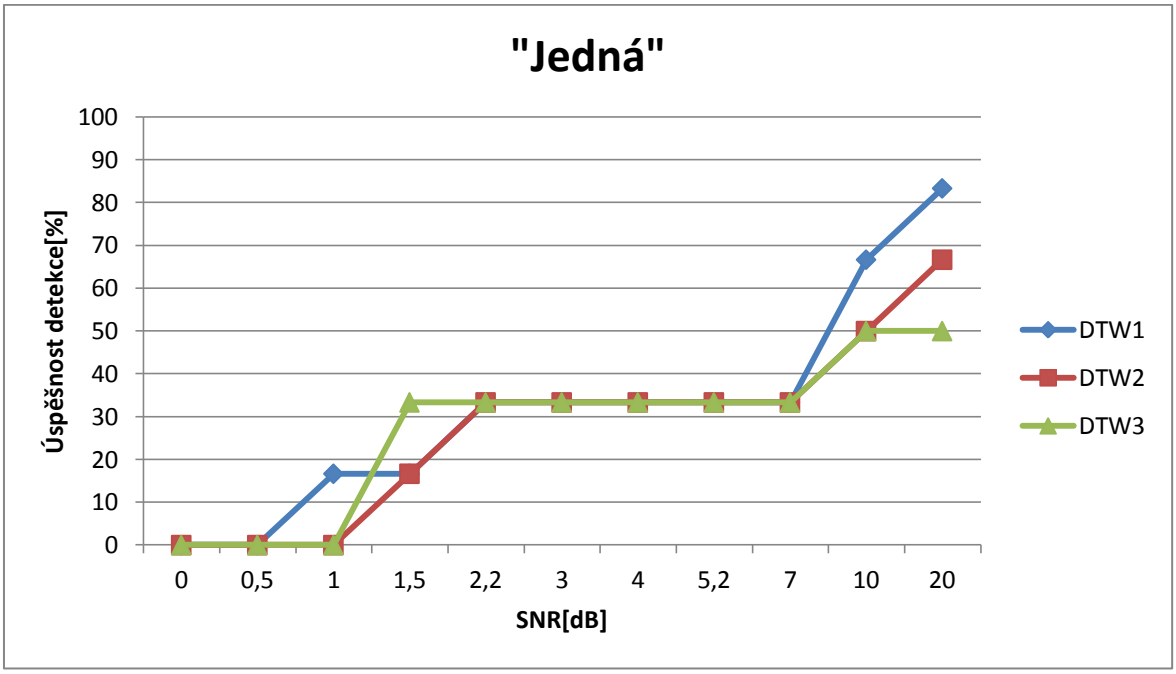


Graf 7- Úspěšnost detekce slova "Vloupání"

Tabulka 8 – Úspěšnost detekce slova "Vloupání"

SNR[dB]	0	0,5	1	1,5	2,2	3	4	5,2	7	10	20
Úspěšnost DTW1 [%]	33,3	66,7	66,7	66,7	66,7	83,3	100	100	100	100	83,3
Úspěšnost DTW2 [%]	50	50	50	66,7	66,7	66,7	66,7	83,3	100	100	83,3
Úspěšnost DTW3 [%]	33,3	33,3	66,7	66,7	66,7	66,7	66,7	66,7	66,7	83,3	83,3

Průměrná úspěšnost detekce slova "Vloupání" všech 3 typů DTW je 70%. Nejvyšší úspěšnost detekce měl 1. typ DTW s průměrnou úspěšností 78,8%. K záměně docházelo se slovy "Rostoucí", "Trhu", "Spolehlivé" a "Zařízení". 2. Typ DTW měl průměrnou úspěšnost detekce 71,2%, a u tohoto typu neklesla úspěšnost rozpoznávání pod 50% ani u nejmenšího odstupu SNR. Tento typ lokálního omezení chybně detekoval slova "Nabídku", "Trhu" a „Zařízení". Nejhorší úspěšnost detekce měl 3. typ DTW s průměrnou úspěšností 63,6%. Stejně jako u obou předchozích typů lokálního omezení nastala chybná detekce se slovem "Zařízení". Navíc byly špatně detekovány ještě slova "Nabídku", "Spolehlivé" a "Jedná".



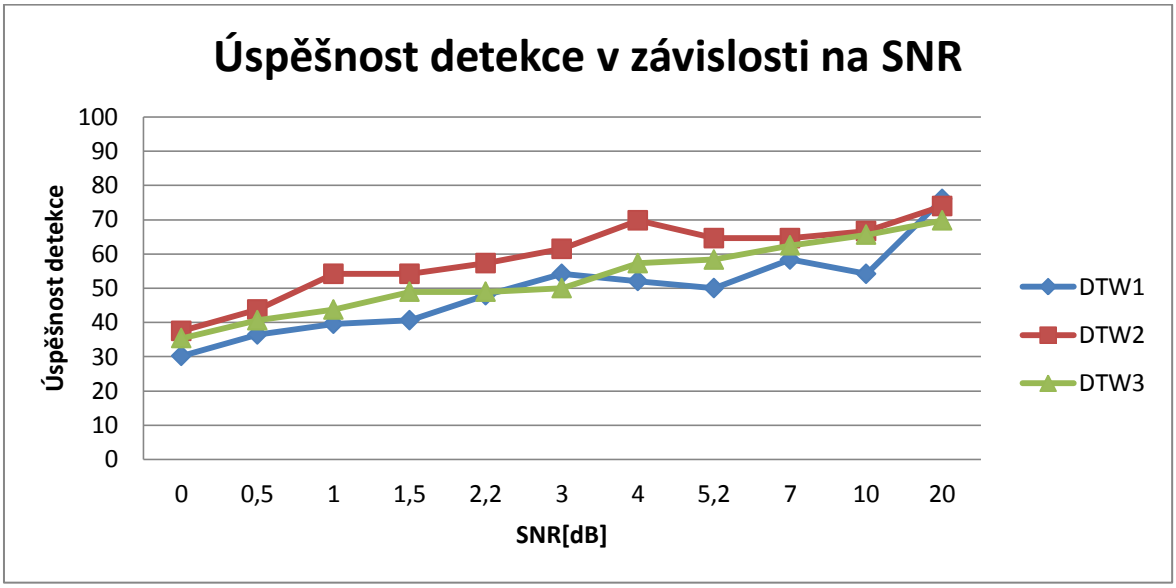
Graf 8 – Úspěšnost detekce slova "Jedná"

Tabulka 9- Úspěšnost detekce slova "Jedná"

SNR[dB]	0	0,5	1	1,5	2,2	3	4	5,2	7	10	20
Úspěšnost DTW1 [%]	0	0	16,7	16,7	33,3	33,3	33,3	33,3	33,3	66,7	83,3
Úspěšnost DTW2 [%]	0	0	0	16,7	33,3	33,3	33,3	33,3	33,3	50	66,7
Úspěšnost DTW3 [%]	0	0	0	33,3	33,3	33,3	33,3	33,3	33,3	50	50

Nejhorší úspěšnost detekce ze všech testovaných slov mělo slovo "Jedná", které mělo průměrnou úspěšnost detekce jen 28,8%. Při odstupu SNR 0 a 0,5 dB mají dokonce všechny typy DTW nulovou úspěšnost. Nejvyšší úspěšnost detekce má 1. typ DTW, s průměrnou úspěšností 31,8%. 2. a 3 typ DTW má průměrnou úspěšnost detekce shodně 28,8%.

6.4.3 Slova s přidaným rušením a AWGN

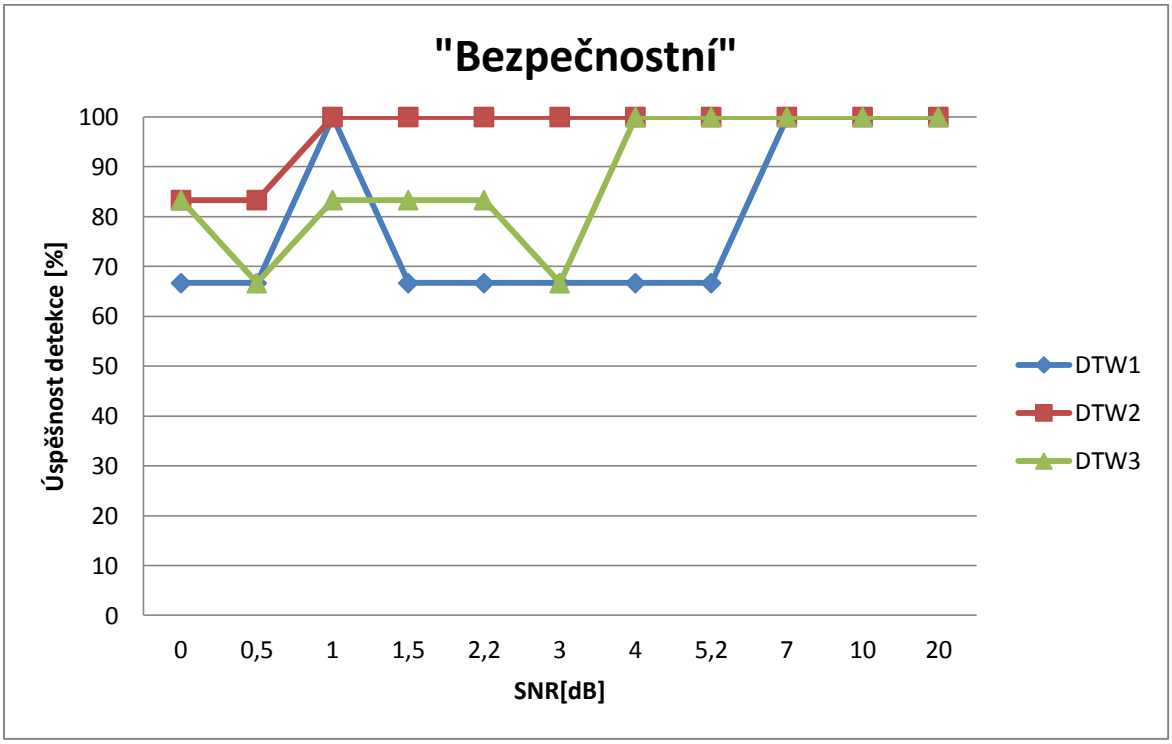


Graf 9 - Průměrná úspěšnost detekce všech slov

Tabulka 10 - Průměrná úspěšnos detekce všech slov

SNR[dB]	0	0,5	1	1,5	2,2	3	4	5,2	7	10	20
Úspěšnost DTW1 [%]	30,2	36,5	39,6	40,6	47,9	54,2	52,1	50	58,3	54,2	76
Úspěšnost DTW2 [%]	37,5	43,8	54,2	54,2	57,3	61,5	69,8	64,6	64,6	66,7	74
Úspěšnost DTW3 [%]	35,4	40,6	43,8	49	49	50	57,3	58,3	62,5	65,6	69,8

V grafu a tabulce je znázorněna průměrná úspěšnost rozpoznávání všech zarušených testovaných slov s přidaným bílým šumem v závislosti na odstupu SNR. Nejlepší úspěšnost rozpoznávání prokazuje 2. typ DTW, který má nejlepší úspěšnost detekce pro všechny hodnoty SNR kromě SNR=20dB, kde byl úspěšnější 1. typ DTW, a v průměru vychází na 58,9%. 3. typ DTW měl úspěšnost rozpoznání 52,8% a nejhorší úspěšnost měl 1. typ DTW, u kterého vycházela průměrná úspěšnost 49,1%. Takže úspěšnost detekce se zhoršila v průměru asi o 8% oproti zarušeným slovům bez bílého šumu.

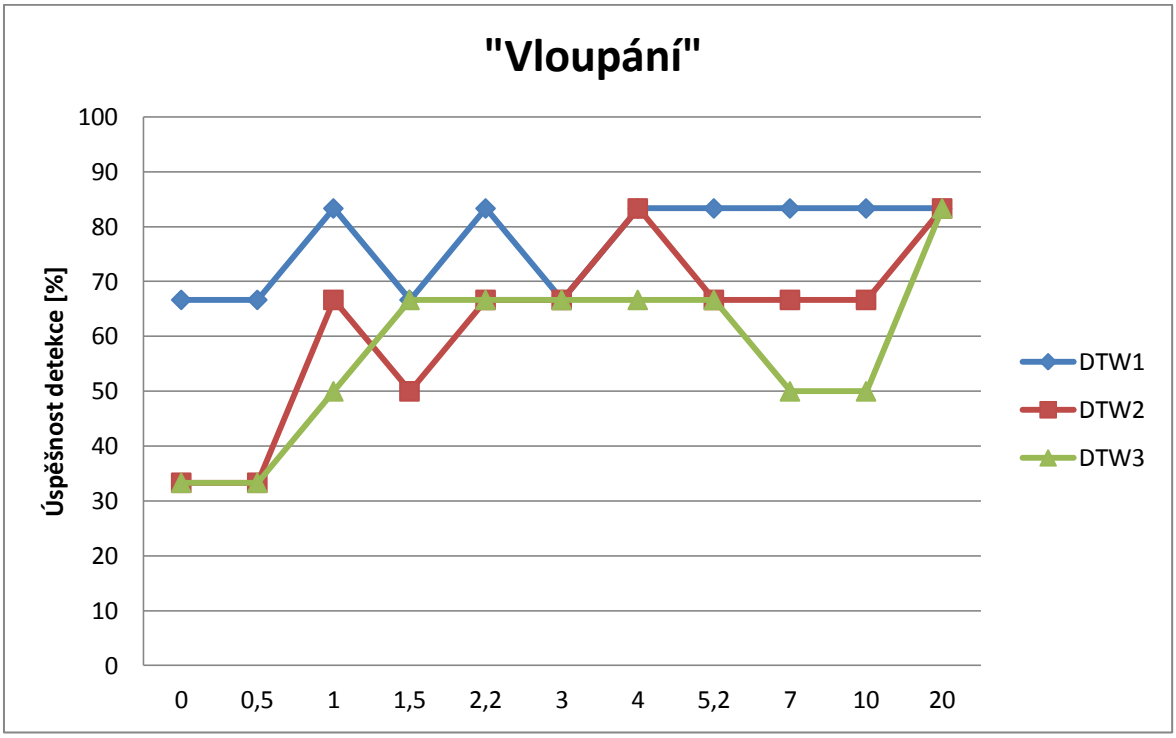


Graf 10 - Úspěšnost detekce slova "Bezpečnostní"

Tabulka 11 - Úspěšnost detekce slova "Bezpečnostní"

SNR[dB]	0	0,5	1	1,5	2,2	3	4	5,2	7	10	20
Úspěšnost DTW1 [%]	66,7	66,7	100	66,7	66,7	66,7	66,7	66,7	100	100	100
Úspěšnost DTW2 [%]	83,3	83,3	100	100	100	100	100	100	100	100	100
Úspěšnost DTW3 [%]	83,3	66,7	83,3	83,3	83,3	66,7	100	100	100	100	100

Nejlepší úspěšnost rozpoznávání ze všech slov mělo slovo “Bezpečnostní“. Vidíme, že u 2. typu DTW proběhlo rozpoznávání téměř bez chyb, pouze pro odstup 0 a 0,5dB chybně detekuje slovo “Zařízení“ a celkově má úspěšnost 97%. I 3. typ DTW dosahuje výborných výsledků a má průměrnou úspěšnost rozpoznávání 87,9%. Stejně jako u prvního typu lokálního omezení dochází k chybné detekci se slovem “Zařízení“. Nejhorší úspěšnost rozpoznávání měl 1. typ DTW s průměrnou úspěšností 78,8%. U tohoto typu docházelo k záměně se slovem “Rostoucí“.

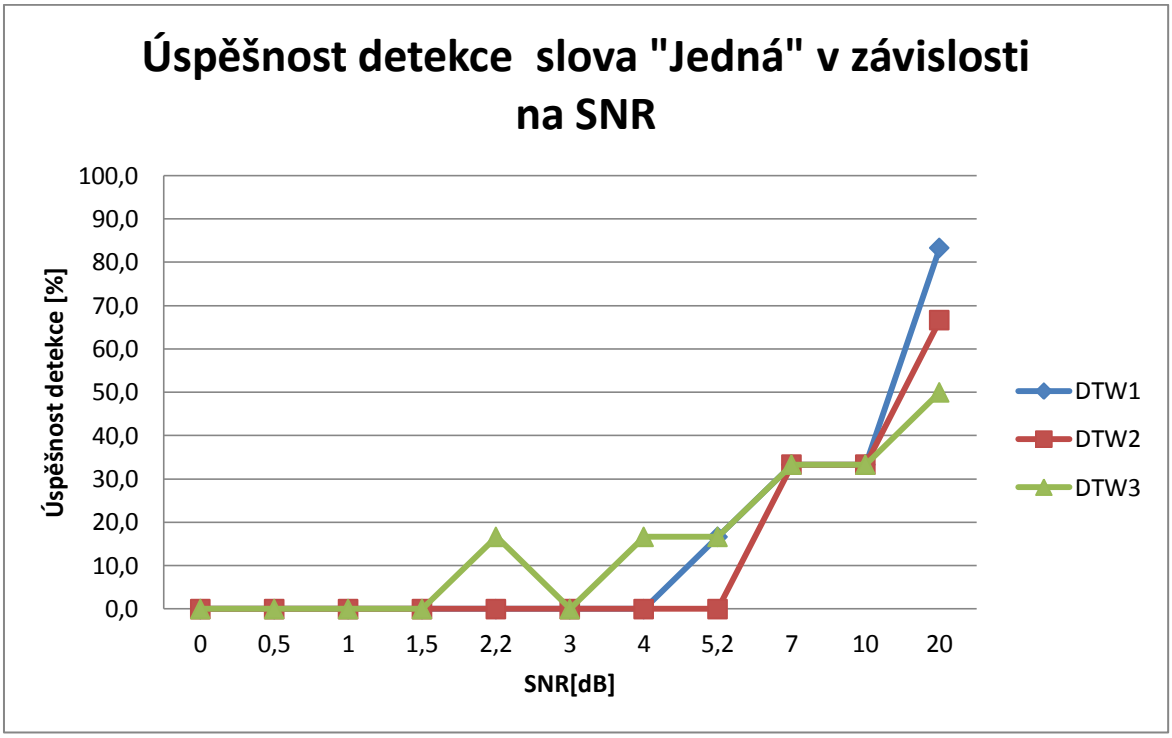


Graf 11 - Úspěšnost detekce slova "Vloupání"

Tabulka 12 - Úspěšnost detekce slova "Vloupání"

SNR[dB]	0	0,5	1	1,5	2,2	3	4	5,2	7	10	20
Úspěšnost DTW1 [%]	66,7	66,7	83,3	66,7	83,3	66,7	83,3	83,3	83,3	83,3	83,3
Úspěšnost DTW2 [%]	33,3	33,3	66,7	50	66,7	66,7	83,3	66,7	66,7	66,7	83,3
Úspěšnost DTW3 [%]	33,3	33,3	50	66,7	66,7	66,7	66,7	66,7	50	50	83,3

Průměrná úspěšnost detekce slova “Vloupání“ všech 3 typů DTW je 65,7%. Nejvyšší úspěšnost detekce měl tentokrát 1. typ DTW s průměrnou úspěšností 77,3% a oproti detekci slov bez přidaného bílého šumu klesla úspěšnost jen o 1,5%. Chybně byly detekovány slova “Trhu“ a “Zařízení“. 2. typ DTW měl průměrnou úspěšnost detekce 62,1%. Stejně jako u 1. typu došlo k záměně se slovy “Zařízení“ a “Trhu“ a k tomu ještě přibylo slovo “Nabídku“. Nejhorší úspěšnost detekce měl 3. typ DTW s průměrnou úspěšností 57,6% . Chybně byla detekována slova “Nabídku“ a “Trhu“.



Graf 12 - Úspěšnost detekce slova "Jedná"

Tabulka 13 - Úspěšnost detekce slova "Jedná"

SNR[dB]	0	0,5	1	1,5	2,2	3	4	5,2	7	10	20
Úspěšnost DTW1 [%]	0	0	0	0	0	0	0	16,7	33,3	33,3	83,3
Úspěšnost DTW2 [%]	0	0	0	0	0	0	0	0	33,3	33,3	66,7
Úspěšnost DTW3 [%]	0	0	0	0	16,7	0	16,7	16,7	33,3	33,3	50

Nejhorší úspěšnost detekce ze všech testovaných slov mělo slovo “Jedná“, které mělo průměrnou úspěšnost detekce jen 14,4%. Při odstupu SNR 0 až 1,5 dB mají dokonce všechny typy DTW nulovou úspěšnost. Všechny typy DTW mají velice podobnou úspěšnost rozpoznávání, ale nejvyšší úspěšnost detekce mají 1. a 3. typ DTW, se stejnou průměrnou úspěšností 15,2% a 3 typ DTW má průměrnou úspěšnost detekce 12,1%.

Závěr

Cílem této diplomové práce bylo analyzovat vliv různých typů rušení na detekci slov. Pro rozpoznávání slov byly použity 3 různé typy lokálního omezení metody borcení časové osy DTW. Zpracování signálů stejně jako rozpoznávání jednotlivých slov bylo prováděno prostřednictvím softwarového programu Matlab.

V teoretické části jsou popsány některé typy zvuku a jeho základní vlastnosti. Dále je popsán způsob, jakým lze pořídit nahrávky od jednotlivých mluvčích, a jsou uvedeny jednotlivé části nahrávacího řetězce. Další kapitola je věnována zpracování akustického signálu a jeho převodu do Mel kepstrálních koeficientů, které slouží jako vstup pro rozpoznávání slov pomocí metody borcení časové osy. Na závěr teoretické části jsou popsány konkrétní příklady aplikací, které pro svoji funkci používají rozpoznávání slov.

Databázi pro rozpoznávání tvoří 16 slov, která jsou vyslovována jednotlivými mluvčími ve formě diskrétního diktátu a neobsahují žádné spřažené spojky. Tato opatření usnadňují detekci jednotlivých slov.

Jako první jsme analyzovali slova zaznamenaná směrovým mikrofonem. Ta jsou zatížena rušením, které jsme získali nahráním náhodné promluvy 15 mluvčích, kteří mluví najednou v uzavřené místnosti. Nejlepší úspěšnost detekce prokazoval 2. typ lokálního omezení DTW s průměrnou úspěšností detekce 68,7%. 3. typ DTW rozpoznával slova s průměrnou úspěšností 63,5% a nejhůře dopadl 1. typ DTW, který detekoval slova s úspěšností 58%. Z průběhu úspěšnosti detekce vidíme, že rušení nejhůře působí na základní 1. typ lokálního omezení. Hlavně při odstupu SNR< 3 dB jsou výsledky minimálně o 10% horší než u 2. a 3. typu DTW.

Druhá analýza se věnovala slovům, která byla nahrána všesměrovým mikrofonem a která jsou zatížena stejným rušením jako v prvním případě. Stejně jako u slov nahraných za pomocí směrového mikrofonu je nejúspěšnější detekce pomocí 2. typu lokálního omezení DTW a nejméně úspěšný je 1. typ DTW. V porovnání se slovy nahranými směrovým mikrofonem je úspěšnost detekce nižší, ale rozdíl není nijak markantní a dosahuje zhruba 2%, což není ve vztahu k počtu provedených pokusů zcela zřejmý přínos použití směrového mikrofonu. Realizované pokusy však probíhaly tak, že testovaný mluvčí i mluvčí tvořících rušivé pozadí byli před mikrofonem

Třetí analýza probíhala tak, že jsme kromě rušení použitého v předchozích 2 případech zatížili slova ještě bílým šumem s odstupem SNR= 10 dB. Pro nahrávání byl použit směrový mikrofon. I v tomto případě slova nejpřesněji detekoval 2. typ DTW a

nejhůře 1. typ lokálního omezení DTW. Přítomnost bílého šumu zhoršila kvalitu detekce u všech typů DTW asi o 8%.

Při porovnání těchto tří analýz vidíme, že druh mikrofonu, kterým byly nahrávky pořízeny, nehraje (při konfiguraci provedených pokusů) pro úspěšnost detekce významnou roli. Je to způsobeno tím, že kvůli usnadnění detekce začátku a konce slova přidáváme rušení do nahrávek jen "uměle" pomocí programu Matlab, takže se neprojeví směrové charakteristiky mikrofonů. Nejlepší úspěšnost detekce bez ohledu na použitý typ rušení vykazoval 2. typ lokálního omezení DTW. Naopak nejhůře dopadlo rozpoznávání pro základní 1. typ omezení. Ze slov mělo největší úspěšnost detekce slovo "Bezpečnostní" a nejvíce chybných detekcí vykazovalo slovo "Jedná".

Literatura

ČERNOCKÝ, Jan. 2006. [Online] 6. Prosinec 2006. [Citace: 3. Červen 2014.] http://www.fit.vutbr.cz/study/courses/ZRE/public/opora/zre_opora.pdf.

HAJKO, Vladimír a DANIEL-SZABO, Juraj. 1973. *Základy fyziky.* Bratislava; : VEDA, 1973.

HORÁK, Zdeněk. Mikrofony. [Online] [Citace: 18. Duben 2013.] http://www.horakzdenek.cz/element_mikrofony.php.

KOČÍ, Miloslav. 2010. Rozpoznávání slov disktrétního diktátu. *Diplomová práce.* Pardubice : Univerzita Pardubice, 2010.

PSUTKA, Josef. 1995. *Komunikace s počítačem mluvenou řečí.* Praha : Academia, 1995. 80-200-0203-0.

PSUTKA, Josef, a další. 2006. *Mluvíme s počítačem česky.* Praha : Academia, 2006. 80-200-1309-1.

SMETANA, Ctirad. 1998. *Hluk a vibrace. Měření a hodnocení.* Praha : Sdělovací technika, 1998. 80-90 1936-2-5.

SpeechTech. SpeechTech ASR - rozpoznávání řeči. *SpeechTechnology.* [Online] [Citace: 8. Srpen 2014.] http://www.speechtech.cz/cs/produkty/rozpoznavani-reci.html.

SVOBODA, Emanuel, a další. 2006. *Přehled středoškolské fyziky.* Praha : Prometheus, 2006. 80-7196-307-0.

ŠIMON, Martin a Jan, POSPÍŠIL. 2012. Dragon Dictation: vyzkoušeli jsme rozpoznávání českého hlasu (iOS). *mobilenet.cz.* [Online] 12. Květen 2012. [Citace: 8. Srpen 2014.] http://mobilenet.cz/clanky/dragon-dictation-vyzkouseli-jsme-rozpoznavani-ceskeho-hlasu-ios-9197?desktop.

VOPIČKA, Josef. 2002. Vyhledávání klíčových elementů v souvislé promluvě. *Disertační práce.* Praha : autor neznámý, 18. Září 2002.

Příloha A – Tabulka výsledků rozpoznávání slova "Vloupání" s úrovní rušení SNR=3dB pro 2.typ lokálního omezení DTW

2. typ omezení DTW	rostoucí	počet	vloupání	zvyšuje	nabídku	bezpečnostní	techniky	na	trhu	ne	vždy	se	jedná	o	spolehlivé	zařízení
mluvčí 1	16,84	20,60	14,89	17,10	16,81	17,09	17,39	100	21,5	100	19,25	20,8	18,8	21,0	17,34	16,39
mluvčí 2	17,24	18,66	16,24	17,18	17,73	16,89	18,13	19,5	17,4	19,8	18,42	20,1	16,7	100	16,30	17,84
mluvčí 3	17,44	18,81	15,47	19,38	18,23	18,78	18	20,3	18,4	21,9	20,90	21,3	16,6	19,0	18,80	17,61
mluvčí 4	16,97	19,37	15,97	19,45	19,05	19,60	19,27	20	19	19,1	19,41	20,7	17,6	21,4	19,32	17,92
mluvčí 5	16,48	19,26	17,60	19,20	18,49	100	19,61	18,3	18,6	19,5	20,49	20,0	17,2	19,9	19,21	18,48
mluvčí 6	16,46	18,92	16	19,46	18,15	18,47	19,12	100	19,9	20,9	21,30	21	17,6	19,6	18,39	18,40
Ref. slovo detekováno[%]	16,66	0	83,33	0	0	0	0	0	0	0	0	0	0	0	0	0

Příloha B – Tabulka výsledků rozpoznávání slova “Techniky“ s úrovní rušení SNR=3dB pro 1. typ lokálního omezení DTW

1. typ omezení DTW	rostoucí	počet	vloupání	zvyšuje	nabídku	bezpečnostní	techniky	na	trhu	ne	vždy	se	jedná	o	spolehlivé	zařízení
mluvčí 1	15,79	15,76	16,78	16,53	16,26	15,90	15,40	17,5	16,0	16,9	16,6	19,1	17,54	17,2	17,70	16,32
mluvčí 2	16,30	18,44	16,72	16,59	16,09	16,49	15,79	17,7	16,5	17,8	17,2	18,5	17,05	100	15,85	16,79
mluvčí 3	16,76	17,98	17,04	16,34	16,29	100	15,94	17,6	16,5	16,5	17,5	19	17,04	17,6	17,61	15,47
mluvčí 4	100	17,39	100	16,72	17,11	100	15,03	19,7	17,1	17,6	15,9	18,4	17,24	16,8	100	16,29
mluvčí 5	17,04	17,95	100	17,19	16,50	100	16,07	18,6	17,4	18,8	17,7	17,9	17,34	19	17,45	16,38
mluvčí 6	17,16	17,67	17,46	17	16,49	100	17,18	18,5	17,1	17,2	18,2	19,6	16,98	18,6	18,21	16,75
Ref. slovo detekováno[%]	0	0	0	0	16,67	0	66,67	0	0	0	0	0	0	0	0	16,67