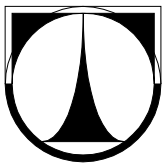
TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky a mezioborových inženýrských studií



BAKALÁŘSKÁ PRÁCE

Liberec 2008 **Jan Pražák**

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky a mezioborových inženýrských studií

Studijní program: B2612 – Elektrotechnika a informatika
Studijní obor: 1802R022 – Informatika a logistika

Návrh, tvorba a analýza řečového korpusu telefonních nahrávek pro úlohu rozpoznávání řeči a mluvčích

Suggestion, creation and analysis of speech corpus of telephonic records for speech and speaker recognition task

Bakalářská práce

Autor: **Jan Pražák**
Vedoucí práce: Ing. Jan Silovský
Konzultant:

V Liberci 16. 5. 2008

Originál zadání práce

Prohlášení

Byl(a) jsem seznámen(a) s tím, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 o právu autorském, zejména § 60 (školní dílo).

Beru na vědomí, že TUL má právo na uzavření licenční smlouvy o užití mé bakalářské práce a prohlašuji, že **s o u h l a s í m** s případným užitím mé bakalářské práce (prodej, zapůjčení apod.).

Jsem si vědom(a) toho, že užít své bakalářské práce či poskytnout licenci k jejímu využití mohu jen se souhlasem TUL, která má právo ode mne požadovat přiměřený příspěvek na úhradu nákladů, vynaložených univerzitou na vytvoření díla (až do jejich skutečné výše).

Bakalářskou práci jsem vypracoval(a) samostatně s použitím uvedené literatury a na základě konzultací s vedoucím bakalářské práce a konzultantem.

Datum

Podpis

Poděkování

Děkuji Ing. Janu Silovskému za odborné vedení při tvorbě této práce. Dále děkuji svým rodičům a přátelům za podporu během bakalářského studia.

Abstrakt

Tato práce se zabývá zejména tvorbou a analýzou konkrétního řečového korpusu telefonních nahrávek. V první části se věnuje úvodu do počítačového zpracování řeči, nastiňuje účel tvorby korpusu a cíle práce. Ve druhé části se věnuje především teoretickému popisu tvorby korpusu a některým parametrům používaných při analýze řečového korpusu. Třetí část práce se zabývá popisem tvorby vytvořeného řečového korpusu. Konkrétně pak zejména vlastnímu pořizování a přepisování nahrávek a dále také softwarové podpoře Uvádí také nabyté zkušenosti při vytváření korpusu. Čtvrtá část práce se pak zabývá analýzou vytvořeného korpusu. Ta představuje analýzu signálů a analýzu fonetické bohatosti. Analýza signálů se zaměřuje na koeficient přebuzení, koeficient vybuzení a odhad odstupu signálu od šumu. Právě v souvislosti s odhadem odstupu signálu od šumu se věnuje také vývoji detektoru řečové aktivity. Analýza fonetické bohatosti se zaměřuje na porovnání výskytu fonémů v korpusu a v českém jazyce. V páté části se práce zaměřuje na přínos vytvoření korpusu pro účely Laboratoře počítačového zpracování řeči. V poslední šesté části shrnuje a diskutuje dosažené výsledky.

Klíčová slova: řečový korpus, analýza signálů, telefonní nahrávky, detektor řečové aktivity

Abstract

This work mostly deals with creation and analysis of concrete speech corpus of telephonic records. In the first part attends to introduction to computer speech processing, describes purpose of corpus creation and targets of the work. In the second part mainly attends to theoretic description of corpus creation and to some parameters used in speech corpus analysis. The third part of the work deals with description of speech corpus creation. Concretely to own records gathering and transcribing mostly and further to software support too. It gives experience gained by corpus creation too. The forth part of the work deals with created corpus analysis. It represents signal analysis and analysis of phonetic richness. Signal analysis is focusing on overexcitation coefficient, energization coefficient and signal-to-noise ratio estimation. Rightly in connection with signal-to-noise ratio estimation attends to making of voice activity detector too. Analysis of phonetic richness is focusing on appearance of fonems in corpus comparison and in the Czech language. In the fifth part work attends to contribution of created corpus for purposes of Laboratory of computer speech processing. In the last sixth part summarises and discusses achieved results.

Keywords: speech corpus, signal analysis, telephonic records, voice activity detector

Obsah:

1. Úvod

Počítačové zpracování řeči zaznamenává v poslední době prudký rozvoj a řečové technologie patří k nejdynamičtěji se rozvíjejícím oborům v oblasti umělé inteligence. Umožňují zautomatizovat řadu činností, které až dosud vyžadovaly přítomnost člověka a jeho intelektu.

Systémy s hlasovým výstupem, jako automatické ohlašování stanic v dopravních prostředcích, ohlašování informací o spojích na vlakových nádražích či některé druhy telefonních služeb, jsou dnes již běžnou součástí našeho života.

Vznikají již ale i systémy pracující s hlasovým vstupem, jejichž tvorba je mnohem složitější. Rozpoznávací systémy Laboratoře počítačového zpracování řeči TUL dosahují v současné době v případě rozpoznávání izolovaných slov (slovních příkazů) od libovolného mluvčího úspěšnosti 90% [SpeechLabTUL2008]. Jedná se přitom o rozpoznávání v reálném čase z rozsáhlých slovníků obsahujících řádově desítky tisíc slov. Představitelem takovéhoto systému je např. program MyVoice, vyvinutý zejména pro účely umožnění handicapovaným lidem ovládat počítač pomocí hlasových příkazů. Ještě o poznání složitější je ale rozpoznávání souvislé (spojité) řeči. Úspěšnost těchto rozpoznávacích systémů Laboratoře počítačového zpracování řeči TUL se pohybuje okolo 70% [SpeechLabTUL2008]. Nejpracnější a zároveň nejdražší fází vývoje systémů rozpoznávání řeči je získání kvalitních anotovaných dat (řečového korpusu), na kterých se vyvíjený systém učí (trénuje) a testuje.

Prioritní motivací pro vytvoření řečového korpusu telefonních nahrávek TULTEL07 byl výzkumný projekt, na kterém začala Laboratoř počítačového zpracování řeči TUL pracovat. Ta na daném projektu spolupracuje s laboratořemi obdobného zaměření vysokých škol VUT Brno a ZČU Plzeň.

Moje práce měla dva hlavní cíle. Prvním cílem bylo pomoci vytvořit tento korpus především formou vlastního pořizování nahrávek a jejich anotace. Druhým cílem pak bylo analyzovat vytvořený korpus pro účely jeho budoucího využití. To vše za softwarové podpory obou těchto cílů. Jako vedlejší cíl bych pak označil vyhodnocení automatického rozpoznávání řeči na vytvořeném korpusu a vyjádření přínosu korpusu pro Laboratoř počítačového zpracování řeči TUL.

2. Obecný popis návrhu, tvorby a analýzy řečového korpusu

Za řečový korpus lze považovat jakýkoli soubor řečových záznamů, jehož nedílnou součástí je anotace a dokumentace dostatečná k tomu, aby řečová data bylo možné znovu využít v budoucnu.

Vývoj řečového korpusu lze rozdělit do tří fází. V přípravné fázi je třeba jasně definovat, k čemu je korpus určen a podle toho stanovit řečový obsah korpusu, počet a typ řečníků, jejichž hlasy mají být v korpusu zaznamenány, a podmínky, za kterých bude korpus nahráván. K vlastnímu nahrávání řečového signálu dochází v průběhu záznamové fáze. Během dokončovací fáze se provádí transkripce zaznamenaného signálu, vytváří se slovník a pořizuje se podrobná dokumentace celého korpusu [Radová2004].

2.1 Přípravná fáze tvorby řečového korpusu

V přípravné fázi je třeba stanovit typ řeči, který bude korpus obsahovat, dále pak počet a typ řečníků, jejichž hlasy budou v korpusu zaznamenány, a také nahrávací podmínky.

2.1.1 Nábor mluvčích

Mezi nejpoužívanější strategie náboru mluvčích při vytváření řečových korpusů patří [Pollák2002]:

Lavinový nábor (snow-ball recruitement) - tato metoda vychází z předpokladu, že jednotliví mluvčí zprostředkují kontakty na další potencionální mluvčí

Pevně řízený nábor - tento způsob náboru spočívá v tom, že se vygenerují formuláře s pevně danými charakteristikami mluvčích (pohlaví, věková skupina, dialekt), které je nutno následně sehnat

Částečně řízený lavinový nábor - spočívá v náboru organizátorů náboru menšího počtu mluvčích

Strategie částečně řízeného lavinového náboru byla použita mj. pro nábor mluvčích do českých telefonních databází SpeechDat (1052 mluvčích) a ČÍSLOVKY (1227 mluvčích).

2.2 Záznamová fáze tvorby řečového korpusu

Do záznamové fáze patří samotné nahrávání řečového signálu, ale také zaznamenání potřebných informací o samotných nahrávkách. Je třeba myslet na to, že vytvářený korpus může posloužit v budoucnu i pro jiné účely, než pro které je bezprostředně vytvářen. Mezi informace, které by při nahrávání řečového korpusu neměly o nahrávkách chybět, patří: datum a čas pořízení, informace o mluvčím (nejčastěji pohlaví, věk a dialekt) a v případě telefonního korpusu ještě parametry telefonu, ze kterého se nahrávka uskutečnila.

2.3 Dokončovací fáze tvorby řečového korpusu

Během dokončovací fáze se vytváří vhodná textová reprezentace nahraných řečových dat. Tato reprezentace se označuje pojmem anotace. Anotace se nevytváří pro jednotlivé nahrávky jako celky, ale po jednotlivých úsecích nahrávek (segmentech), na které jsou nahrávky rozděleny. Velikost těchto segmentů odpovídá typicky větám nebo slovům. Význam segmentace v přepisech spočívá v synchronizaci přepisu a signálu. Důvodem pro vytvoření anotace je potřeba znalosti obsahu příslušných nahrávek, na jejímž základě se trénují a testují vyvíjené systémy.

Standardním postupem pro vytvoření anotace je ortografická transkripce, tj. přepis nahrávek dle pravidel pravopisu daného jazyka. Mluvená řeč obsahuje ale navíc specifika, která je při anotaci třeba zohlednit. Jedná se např. o ruchy na pozadí nahrávky (šustění papíru, bouchnutí dveří, projíždějící sanitka atd.), neřečové události (smích, zakašlání, mlasknutí apod.), ale například i skutečnost, že číslovky mají různé výslovnostní varianty a v anotacích se proto musí vyskytovat pouze v textové podobě (např. číslovka čtyři se běžně vyslovuje ve variantách: "čtyři", "čtyry", "štyři" a "štyry"). Proto se pro ortografickou anotaci používají kromě pravopisných pravidel ještě další pravidla zohledňující odchylky psané a mluvené podoby jazyka.

V dokončovací fázi je třeba také dokončit podrobnou dokumentaci vytvořeného korpusu. Zaznamenat informace o druhu a účelu korpusu. Dále o typu řečového materiálu, který je v korpusu obsažen, pravidlech, která byla použita pro vytvoření anotace, a v případě veřejného korpusu uvést také kontaktní adresy, kde je možné získat o korpusu případně další informace.

2.4 Základní parametry signálů

Pod pojmem základní parametry signálů rozumíme elementární číselné charakteristiky signálů používané pro popis kvality signálů. Často používanými charakteristikami jsou střední hodnota (pro akustický signál je nulová), koeficient vybuzení, koeficient přebuzení, odstup signálu od šumu a další.

2.4.1 Koeficient přebuzení a koeficient vybuzení

Při nevhodném nastavení vstupní části snímacího zařízení může docházet k přebuzení. V případě signálů získaných prostřednictvím telefonní sítě je potom přebuzení často způsobeno blízkostí sluchátka a úst mluvčího.

Pro kvantifikaci úrovně vybuzení daného signálu se používají např. následující parametry:

- koeficient přebuzení - vyjadřuje četnost vzorků s maximální možnou amplitudou

$$koeficientPrebuzeni = \frac{N_{MAXAMP}}{N} 100[\,v\ procentech\,], \tag{2.1}$$

kde N_{MAXAMP} je počet vzorků s maximální možnou amplitudou, tj. pokud platí $|x[n]| = MAXAMP$.

- koeficient maximálního vybuzení - udává poměr mezi maximální absolutní hodnotou signálu a maximální možnou amplitudou použitého formátu

$$koeficientMaximalnihoVybuzeni = \frac{max|x[n]|}{MAXAMP} 100[\,v\ procentech\,], \tag{2.2}$$

- koeficient průměrného vybuzení - udává poměr mezi aritmetickou střední hodnotou (z absolutní hodnoty signálu) a maximální možnou amplitudou použitého formátu

$$koeficientPrumernehoVybuzeni = \frac{\sum_{n=0}^{N-1} |x[n]|}{N * MAXAMP} 100[\,v\ procentech\,]. \tag{2.3}$$

2.4.2 SNR - odstup signálu od šumu pro řečový signál

Měření úrovně šumu v signálů je nezbytnou součástí tvorby a anotací řečových databází. Standardním kritériem pro měření úrovně šumu v signálu je odstup signálu od šumu neboli SNR (z anglického Signal-to-Noise Ratio).

Pomocí kritérií na bázi SNR lze kvantifikovat aditivní šum v signálu. Vychází se přitom z modelu směsi daného jako [Pollák2001]:

$$x[n]=s[n]+n[n], \tag{2.4}$$

kde $s[n]$ představuje užitečný řečový signál, $n[n]$ rušivý signál a $x[n]$ jejich směs.

Základní definice pro výpočet SNR je dána vztahem:

$$SNR=10\log\frac{P_s^2}{P_n^2}, \tag{2.5}$$

kde P_s^2 je výkon užitečného signálu a P_n^2 je výkon šumu.

Globální SNR

Globální SNR (GSNR) dostaneme aplikací rovnice (2.5) na řečový signál, počítáme-li výkony řeči a šumu přes celý signál, tj:

$$GSNR=10\log\frac{P_s^2}{P_n^2}=10\log\frac{\sum_{n=0}^{l-1}s^2[n]}{\sum_{n=0}^{l-1}n^2[n]}=10\log\frac{E_s}{E_n} \tag{2.6}$$

Pro vyjádření globálního SNR se pracuje s analyzovaným signálem jako s celkem. Počítají-li se odhady výkonů signálu a šumu ze signálů stejné délky, je možné uvažovat namísto poměru výkonů poměr energií daných segmentů [Pollák2001].

Toto kritérium je ale zatíženo chybou, protože do výpočtu výkonu řeči jsou zahrnuty také části signálu bez řečové aktivity, které snižují celkový výkon řečového signálu. Správný výpočet SNR pro řečový signál je definovaný jako:

$$SNR=10\log\frac{\sum_{n=0}^{l-1}s^2[n]vad[n]}{\sum_{n=0}^{l-1}n^2[n]vad[n]}=10\log\frac{E_s'}{E_n'}, \tag{2.7}$$

kde $vad[n]$ nese informaci o řečové aktivitě pro daný vzorek signálu (1 - řeč, 0 - pauza).

Lokální SNR

Nestacionarita řečového signálu vede často na požadavek podchycení změn v SNR v závislosti na čase. Vzhledem ke kvazistacionaritě řeči pro framy (segmenty)

délky cca 10-25 ms, lze vystačit s vyčíslením SNR pro tyto framy. Mluví se o lokálním SNR pro i-tý frame definovaným jako:

$$SNR_i = 10\log\frac{P^2_{(s,i)}}{P^2_{(n,i)}} = 10\log\frac{\sum_{n=0}^{M-1} s^2[m\cdot i+n]}{\sum_{n=0}^{M-1} n^2[m\cdot i+n]}, \tag{2.8}$$

kde M je délka framu a m je krok segmentace.

Segmentální SNR

Segmentální SNR (SSNR) dostaneme, zprůměrujeme-li lokální SNR v jednotlivých framech. Toto průměrování se přitom provádí pouze přes framy s řečovou aktivitou (lokální SNR není v řečových pauzách definované). SSNR může být potom počítáno dle vztahu:

$$SSNR = \frac{1}{K}\sum_{i=0}^{L-1} SNR_i \cdot VAD_i, \tag{2.9}$$

kde VAD_i nese informaci o řečové aktivitě v i-tém framu (1 - řeč, 0 - pauza), L je celkový počet analyzovaných framů v signálu a K je počet framů s řečovou aktivitou, které se fakticky průměrují. Takto definované kritérium je z dlouhodobého hlediska málo ovlivněno nestacionaritou řečového signálu [Pollák2001].

Při měření SNR v reálných signálech řečových databází je k dispozici však typicky pouze směs řeči a rušivého pozadí $x[n]$ a jak výkon řeči P_s^2 , tak výkon šumu P_n^2 je nutné z dostupné směsi odhadnout.

Uváží-li se předpoklad nekorelovanosti řeči a šumu, lze SNR obecně vyjádřit jako:

$$SNR = 10\log\frac{P_s^2}{P_n^2} = 10\log\frac{P_x^2 - P_n^2}{P_n^2} \tag{2.10}$$

a úloha se tak zjednoduší pouze na odhad výkonu šumu.

Existují dva základní přístupy k odhadu výkonu šumu. První skupinu tvoří algoritmy využívající informace o řečové aktivitě a získávající výkon šumu průměrováním v řečových pauzách. Druhou skupinu pak tvoří metody hledající výkon šumu sledováním minima krátkodobého výkonu [Pollák2001].

Odhad globálního SNR s detektorem

Pro relativně krátké signály se často počítá globální SNR s detektorem řečové aktivity. Principem tohoto algoritmu je odhad výkonu šumu z pauz v dané promluvě a

odhad výkonu řeči pouze z řečových úseků signálu [Pollák2001]. Algoritmus odhadu SNR lze shrnout následujícími třemi vztahy:

$$\widehat{SNR} = 10 \log \frac{\widehat{P_s^2}}{\widehat{P_n^2}}, \tag{2.11}$$

$$\widehat{P_n^2} = \frac{1}{l_n} \sum_{n=0}^{l-1} x^2[n] \cdot |1 - vad[n]|, \tag{2.12}$$

$$\widehat{P_s^2} = \frac{1}{l_s} \sum_{n=0}^{l-1} x^2[n] \cdot vad[n] - P_n^{2\cdot} \tag{2.13}$$

Pří určování SNR s detektorem řečové aktivity se může stát, že detektor v nahrávce detekuje pouze řeč či pouze šum. To se může stát jednak proto, že nahrávky tento obsah skutečně mají, či na základě selhávání detekce řečové aktivity. V takovýchto případech je SNR teoreticky rovno +/- ∞. V praktických aplikacích se ale tato hodnota obvykle nastavuje na konečnou mez (např +/- 30 dB).

Odhad globálního SNR z histogramů krátkodobého výkonu

Tento algoritmus vychází z odhadu výkonu šumu na základě hledání minima výkonu směsi. Vychází se přitom z krátkodobého výkonu celého signálu a předpokládá se, že určitá část nejnižších hodnot reprezentuje šum. Výkon šumu je poté dán průměrem L_n nejmenších hodnot krátkodobého segmentálního výkonu zašuměného signálu $P_{(x,i)}^2$. Hodnota L_n se volí z celkového počtu framů jako $L_n = b \cdot L$, $b \in (0,1)$. Algoritmus odhadu SNR z histogramů krátkodobého výkonu lze shrnout následujícími třemi vztahy:

$$\widehat{SNR_h} = 10 \log \frac{P_x^2 - \widehat{P_n^2}}{\widehat{P_n^2}}, \tag{2.14}$$

$$\widehat{P_n^2} = \frac{1}{L_n} \sum_{i=0}^{L_n-1} P_{(x-sort,i)}^2, \tag{2.15}$$

kde $P_{(x-sort,i)}^2$ představuje vzestupně seřazené krátkodobé segmentální výkony signálu.

$$\widehat{P_x^2} = \frac{1}{L} \sum_{n=0}^{L-1} P_{(x,i)}^2. \tag{2.16}$$

O metodách odhadu lokálního a segmentálního SNR se lze dočíst v [Pollák2001] či v [Pollák2002].

Detekce řečové aktivity pro výpočet SNR

Klíčovou součástí většiny algoritmů odhadu odstupu signálu od šumu je detekce řečové aktivity (VAD - Voice Activity Detection). Řečové detektory se dají rozdělit do tří skupin:

ideální detekce - typicky ruční nastavení řečových úseků

kepstrální detektor - indikuje řeč na základě vzdálenosti mezi kepstrem aktuálně zpracovávaného framu a kepstrem průměrovaným v řečových pauzách

energetický detektor - stanovuje práh pro detekci řeči na základě sledování minima a maxima krátkodobého výkonu signálu

2.5 Foneticky bohatý korpus

Naprosto dokonalé pokrytí všech možných promluv je v korpusu samozřejmě nemožné. To je i jedním z důvodů, proč je rozpoznávání založeno na rozpoznávání úseků řeči menších než jsou slova. Jedná se o rozpoznávání fonémů, difónů či trifónů (kontextově závislých fonémů). Tyto úseky slov jsou poté brány jako základní akustické elementy promluvy s následným jazykovým modelem či gramatikou pro určení rozpoznané promluvy. Pro tyto akustické elementy jsou potom trénovány např. modely nejčastěji používaných rozpoznávačů řeči na bázi skrytých Markovských modelů (HMM).

Kvalitní natrénování všech možných akustických elementů zajišťuje tzv. foneticky bohatý materiál. Tj. korpus vět nebo slov pokrývající v dostatečné míře všechny fonémy vyskytující se v daném jazyce.

Všeobecným požadavkem pro foneticky bohatý korpus je takové pokrytí fonémů, které odpovídá jejich přirozené četnosti v mluveném jazyce. V praktickém korpusu je však nutné četnost zastoupení jednotlivých fonémů mírně modifikovat. To především z důvodu omezenosti navrhovaného korpusu. Tj. pokud není korpus dostatečně velký (což není prakticky nikdy) je nutné výskyt řídce se vyskytujících fonémů nadhodnotit na úkor fonémů s nejvyšší relativní četností výskytu.

3. Tvorba korpusu TULTEL07

Práce na tvorbě korpusu představovala nejprve vytvoření scénářů nahrávek, dle nichž mluvčí při nahrávce postupovali, dále nábor mluvčích a pořizování samotných nahrávek, zaznamenávání informací o nahrávkách a následně anotaci nahrávek. Pro podporu uvedených činností byly také vytvořeny dva programy.

3.1 Přípravná fáze

Plánovaný počet mluvčích v korpusu byl stanoven na 160. Vzhledem k tomu, že měl korpus sloužit mimo jiné pro vývoj systému rozpoznávání řečníka, měl každý z plánovaných 160ti mluvčích pořídit 2 nahrávky (telefonáty). Kvůli plánované robustnosti systému měl být přitom mezi těmito dvěma nahrávkami časový odstup alespoň pěti dnů a obě nahrávky měly být pořízeny z jiného typu telefonní sítě (nejlépe pak jedna z pevné linky a druhá ze sítě mobilní). Kvůli alespoň částečné věkové vyváženosti mluvčích v korpusu mělo být aspoň 25% mluvčích nad 40 let. Jednotlivé části telefonátu (viz. Tab 3.2) od sebe měly být odděleny tichem (cca 3 vteřiny). Jednotlivé položky scénáře nahrávky (3.1.2) měly být vyslovovány zcela libovolně (např. výraz "1K8 9819" se mohl číst jako: "jedna ká osm devadesát osm devatenáct", ale stejně tak např. i: "jedna k osum devět osum jedna devět" apod.).

Stanovení základních požadavků na korpus TULTEL07 dle [Radová2004] je v následujícím přehledu:

Tab. 3.1: Základní požadavky na korpus TULTEL07

typ	telefonní
účel	pro úlohy rozpoznávání řeči a rozpoznávání řečníka
počet mluvčích	cca 160
typ mluvčích	rovnoměrné zastoupení pohlaví, přiměřeně věkově vyváženo
nahrávací podmínky	požadavek na ticho v pozadí
typ řeči	viz. Tab. 3.2

3.1.1 Praktický nábor mluvčích

Pro nábor mluvčích byla použita metoda částečně řízeného lavinového náboru. Plánovaných 160 mluvčích se rozdělilo rovnoměrně na 10 organizátorů náboru mluvčích, kteří měli mj. za úkol kromě samotného pořízení nahrávek i jejich anotaci.

Organizátory náboru se stali studenti Fakulty mechatroniky TUL. Všichni organizátoři náboru se na této konkrétní práci podíleli formou placené spolupráce (spolupráci financoval Ústav ITE, který je součástí Fakulty mechatroniky).

3.1.2 Přípravná fáze z pohledu organizátora nahrávek

Úkolem organizátorů náboru mluvčích v rámci přípravné fáze tvorby korpusu bylo vytvořit pro své mluvčí scénáře hovorů. Tyto scénáře měly následující strukturu:

Tab. 3.2: Scénář hovoru

	položka scénáře hovoru	**specifikace**
1	uvítací pozdrav	libovolný, např. „ahoj", „zdar", „čus", „dobrý den", ...
2	datum	měsíc slovy, např. „úterý 12. května 1962"
3	identifikátor-čísla	8-10 číslic
4	identifikátor-kombinace	kombinace číslic a písmen (bez diakritiky), 8-10 znaků
5	SPZ	nový tvar, např. „1L1 1234"
6	telefonní číslo	národní tvar => 9 číslic => bez mezinárodní předvolby
7	adresa	ulice, číslo popisné, město, PSČ např. „Hálkova 6, Liberec 1, 460 01"
8	souvislý text	libovolný, doba čtení cca 3 minuty, bez cizích slov (především vlastních jmen)
9	loučící pozdrav	libovolný, např. „ahoj", „na shledanou", „čau", „zdar", ...
10	spontánní promluva	výzva ke spontánní promluvě charakteru objednávky jízdenky/letenky (odkud, kam, v kolik odjezd, ...)

Souvislé texty scénářů hovoru, které tvořily největší část těchto scénářů, bylo doporučeno čerpat z internetu, protože se tam vyskytovaly již v elektronické podobě. Ve snaze o co nejširší slovník korpusu (seznam slov obsažených v korpusu) byl na serveru Ústavu ITE vytvořen web, na kterém si organizátoři tyto souvislé texty zamlouvali. Přínos tohoto webu byl v tom, že při správném použití neumožňoval zamluvit si jeden souvislý text z internetu dvakrát, čímž byla zvýšena šance na různorodost slov obsažených v korpusu.

Každý mluvčí měl ke své první nahrávce jiný scénář hovoru než ke své druhé nahrávce. S ohledem na prakticky vyloučenou možnost duplicity souvislých textů v korpusu to tak znamenalo pro každého z organizátorů připravit celkem 32 různých scénářů hovoru.

Tyto scénáře organizátoři zároveň také odevzdávali z důvodu budoucí možnosti porovnání, kolik ze čtených položek scénáře přečetli mluvčí správně. Pro tento účel se jevil pro scénáře hovoru ideální formát typu prostý text. Zároveň ale bylo potřeba, aby

se daly tyto scénáře jednoduše vytisknout. Z těchto důvodů byl pro scénáře hovoru vybrán formát HTML.

3.1.3 Program na vytváření koster scénářů hovoru

Pro účely ušetření práce organizátorů a zároveň téměř jistoty správného obsahu scénářů byl vyvinut v programovacím jazyce Java program, který vytvářel jakési kostry těchto scénářů. Ty se od scénářů popsaných v Tab. 3.2 lišily tím, že neobsahovaly adresu a souvislý text. Všechny ostatní položky scénářů kromě spontánní promluvy (viz dále) program náhodně generoval. Způsob implementace jednotlivých položek scénáře hovoru v programu shrnuje Tab 3.3.

Tab. 3.3: Způsob implementace jednotlivých položek scénáře hovoru

	položka scénáře hovoru	**implementace v programu**
1	uvítací pozdrav	výběr z devíti různých uvítacích pozdravů
2	datum	den v týdnu a měsíc - výběr ze všech příslušných variant den v měsíci - výběr mezi čísly 1 - 30 rok - výběr mezi čísly 1951 - 2015
3	identifikátor-čísla	nejprve výběr počtu číslic z čísel 8 - 10 poté náhodný výběr číslic ze všech možných variant
4	identifikátor-kombinace	nejprve výběr počtu znaků z čísel 8 - 10 poté výběr ze všech možných variant tak, aby identifikátor obsahoval stejný počet číslic a písmen (v případě devíti znaků o jednu číslici/písmeno navíc)
5	SPZ	výběr první číslice SPZ z čísel 1 - 5 výběr písmena kraje ze všech možných variant pro ČR výběr ostatních pěti číslic SPZ ze všech možných variant
6	telefonní číslo	výběr první číslice z čísel 1 - 9 (nulou tel. čísla v ČR již nezačínají) výběr ostatních osmi číslic ze všech možných variant
7	adresa	neimplementováno
8	souvislý text	neimplementováno
9	loučící pozdrav	výběr z devíti různých loučících pozdravů
10	spontánní promluva	pro první čtení mluvčího vygenerován text vybízející ke spontánní promluvě na téma objednání jízdenky, pro druhé čtení mluvčího vygenerován text vybízející ke spontánní promluvě na téma objednání letenky

Spontánní promluva byla jedinou položkou scénáře hovoru, která nebyla přímo čtena. Scénář hovoru obsahoval na místě spontánní promluvy vyzvání k tomu, aby si mluvčí vlastními slovy objednal jízdenku/letenku s libovolnými parametry objednávky.

Tvorba scénářů hovoru se tak pro organizátory díky tomuto programu zjednodušila na to, doplnit do vytvořených koster scénářů adresu a souvislý text. Ukázka vytvořené kostry scénáře hovoru je v Příloze A.

3.2 Záznamová fáze

Samotné nahrávání telefonních hovorů bylo uskutečněno prostřednictvím nahrávacího automatu. Mluvčí vytočil konkrétní telefonní číslo, na kterém se mu v případě dovolání ozval nahrávací automat a vyzval ho k začátku telefonního hovoru. Nahrávací automat byl reprezentován PC.

3.2.1 Záznamová fáze z pohledu organizátora nahrávek

Úkolem organizátorů nahrávek v záznamové fázi bylo nejprve sehnat 16 mluvčích, kteří byli ochotni uskutečnit nahrávky (dle 3.1.). Před samotným uskutečněním nahrávek (zvláště té první) mluvčí seznámit s tím, co se po nich chce. A po uskutečnění nahrávky zaznamenat o nahrávce následující údaje: datum a čas pořízení, jméno, příjmení, věk a dialekt mluvčího, druh spojení (mobil/pevná/VoIP/Skype), výrobce telefonního přístroje (případně i typ) a v případě pevné linky ještě údaj o přenositelnosti přístroje.

Při uskutečňování nahrávek bylo důležité zajistit ticho na pozadí (viz 3.1). To prakticky znamenalo např. zavřít okna v místnosti, uvědomit ostatní obyvatele bytu, ve kterém se telefonát uskutečňoval, aby zůstali po dobu nahrávky v jiném pokoji či zajistit ostatní možné hluky (hudba, domácí spotřebiče, telefony). Dobré bylo pro organizátora i pro mluvčího, když byl mluvčí se scénářem nahrávky seznámen už před nahrávkou. Mluvčí se tak v rámci možností uklidnil a zároveň byla i zmenšena šance na to, že bude při nahrávce ve čtených položkách scénáře chybovat. To znamenalo následně i méně práce pro organizátora nahrávky při jejím přepisování viz. 3.3. Při pořizování prvních nahrávek se stávalo, že když měl mluvčí scénář hovoru na papíře, který držel samotný v ruce, nevyhnul se přitom šustění, které bylo poté slyšet i v nahrávce. Proto bylo pro ostatní nahrávky doporučeno, aby měl mluvčí při čtení papír buď ve fólii či na nějaké pevné podložce (např. stůl nebo kniha).

Problémem se ukázalo být splnění požadavku na pořízení každé nahrávky mluvčího z jiného typu telefonní sítě. Jako nejrozšířenější síť současnosti se projevila jednoznačně síť mobilní a problém nezřídka byl s pořízením telefonátů z jiného typu sítě. A to včetně pevné linky, která byla jako doplnění sítě mobilní pro nahrávky

preferovaná. Řešením v případě, kdy mluvčí neměl pevnou linku doma, bylo např. uskutečnění hovoru z pevné linky ze zaměstnání či od známých. Při uskutečňování hovorů z pevných linek ze zaměstnání se dal přitom obtížněji zajistit požadavek na ticho na pozadí. Snaha v takových případech byla rušivé pozadí v dané kanceláři minimalizovat.

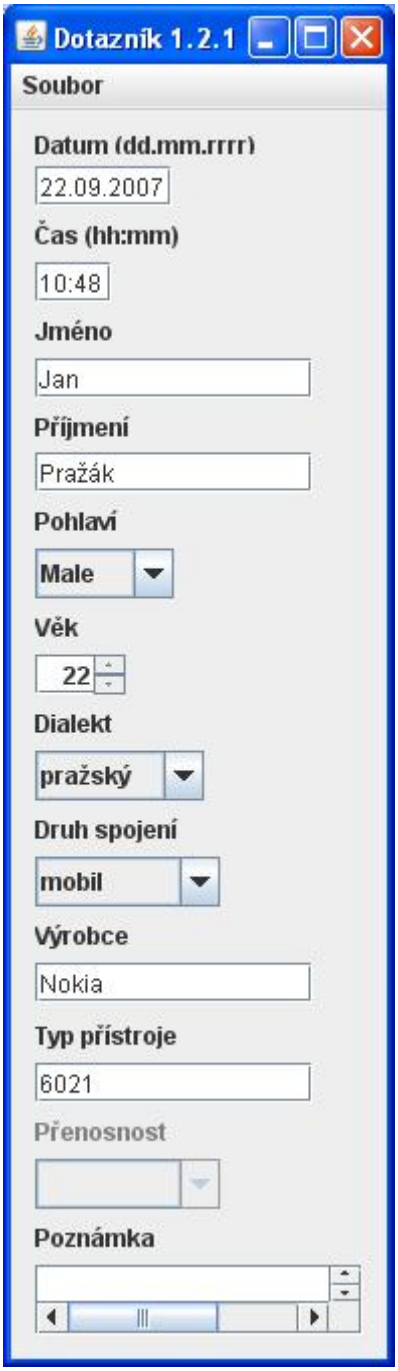
3.2.2 Program Dotazník

Při vytváření korpusu TULTEL07 byl vznesen požadavek na to, aby byly informace o nahrávkách k dispozici v elektronické podobě. Z toho důvodu byl vyvinut (v programovacím jazyce Java) program Dotazník. Prostřednictvím něj pak organizátoři nahrávek zaznamenávali informace o nahrávkách. Okno tohoto programu je na Obr. 3.1. Výstup tohoto programu byl ve formátu XML a je zobrazen v Příloze C.

Program zajišťoval především:

1) vytvoření a zpětnou editaci formuláře k nahrávce
2) základní prevenci správnosti vyplnění a kontrolu úplnosti vyplnění formuláře
3) generování názvu souboru formuláře dle smluvené konvence (datem a časem pořízení nahrávky pevně daných 18 znaků + koncovka ".xml").

Tento program je k dispozici na přiloženém CD.



Obr. 3.1: Okno programu Dotazník

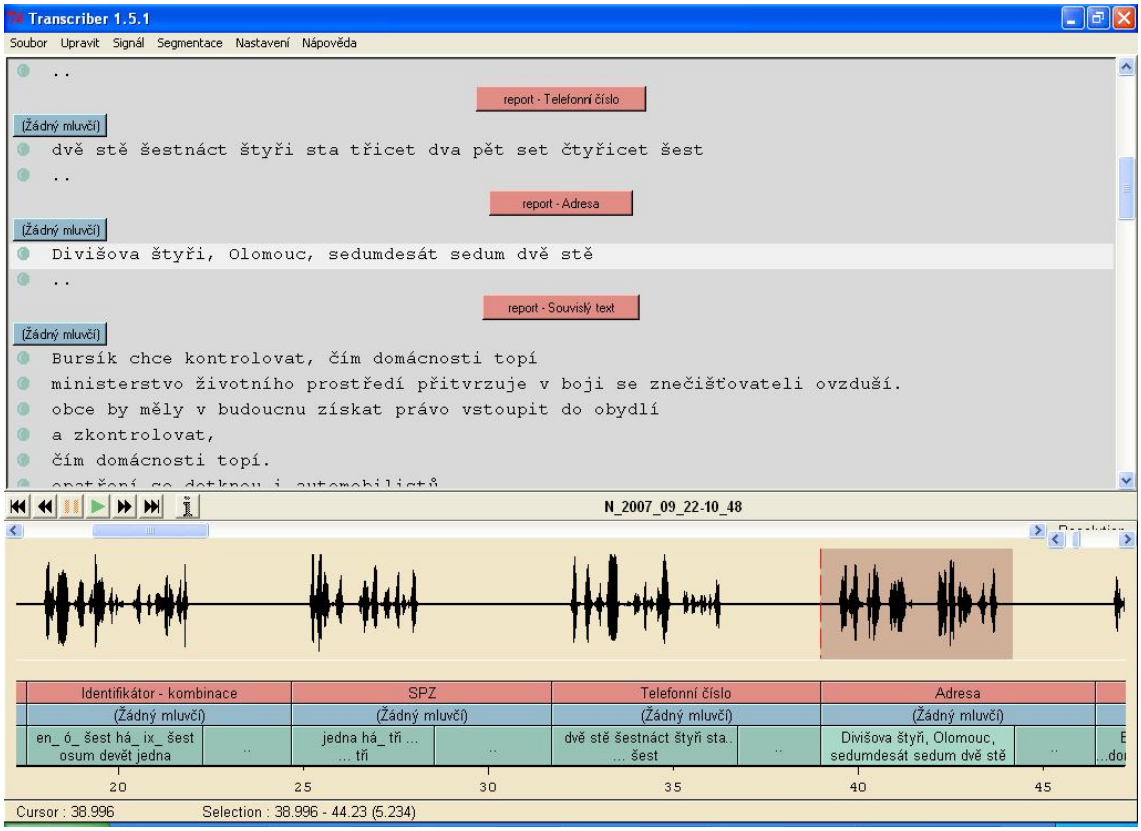
3.3 Dokončovací fáze

Dokončovací fáze představovala anotaci pořízených nahrávek a vytvoření dokumentace korpusu. Nejprve se provedla ortografická anotace, kterou prováděli organizátoři nahrávek. Z té se poté pomocí specializovaného software vygenerovala

anotace fonetická, která byla následně ještě ručně zkontrolována. Dokumentace korpusu byla vytvořena v rámci této práce.

3.3.1 Dokončovací fáze z pohledu organizátora nahrávek

Úkolem organizátora nahrávek v rámci dokončovací fáze tvorby korpusu byla ortografická anotace nahrávek (telefonátů) dle anotačních pravidel konsorcia vysokých škol TU Liberec, VUT Brno a ZČU Plzeň. Tato anotace byla prováděna v programu Transcriber. Jedná se o open source, který je k dispozici např. na serveru sourceforge.net. Okno programu Transcriber je na Obr. 3.2.



Obr. 3.2: Okno programu Transcriber

Tento program umožňuje jednoduše přehrát i anotovat libovolnou část signálu. Horní část okna představuje textový editor, do kterého se zapisuje výsledná anotace nahrávky. Ve spodní části okna je zobrazen signál, který je anotován. V tomto okně se dá signál velmi jednoduše segmentovat, což se ihned projeví i v okně horním.

V tomto programu měli anotátoři za úkol především tyto činnosti:

1) rozdělit signál na menší části v závislosti na promluvách mluvčího (segmentaci signálu)

2) přidělit vzniklé segmenty jednotlivým částem scénáře hovoru (na Obr. 3.2 s červeným pozadím)

3) anotovat segmenty dle požadovaných pravidel

Prakticky to znamenalo nejprve si stáhnout ze serveru Ústavu ITE vlastní pořízenou nahrávku (identifikovanou v názvu dle data a času pořízení). Následně otevření nahrávky v programu Transcriber a vykonání tří výše uvedených činností.

Segmentace byla prováděna tak, aby před i za každou promluvou v rámci jednoho segmentu bylo ticho (cca desetiny sekundy). Tím se předcházelo možnému nechtěnému rozdělení jednotlivých promluv do více segmentů, které by se v anotacích velmi pravděpodobně nezřídka vyskytovalo, kdyby se hranice segmentů umísťovaly přímo před a za jednotlivé promluvy.

Přidělení vzniklých segmentů jednotlivým částem scénáře hovoru obnášelo v době prvních anotací každou část přepisu ručně pojmenovat. Vzhledem k počtu částí scénáře hovoru (10) a k počtu nahrávek každého organizátora (32) to ale byla jednak poměrně zdržující činnost a jednak bylo pravděpodobné, že se při takovém množství těchto ručních pojmenování organizátoři nevyvarují překlepů. Z těchto důvodů byl vytvořen skript, po jehož aplikaci se přidělení segmentů pasážím scénáře hovoru zjednodušilo na výběr názvu pasáže pro danou část scénáře hovoru ze všech příslušných variant. Za pomoci tohoto skriptu pak byly prováděny všechny ostatní anotace.

Při anotaci segmentů se vycházelo ze scénáře hovoru k anotované nahrávce, který měl anotátor v elektronické podobě k dispozici. Anotace probíhala typicky tak, že si anotátor přehrál daný segment nahrávky a přidělil mu odpovídající část scénáře hovoru. Spolu s tím kontroloval, jestli mluvčí přečetl danou část scénáře hovoru správně a pokud ne, zapsal do anotace skutečný obsah jeho promluvy. V případě spontánní promluvy neměl anotátor pochopitelně předlohu v elektronické podobě k dispozici a byl nucen její obsah přepsat pouze na základě poslechu. Uvedené činnosti prováděl anotátor přitom s ohledem na anotační pravidla. Nejčastěji používaná anotační pravidla shrnuje Tab. 3.4.

Tab. 3.4: Nejčastěji používaná anotační pravidla

Objekt přepisu	Pravidlo přepisu	Příklad	
		obsah nahrávky	přepis
číslovky	pouze v textové podobě	"štyřicet"	štyřicet
hláskování	výslovnostně	"čé er"	čé_ er_
nedokončené slovo	speciální značka	"str"	str+
nesrozumitelné slovo	speciální značka	něco jako "šrakrš"	??
neutrální hláska	speciální značka	"čr" (ne "čé er")	č=_ r=_
novotvary, neznámá a cizí slova	speciální párová značka	"řikl, že"	(řikl), že
polykání hlásek	v případě nepolknutí celé slabiky zanedbat	"devatenácet"	devatenáct set
ruchy okolí	speciální párová značka	průjezd sanitky	%%
ruchy řečníka	speciální značka	zakašlání	§
slangová a nářeční výslovnost	výslovnostně	"bejvák"	bejvák
souhlas, nesouhlas a váhací zvuky	speciální značky	znělé váhání	2=
ticho	speciální značka pro segment ticha	ticho	..
věty	s malým písmenem na začátku	"má rád sport"	má rád sport.

Pracný byl zejména přepis číslovek, které se ve scénářích hovoru objevily v číselné podobě (v Tab. 3.2 položky 2 - 7 vždy, souvislý text a spontánní řeč téměř vždy). Ta se v anotacích nesmí vyskytovat a v těchto případech bylo potřeba vyslovenou číslovku vždy ručně vypsat v její příslušné výslovnostní variantě (viz 2.3).

Občas se v souvislých textech scénářů hovoru vyskytovaly pravopisné chyby. Ty byly v přepisech nežádoucí a byla proto snaha je tam opravovat.

3.3.2 Dokumentace korpusu

Součástí dokončovací fáze tvorby korpusu je také vytvoření dokumentace korpusu. Ta se vytváří zejména pro účely budoucího využití korpusu. Dokumentace ke korpusu TULTEL07 je shrnuta v Tab 3.5. Obsah scénáře nahrávek, který se v dokumentaci korpusu také obvykle uvádí, je již v Tab. 3.2.

Tab 3.5: Dokumentace korpusu

Identifikace korpusu	TULTEL07
Dostupnost	neveřejný
Účel	úloha rozpoznávání řeči úloha rozpoznávání řečníka
Počet mluvčích Celkový počet nahrávek Celkový počet promluv	160 322 16976
Počet nahrávaných položek	10
Průměrná délka řeči od jednoho mluvčího	8,13 minuty
Celkový rozsah DB	21,69 hodin řeči cca 1,23 GB řeči
Rovnoměrné pokrytí od obou pohlaví	TÉMĚŘ ANO (81 mužů, 79 žen)
Proporcionální pokrytí věku Proporcionální pokrytí dialektu	ANO NE (informace o dialektu ANO)
Anotace	ortografická transkripce fonetická transkripce značky pro neřečové události značky pro šum značená přeřeknutí apod.
Pravidla použitá pro anotaci	anotační manuál projektu konsorcia škol VUT Brno, TU Liberec a ZČU Plzeň
Vzorkovací frekvence Počet kanálů kvantování Počet bytů na vzorek	8000 Hz 1 a-law 1
Prostředí	telefonní kanál
Typy sítí obsažených v DB	mobilní pevná VoIP

4. Analýza korpusu TULTEL07

Analýza korpusu je rozčleněna na dvě části. V první části se zabývá analýzou signálů korpusu, která je dále rozdělena na přebuzení a vybuzení signálů a na odhad odstupu signálu od šumu. Druhá část se týká analýzy fonetické bohatosti korpusu.

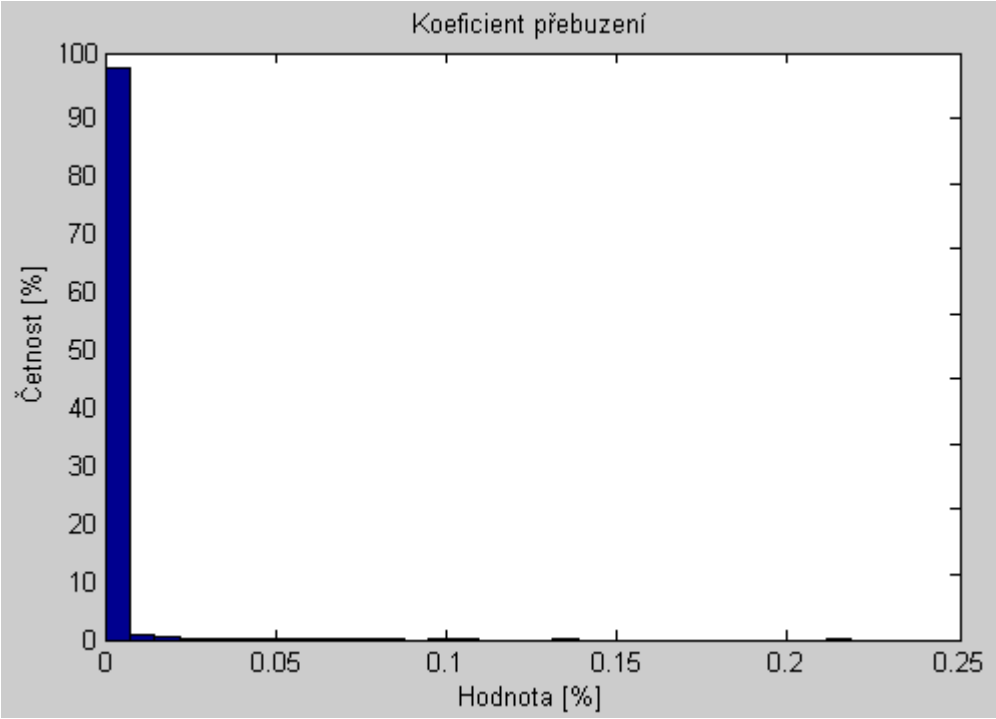
4.1 Analýza signálů v databázi

Všechny nahrávky (telefonáty) korpusu byly v rámci ortografické anotace rozsegmentovány (viz. 3.3.1) a analýza signálů se týkala již pouze řečových segmentů těchto nahrávek. Z 322 telefonátů tak vzniklo 16 976 promluv, které byly analyzovány. Kompletní analýza signálů byla provedena v prostředí Matlab.

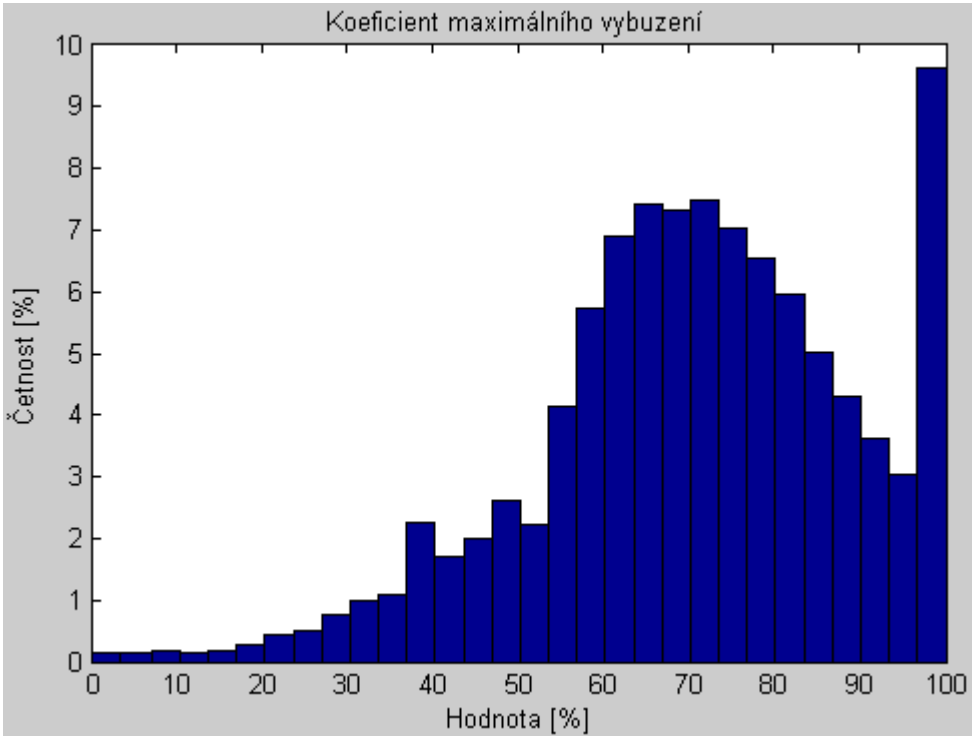
Z důvodu množství analyzovaných dat bylo se postupovalo tak, že se nejprve pro každý analyzovaný parametr spočítaly hodnoty ve všech analyzovaných promluvách a tyto hodnoty byly následně uloženy (vždy vektor hodnot s 16 976 prvky). Jako formát exportu dat byl zvolen mat soubor (nativní matlabovský formát pro import/export dat). Poté se tento mat soubor načetl a provedla se vlastní analýza. Tj. samotný výpočet vektorů jednotlivých parametrů proběhl pouze jednou (na počítači s procesorem 1,4GHz a 512MB RAM trval řádově desítky minut).

4.1.1 Koeficient přebuzení a koeficient vybuzení

Koeficient přebuzení procentuálně udává počet přebuzených vzorků v signálu (viz. 2.4.1). Z Obr. 4.1 je patrné, že pouze minimum řečových segmentů nahrávek v korpusu je přebuzených (obsahuje alespoň jeden přebuzený vzorek). Takovýchto nahrávek jsou v korpusu cca 4%.



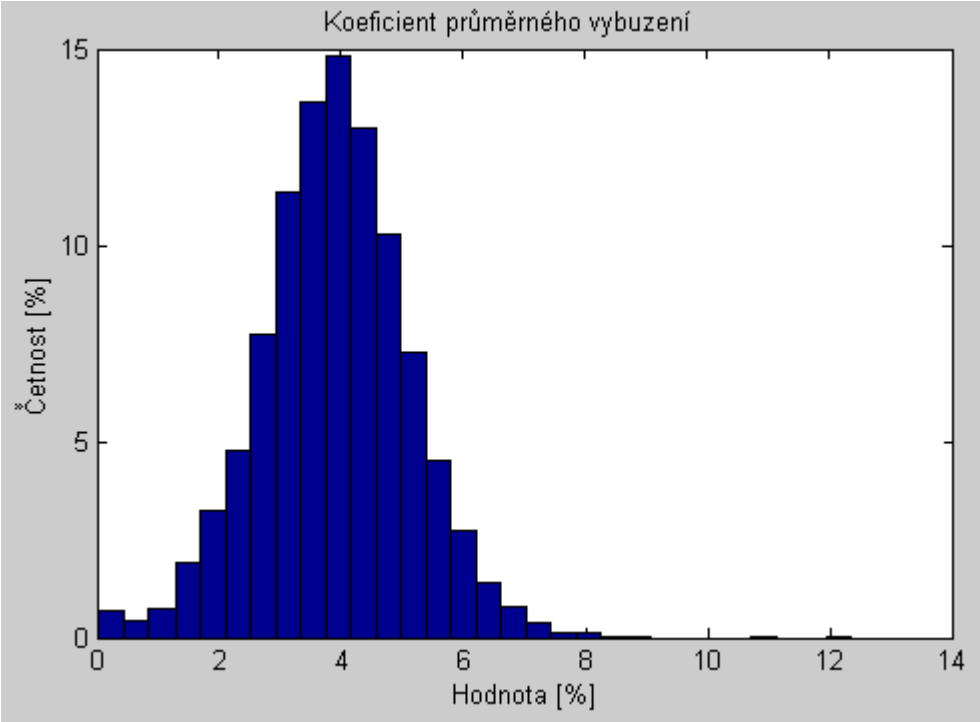
Obr. 4.1: Histogram koeficientů přebuzení řečových segmentů nahrávek



Obr. 4.2: Histogram koeficientů maximálního vybuzení řečových segmentů nahrávek

Koeficient maximálního vybuzení udává poměr mezi maximální absolutní hodnotou amplitudy v signálu a maximální možnou amplitudou daného formátu (viz. 2.4.1). Představuje tak tedy využití dynamického rozsahu použitého záznamového formátu. Rozložení hodnot tohoto parametru v databázi znázorňuje Obr. 4.2. Skutečnost, že necelých 10% analyzovaných řečových segmentů nahrávek má hodnotu koeficientu cca mezi 97% - 100% je dána mj. již zmíněnou přebuzeností cca 4% z nich. Průměrná hodnota tohoto parametru v databázi je 69,8%.

Koeficient průměrného vybuzení udává poměr mezi aritmetickou střední hodnotou z absolutní hodnoty signálu a maximální možnou amplitudou použitého formátu (viz. 2.4.1). Rozložení hodnot tohoto parametru v databázi vyjadřuje Obr. 4.3. Průměrná hodnota tohoto parametru v databázi je 3,9%.



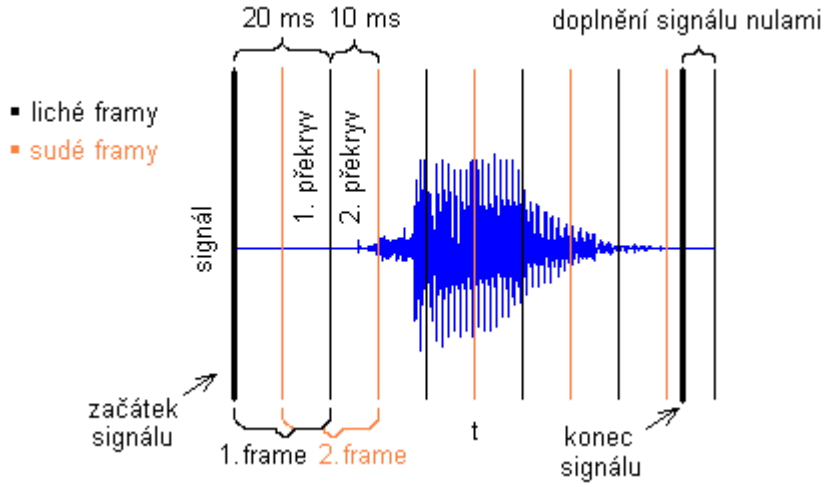
Obr. 4.3: Histogram koeficientů průměrného vybuzení řečových segmentů nahrávek

4.1.2 SNR - odstup signálu od šumu

Odstup signálu od šumu charakterizuje úroveň šumu v signálu (viz. 2.4.2). V rámci analýzy signálů byla u řečových segmentů nahrávek odhadnuta dvě globální SNR - SNR s detektorem řečové aktivity a SNR z histogramů krátkodobého výkonu. Globální SNR s detektorem řečové aktivity se u relativně krátkých signálů určuje často [Pollák2001]. Přesnost jeho odhadu se ale odvíjí od kvality použitého detektoru řečové

aktivity. Pro možnost srovnání byl implementován i odhad globálního SNR z histogramů krátkodobého výkonu, k jehož určení není detektoru řečové aktivity potřeba.

Při odhadu obou globálních SNR byla zvolena délka jednotlivých framů 20ms a překrývání framů 50%. Signály byly případně doplněny (na délku odpovídající segmentaci) nulami. Popsané údaje shrnuje Obr. 4.4.



Obr. 4.4: Zpracování signálů pro odhad SNR

Vývoj detektoru řečové aktivity

Prioritně pro odhad globálního SNR s detektorem řečové aktivity byl realizován (v Matlabu) energetický řečový detektor s adaptivním prahem doplněný na základě experimentů pěti funkcemi pro zlepšení detekce řeči. Pro testování tohoto detektoru bylo vybráno z celé databáze 30 promluv, které alespoň přibližně reprezentovaly celou databázi. Na těchto řečových segmentech nahrávek byla následně učiněna ruční detekce řeči. Ta byla přitom vztažena k jednotlivým framům promluvy (v závislosti na segmentaci viz. výše). Tato detekce byla dále považována jako referenční. Poté byl naprogramován jednoduchý vyhodnocovač úspěšnosti detektoru, který vracel jeho rozpoznávací skóre (počet správně detekovaných framů / počet všech framů). A následně byl po jednotlivých krocích vyvíjen detektor tak, aby mu rostlo rozpoznávací skóre. Zároveň byla u vybraných segmentů nahrávek neustále zobrazována referenční detekce v porovnání s detekcí detektoru (viz. Příloha B).

Jednotlivé kroky vývoje detektoru byly následující:

1) realizace energetického detektoru s adaptivním prahem

Tento základní krok spočíval v těchto třech podkrocích:

1a. Nalezení průměrné maximální a průměrné minimální hodnoty logaritmické energie framů v signálu zprůměrováním N nejvyšších (resp. N nejnižších) hodnot logaritmických energií framů v signálu. U N nejvyšších hodnot se předpokládalo, že budou reprezentovat řeč a u N nejnižších hodnot, že budou reprezentovat šum.

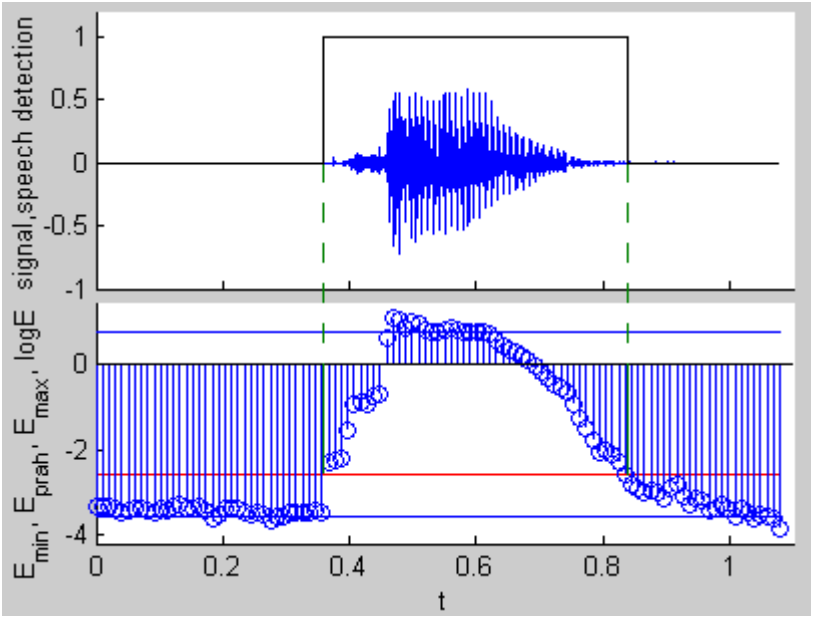
1b. Určení hodnoty prahové energie framů dané vztahem:

$$E_{prah} = \bar{E}_{min} + (\bar{E}_{max} - \bar{E}_{min}) * konst \quad , \qquad (4.1)$$

kde $\bar{E}_{min}$ je průměrná minimální hodnota logaritmické energie framů v signálu, $\bar{E}_{max}$ je průměrná maximální hodnota logaritmické energie framů v signálu a $konst \in (0,1)$ a představuje skutečnost, že hodnota prahové energie bude ležet mezi hodnotami $\bar{E}_{min}$ a $\bar{E}_{max}$.

1c. Všechny framy, u nichž je hodnota logaritmické energie větší, než E_{prah} , označ jako řečové. Ostatní framy v signálu označ jako šum.

Úspěšnost detektoru na reprezentativní množině v této fázi : cca 89,5 %.



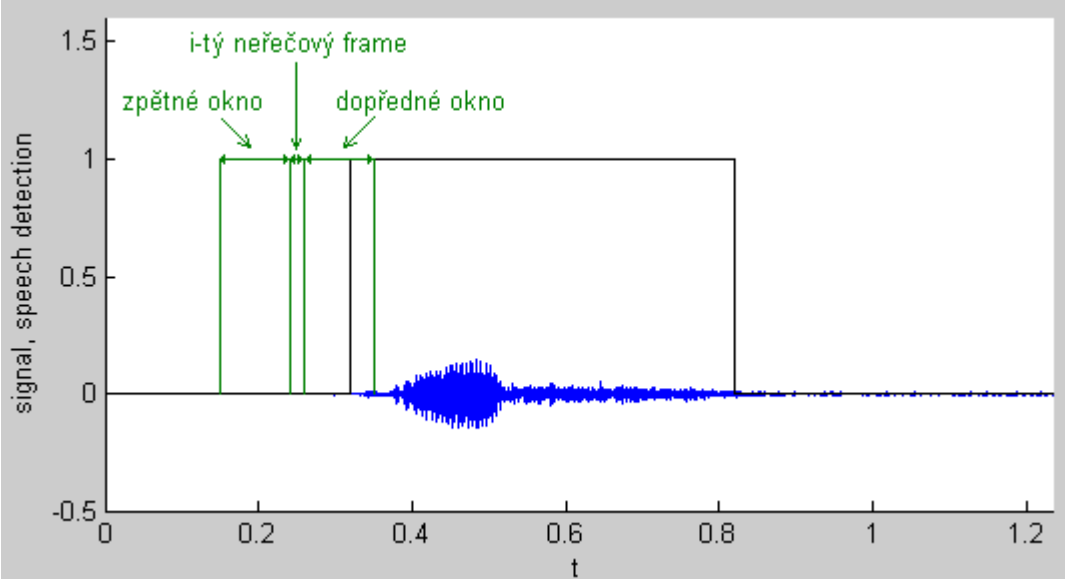
Obr. 4.5: Realizace energetického detektoru s adaptivním prahem

2) změna detekce u framů šumu obklopených framy řeči

V druhém kroku vývoje detektoru bylo pro každý neřečový frame zjištěno, jestli se v okně před ním i za ním nachází alespoň jeden frame řeči a pokud ano, byla u takových framů změněna detekce na framy řečové.

Velikost zpětného i dopředného okna: 9 framů

Úspěšnost detektoru na reprezentativní množině v této fázi : cca 93 %.



Obr. 4.6: Změna detekce u framů šumu obklopených framy řeči

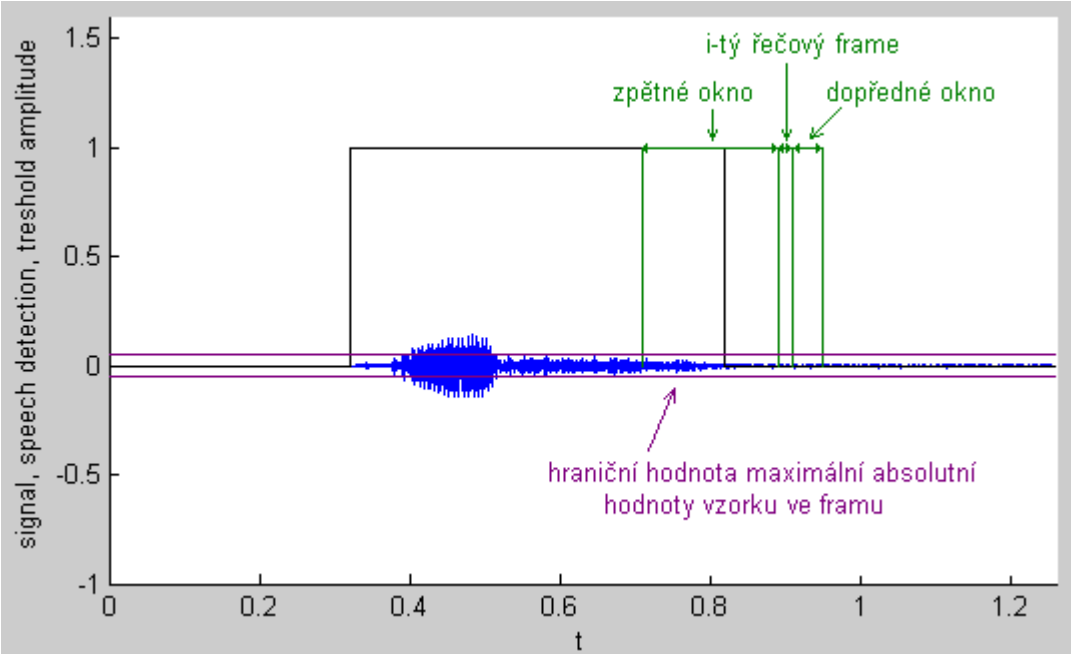
3) změna detekce u framů řeči obklopenými framy šumu – snaha o základní odstranění nesprávné detekce řeči u nádechů a ostatních ruchů

Detektor ve fázi 2 často nesprávně detekoval nádechy, výdechy a ostatní menší ruchy jako řeč. Tyto drobné ruchy měly často společné to, že byly reprezentovány vzorky s nízkou amplitudou. Proto bylo v tomto kroku vývoje detektoru pro všechny řečové framy zjištěno, jestli se na časové ose okolo nich nacházejí pouze framy s nízkými hodnotami amplitud. Pokud tomu tak bylo, byla u procházeného framu změněna detekce řeči z 1 na 0.

Velikost zpětného okna : 18 framů

Velikost dopředného okna: 4 framy

Úspěšnost detektoru na reprezentativní množině v této fázi : cca 96,5 %.



Obr. 4.7: Změna detekce u framů řeči obklopenými framy šumu

4) vyhlazení náhlých změn v detekci

I po aplikaci 2. a 3. fáze vývoje detektoru detektor nezřídka nesprávně detekoval náhlé změny v detekci. Tento jev byl v tomto kroku vývoje detektoru minimalizován napevno nastaveným minimálním počtem stejně detekovaných framů za sebou.

Minimální počet stejně detekovaných framů za sebou : 10

Úspěšnost detektoru na reprezentativní množině v této fázi : cca 97 %.

Poslední dva experimentálně vyvinuté algoritmy opět pracovaly s dopředným i se zpětným oknem a realizovaly následující snahy (blíže zde kvůli jejich relativní složitosti popisovány nebudou):

5) minimalizování nesprávné detekce řeči u nádechů

Úspěšnost detektoru na reprezentativní množině v této fázi : cca 97,5 %

6) minimalizace nesprávné detekce na začátcích a koncích promluv

Úspěšnost detektoru na reprezentativní množině v této fázi : cca 98 %

K výše popsanému detektoru je třeba uvést, že by byl naprosto nevhodný pro online zpracování dat. Jednak byl realizován v Matlabu, což je interpretační programovací jazyk, a již z toho důvodu by online výpočet trval nepřijatelně dlouhou dobu. A jednak jsou uvedené experimentální algoritmy navzájem neoptimalizované. To

pro účel odhadu SNR v korpusu TULTEL07 nevadilo, protože se v tomto případě jednalo o offline zpracování a doba výpočtu tak nebyla klíčovým faktorem.

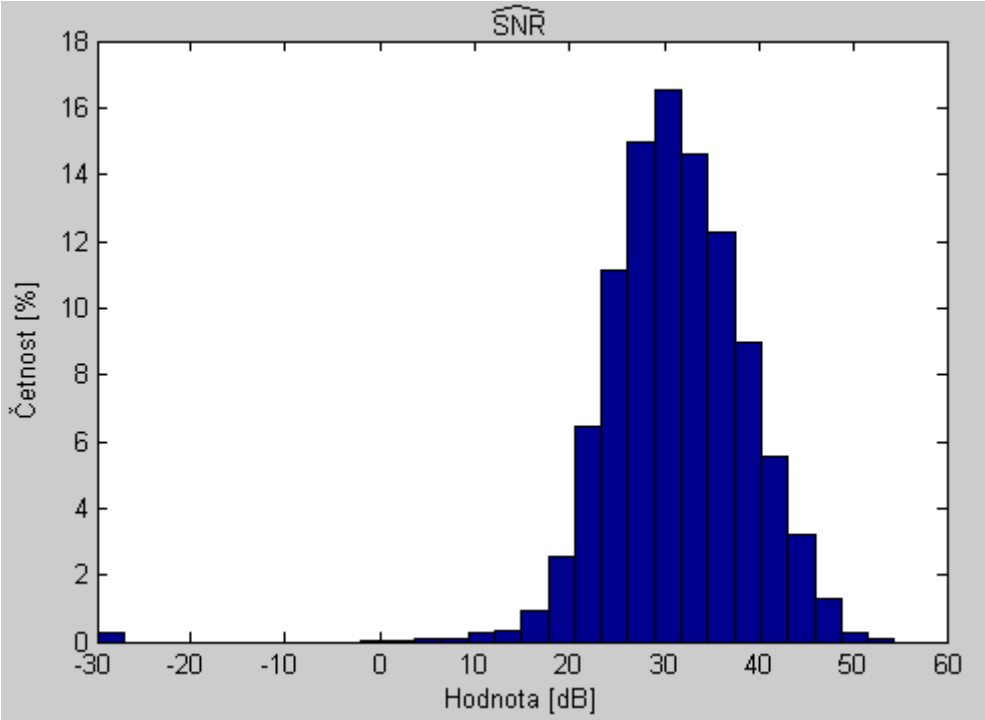
Při implementaci výpočtu globálního SNR s detektorem byla nastavena hodnota SNR v nahrávkách, detekovaných pouze jako řeč, či pouze jako šum, na + 30 dB (řeč) a -30dB (šum). Takové nahrávky se v databázi opravdu vyskytovaly (nahrávky detekované pouze jako šum jsou na Obr. 4.8 viditelné vlevo dole).

Při implementaci výpočtu SNR z histogramů krátkodobého výkonu je třeba v programu zadat, jak velká část framů nahrávky je odhadována jako šum. Ta byla u jednotlivých nahrávek ale dosti proměnlivá mj. i v závislosti na aktuálně čtené položce scénáře hovoru (viz Tab 3.2), kterou nahrávka obsahovala. U krátkých promluv, jako byly například uvítací a loučící pozdravy, byla typicky tato část framů vyšší, než u promluv dlouhých (například části souvislého textu). To bylo dáno mj. snahou anotátorů při segmentaci telefonátů neuseknout začátky a konce jednotlivých promluv a umísťování několika desetin sekundy ticha před i za jednotlivé promluvy v rámci jejich segmentů (viz. 3.3.1). Toto umísťování cca konstantní doby ticha se pak v daném ohledu více projevilo u kratších nahrávek. Tento problém byl vyřešen následovně. Vyvinutým detektorem řeči byly projety všechny analyzované segmenty nahrávek a byla u nich zjištěna procentuální část výskytu šumu. Tyto hodnoty byly poté zprůměrovány a výsledná hodnota považována za reprezentativní pro celou databázi. Tato hodnota byla cca 25% (tj. v průměru cca 25% každé nahrávky neobsahovalo řeč). S tímto odhadem šumu v nahrávce bylo poté vypočteno $\widehat{SNR_h}$ v databázi.

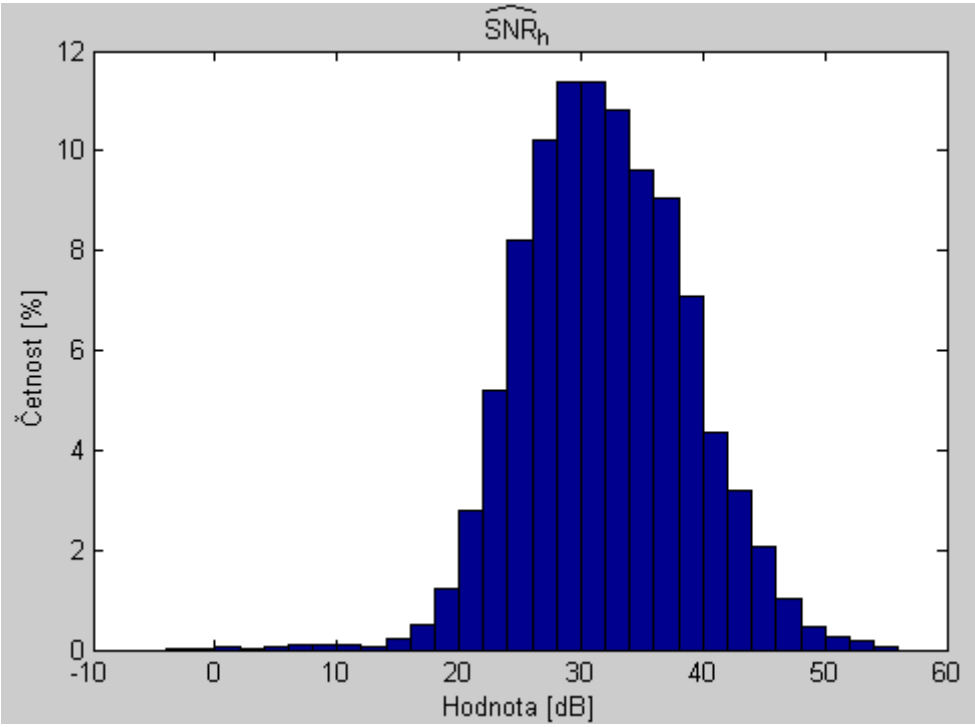
Tab. 4.1: Průměrné hodnoty odhadů SNR v databázi

Algoritmus	Hodnota
$\widehat{SNR}[dB]$	31,33
$\widehat{SNR_h}[dB]$	31,96

Hodnoty odhadovaných SNR napříč všemi řečovými segmenty nahrávek jsou vyjádřeny na Obr. 4.8 a Obr. 4.9.



Obr. 4.8: Histogram odhadu globálního SNR s detektorem řečové aktivity



Obr. 4.9: Histogram odhadu globálního SNR z histogramu krátkodobého výkonu

4.2 Analýza fonetické bohatosti korpusu

Prostřednictvím fonetické anotace (viz 3.3) byl znám fonetický obsah řečových segmentů nahrávek v databázi. Pro účel vyjádření relativní četnosti výskytu fonémů v databázi byl v programovacím jazyce Javě napsán program, který sečetl jednotlivé fonémy obsažené ve všech řečových segmentech nahrávek databáze a vyjádřil jejich relativní četnosti v databázi.

V Tab 4.3 jsou tyto relativní četnosti vyjádřeny společně s relativními četnostmi jednotlivých fonémů v českém jazyce. Údaje o relativní četnosti fonémů v českém jazyce byly čerpány z [Pollák2002]. Tam se lze také dočíst, že tyto údaje byly získány analýzou velkého množství novinových textů a použity pro porovnání výskytu fonémů v jazyce a v databázi SpeechDat.. Porovnání přitom probíhalo za použití fonetické abecedy SAMPA. Ta má ale v porovnání s fonetickou abecedou PAC určité odlišnosti. Jednak obsahuje navíc oproti PAC fonémy "ou" "au" a "eu" a naopak PAC obsahuje navíc oproti SAMPĚ neutrální hlásku označovanou jako schwa (v PAC vyjádřen pomocí "E") a rozlišuje výslovnostní varianty hlásky ř. První varianta hlásky ř (v PAC vyjádřena pomocí "ř") vznikne např. při vyslovení slova moře a druhá varianta (v PAC vyjádřena pomocí "Ř") vznikne např. při vyslovení slova keř.

Při tvorbě korpusu TULTEL07 nebyla fonetická bohatost vytvářeného materiálu uvažována. Tato skutečnost je v Tab 4.3 pozorovatelná např. tím, že výskyt nejméně zastoupených fonémů v jazyce není v korpusu oproti výskytu v jazyce patřičným způsobem nadhodnocen (viz 2.5). Některé fonémy (např. "o", "a" a "i") jsou v korpusu zastoupeny způsobem odpovídajícím foneticky bohatému korpusu a u některých fonémů (např. "e", "d", fonémy nejméně zastoupené v jazyce) je tomu zase naopak. Požadavky na foneticky bohatý korpus uvedené v [Pollák2002] tak u korpusu TULTEL07 splněny nejsou.

Tab. 4.2: Četnost fonémů teoretická a v databázi TULTEL07

SpeechDat SAMPA	PAC	Četnost výskytu v jazyce [%]	Četnost výskytu v databázi [%]
e	e	9,325	10,235
o	o	7,499	7,462
a	a	6,707	6,536
i	i	6,443	6,198
t	t	5,048	5,517
s	s	5,026	5,064
n	n	4,471	4,434
l	l	4,417	3,868
r	r	3,940	3,292
i:	í	3,890	3,557
k	k	3,805	3,073
v	v	3,662	3,778
p	p	3,485	3,390
m	m	3,235	3,166
d	d	2,846	3,959
j	j	2,548	2,891
u	u	2,390	3,333
J	ň	2,287	1,897
a:	á	2,140	2,404
z	z	1,711	1,626
t_s	c	1,579	1,720
b	b	1,574	1,492
h\	h	1,289	1,244
P\	ř Ř	1,243	1,483
e:	é	1,179	1,236
S	š	1,127	1,133
f	f	1,133	0,683
t_S	č	1,000	1,005
x	X	0,974	0,878
c	ť	0,865	0,892
Z	ž	0,753	0,717
u:	ú	0,695	0,609
o_u	-	0,665	-
J\	ď	0,492	0,648
g	g	0,291	0,347
N	N	0,142	0,126
a_u	-	0,060	-
o:	ó	0,035	0,056
e_u	-	0,030	-
d_z	C	0,008	0,023
F	M	0,006	0,003
d_Z	Č	0,004	0,006
-	E	-	0,019

5. Rozpoznávání řeči na korpusu TULTEL07

V rámci porovnání přínosu korpusu TULTEL07 pro Laboratoř počítačového zpracování řeči TUL bylo provedeno testování automatického rozpoznávání řeči. Údaje uvedené v této kapitole byly poskytnuty vedoucím mé bakalářské práce.

Testování proběhlo na 500ti segmentech nahrávek (telefonátů) z korpusu TULTEL07. Tyto segmenty nahrávek se neúčastnily trénování akustického modelu. Rozpoznávání proběhlo rozpoznáváním spojité řeči se slovníkem čítajícím cca 312 000 slov. Tento slovník je Laboratoří počítačového zpracování řeči TUL používán pro přepis televizních a rozhlasových pořadů. I když je cílem výzkumného projektu rozpoznávání běžné spontánní řeči, jemuž se obsah korpusu TULTEL07 příliš nepodobá, hlavním přínosem vytvoření tohoto korpusu je získání akustických dat z telefonního kanálu. Do té doby měla totiž Laboratoř počítačového zpracování řeči TUL k dispozici pouze data převzorkovaná z 16000 Hz na 8000 Hz, která vznikla přepisem televizních a rozhlasových pořadů, či přepisem nahrávek pořízených prostřednictvím mikrofonu a pouze minimum telefonních dat. Proto jsou experimenty v této práci prováděny na segmentech nahrávek charakteru čtené řeči, aby bylo možné jednoznačně porovnat přínos nových akustických modelů.

Zmíněných 500 segmentů nahrávek obsahuje celkem 4725 slov. Při použití původních modelů (převzorkovaná data a minimum dat telefonních) bylo dosaženo rozpoznávacího skóre 48,2%. V případě použití akustických modelů trénovaných na korpusu TULTEL07 se rozpoznávací skóre zvýšilo na 56,7%. Je tedy zřejmé, že došlo k výraznému zvýšení úspěšnosti rozpoznávání. Nicméně v porovnání s úspěšností rozpoznávání řeči u televizních a rozhlasových pořadů (cca 80%) jsou obě uvedené hodnoty poměrně nízké. To mj. dokumentuje obtížnost úlohy rozpoznávání řeči v telefonních nahrávkách.

6. Závěr

V rámci bakalářské práce se autor podílel na vytvoření telefonního řečového korpusu. Pořídil dle požadovaných podmínek (viz 3) celkem 32 telefonních hovorů, které zároveň anotoval. Vyvinul také dva programy, které našly při tvorbě korpusu uplatnění. První z nich (3.1.3) vytvářel scénáře nahrávaných hovorů a druhý (3.3.2) sloužil k zaznamenávání informací o nahrávkách. Oba tyto programy byly vytvořeny v programovacím jazyce Java za použití vývojového prostředí NetBeans IDE..

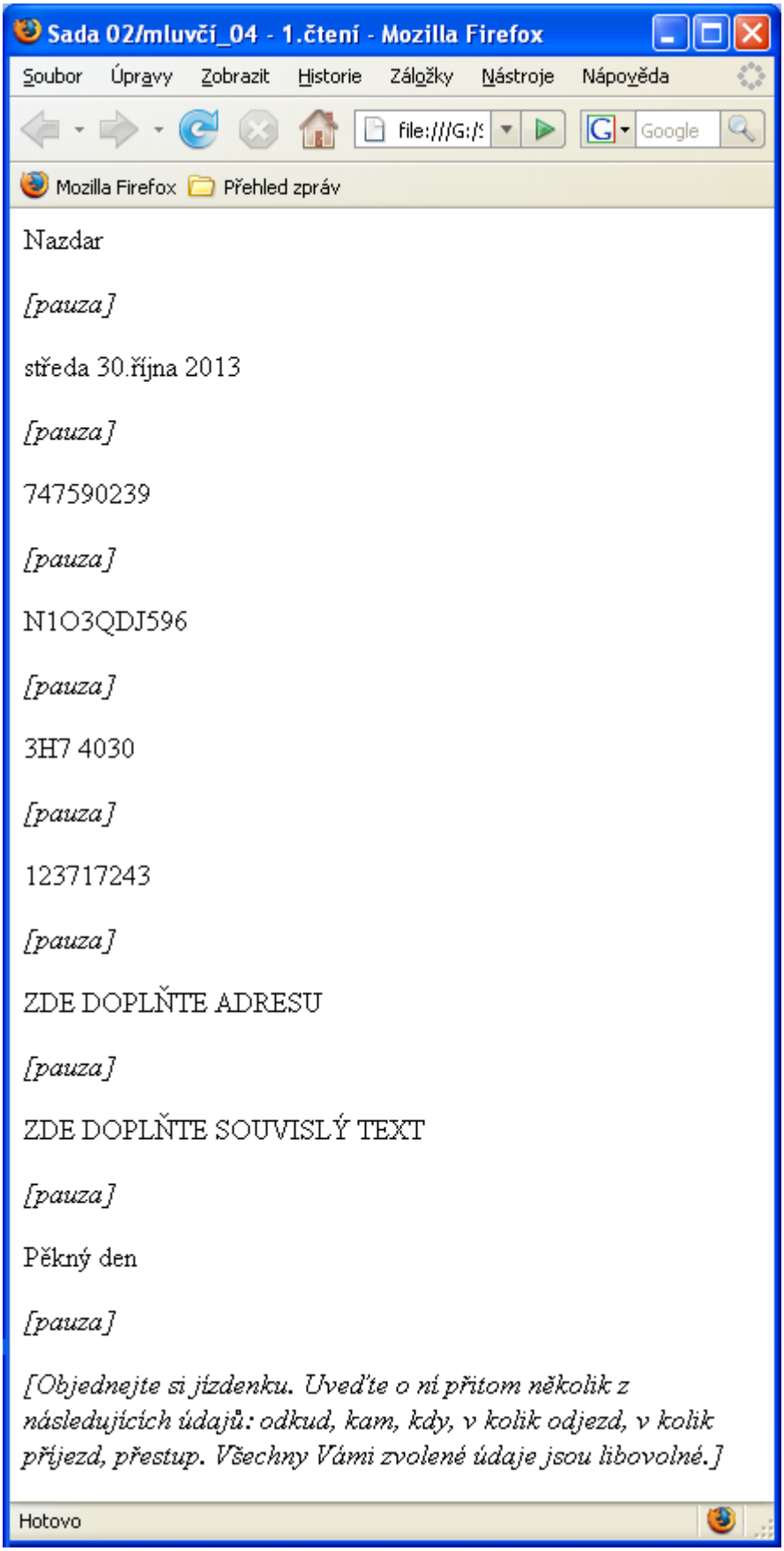
V této práci byly také vyjádřeny číselné charakteristiky řečových segmentů nahrávek popisující jejich kvalitu. Jednalo se o koeficient přebuzení, koeficient vybuzení a odhad odstupu signálu od šumu. Jejich hodnoty byly zjištěny na základě implementace vztahů uvedených v (2.4) v prostředí Matlab. Koeficient přebuzení byl v databázi nenulový pouze u cca 4% segmentů nahrávek. To je poměrně příznivý výsledek. Databáze Speech Dat, jež je rovněž tvořena nahrávkami z telefonního kanálu, měla tento koeficient nenulový u 24% nahrávek [Pollák2002]. Je třeba ale brát v úvahu odlišný charakter těchto nahrávek popsaných v [Pollák2002] v porovnání se segmenty nahrávek korpusu TULTEL07. Průměrné hodnoty odhadu odstupu signálu od šumu v databázi se pohybovaly nad 31 dB, což je velmi vysoká hodnota. Ta je zjevně daná tím, že při mluvení do telefonu je mikrofon velmi blízko ústům mluvčího, a tak je řečový signál v porovnání s případným rušivým pozadím výrazně silnější. V souvislosti s implementací odstupu signálu od šumu byl vyvinut také detektor řečové aktivity. Jeho základem byl energetický detektor s adaptivním prahem. Ten byl za účelem zlepšení detekce rozšířen o 5 algoritmů. K tomuto detektoru je třeba uvést, že je užitečný zejména pro detekci užitečné informace v řeči. V jeho rozšiřujících algoritmech je totiž obsaženo mj. i odstraňování nádechů a výdechů mluvčího z řečové detekce. Pro účely odhadu odstupu signálu od šumu ale nejsou nádechy a výdechy v řeči mluvčích brány jako rušivé pozadí. Tato vlastnost detektoru tak může vést do jisté míry ke zkreslení uvedených hodnot SNR.

V rámci práce byly také vyjádřeny relativní četnosti výskytu fonémů v korpusu.

Výsledky uvedené v (5) ukazují, že po natrénování akustických modelů z korpusu TULTEL07 vzrostlo rozpoznávacímu systému Laboratoře počítačového zpracování řeči TUL výrazně rozpoznávací skóre pro telefonní kanál. Lze ale soudit, že pro natrénování

kvalitních akustických modelů pro telefonní kanál je třeba výrazně vyššího množství dat než je tomu např. v případě modelů pro přepis televizních a rozhlasových pořadů.

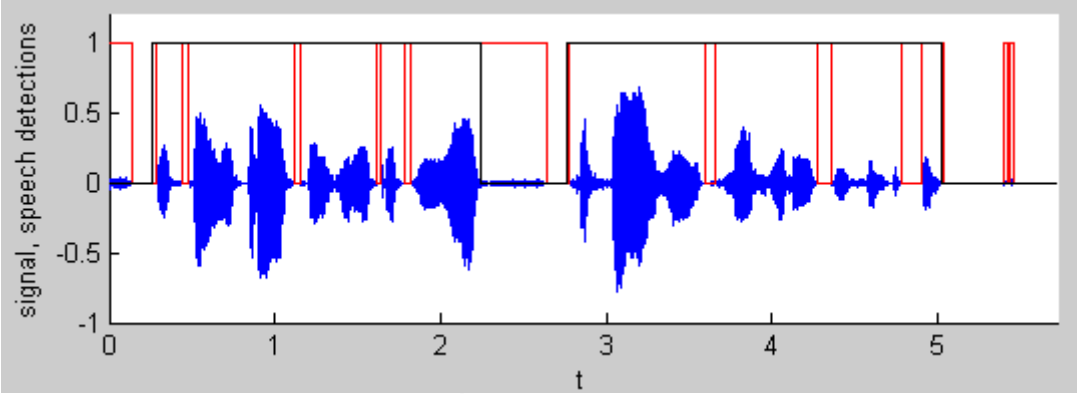
Příloha A - výstup programu na vytváření koster scénářů hovoru



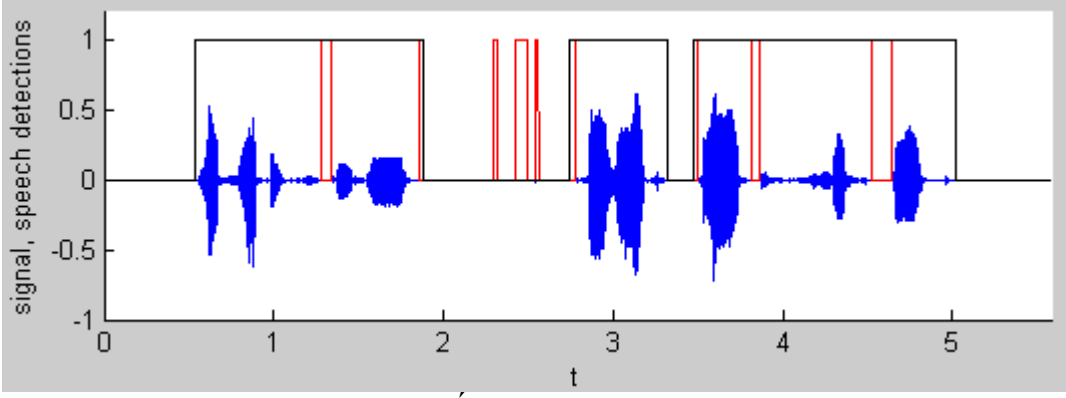
Příloha B - grafické výsledky řečového detektoru v jednotlivých fázích jeho vývoje

Na uvedených obrázcích představuje černá barva referenční detekci řeči a červená barva detekci vyvíjeného detektoru. V místech časové osy, kde je zobrazena pouze referenční detekce, jsou tyto detekce shodné. Jednotlivé fáze vývoje detektoru jsou dokumentovány na příkladech detekce u třech signálů z vytvořeného korpusu.

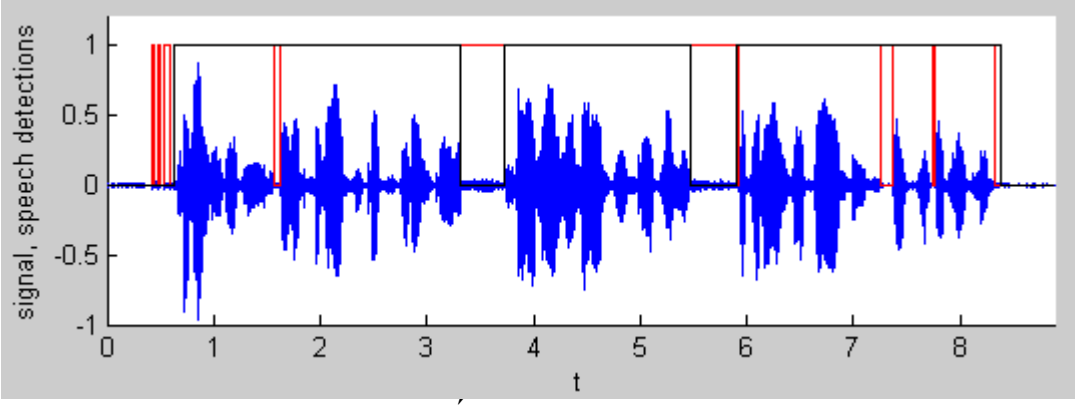
1) Fáze 1 - realizace energetického detektoru s adaptivním prahem



Obr. Detekce řeči na SIGNÁLU 1 – Fáze 1

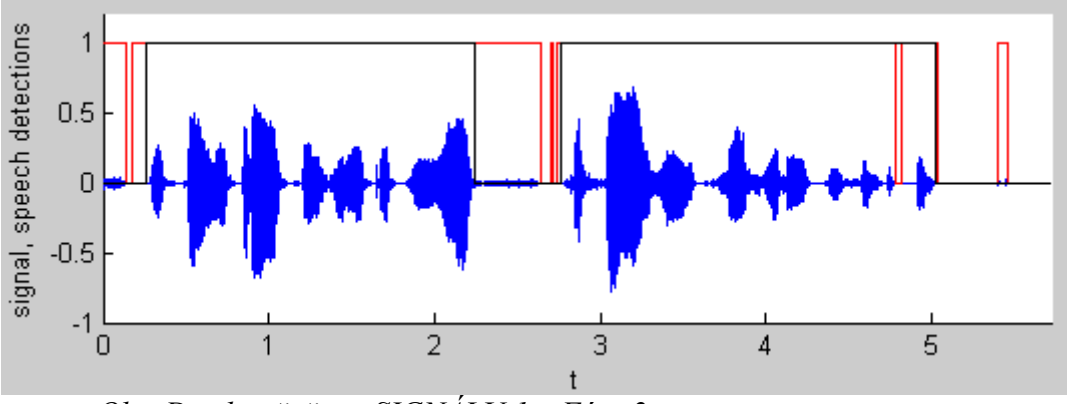


Obr. Detekce řeči na SIGNÁLU 2 – Fáze 1

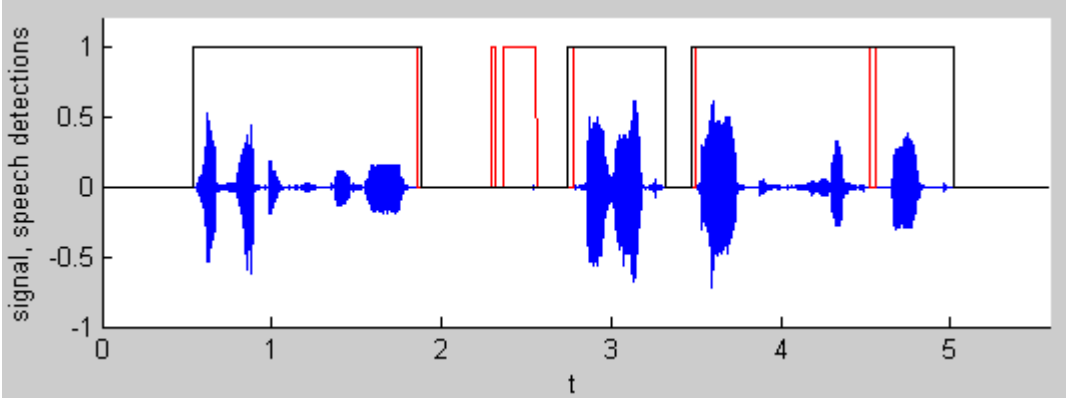


Obr. Detekce řeči na SIGNÁLU 3 – Fáze 1

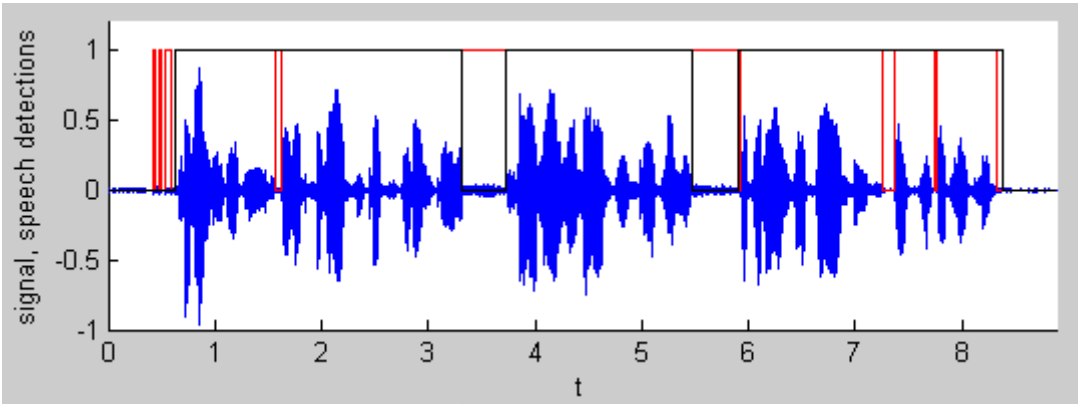
2) Fáze 2 - změna detekce u framů šumu obklopených framy řeči



Obr. Detekce řeči na SIGNÁLU 1 – Fáze 2

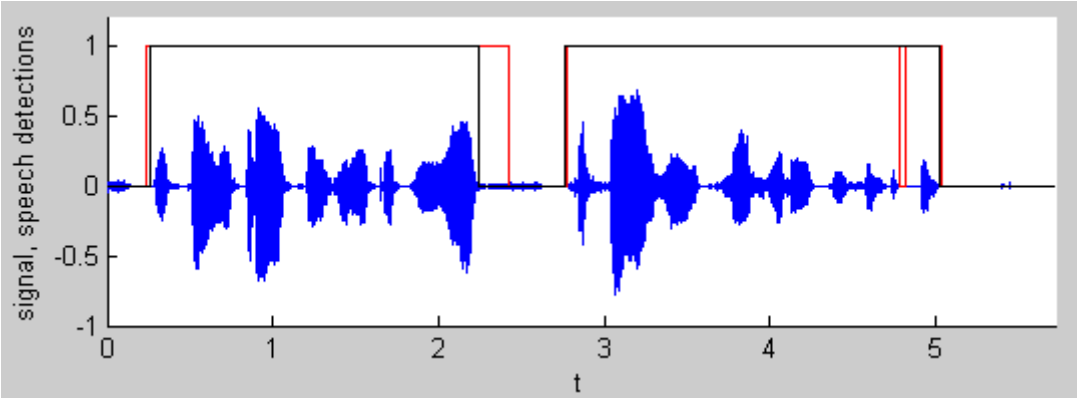


Obr. Detekce řeči na SIGNÁLU 2 – Fáze 2

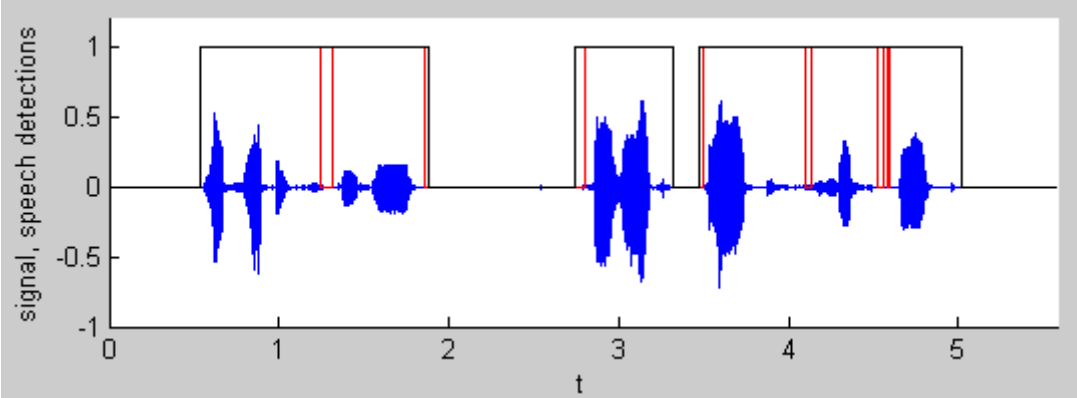


Obr. Detekce řeči na SIGNÁLU 3 – Fáze 2

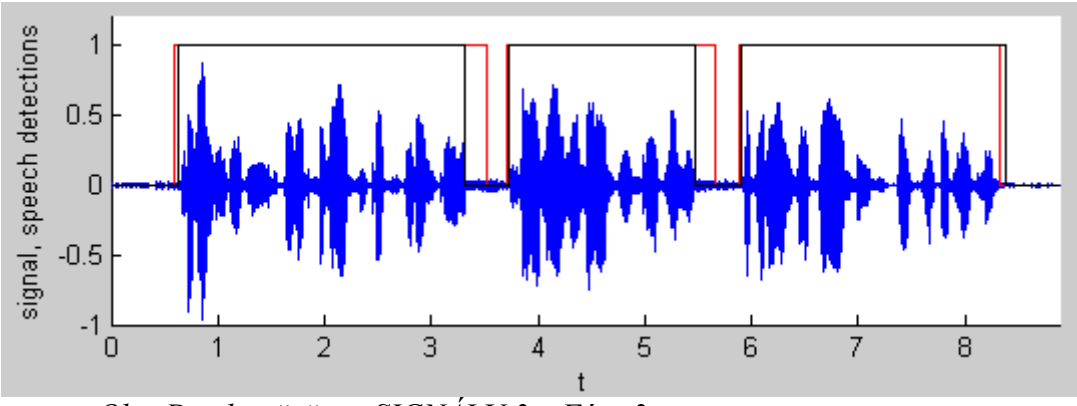
3) Fáze 3 - změna detekce u framů řeči obklopených framy šumu – snaha o základní odstranění nesprávné detekce řeči u nádechů a ostatních ruchů



Obr. Detekce řeči na SIGNÁLU 1 – Fáze 3

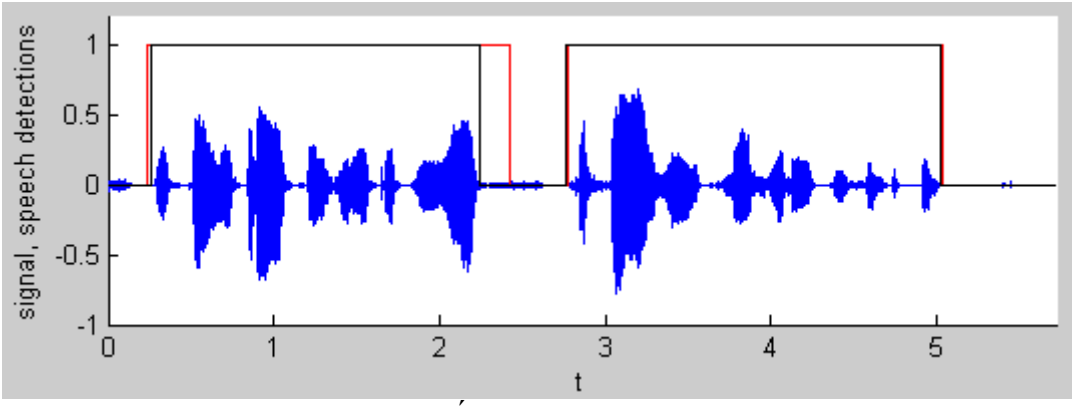


Obr. Detekce řeči na SIGNÁLU 2 – Fáze 3

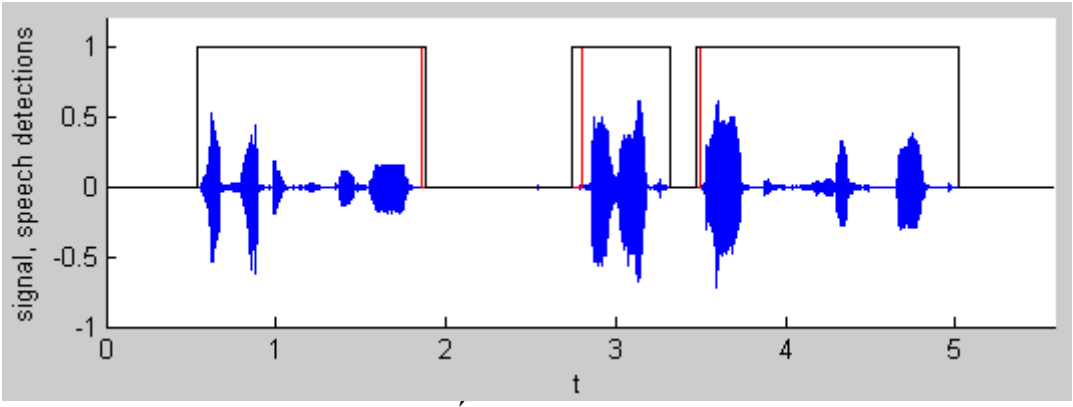


Obr. Detekce řeči na SIGNÁLU 3 – Fáze 3

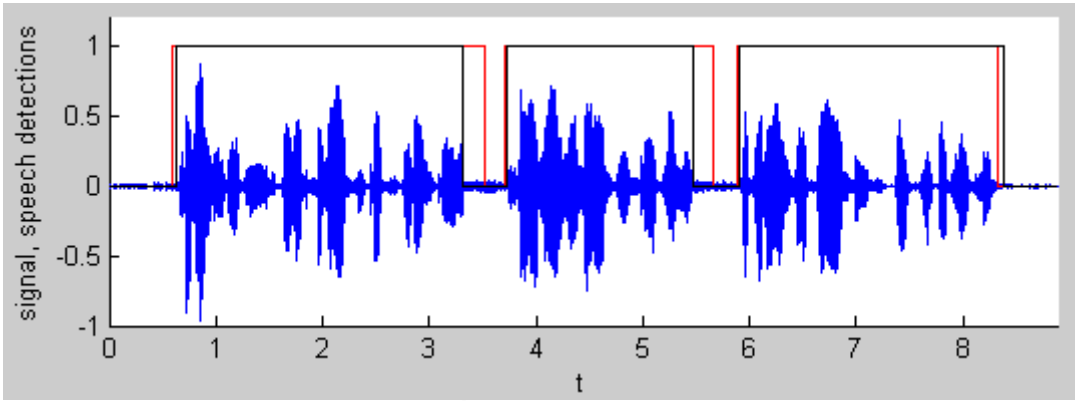
4) Fáze 4 - vyhlazení náhlých změn v detekci



Obr. Detekce řeči na SIGNÁLU 1 – Fáze 4

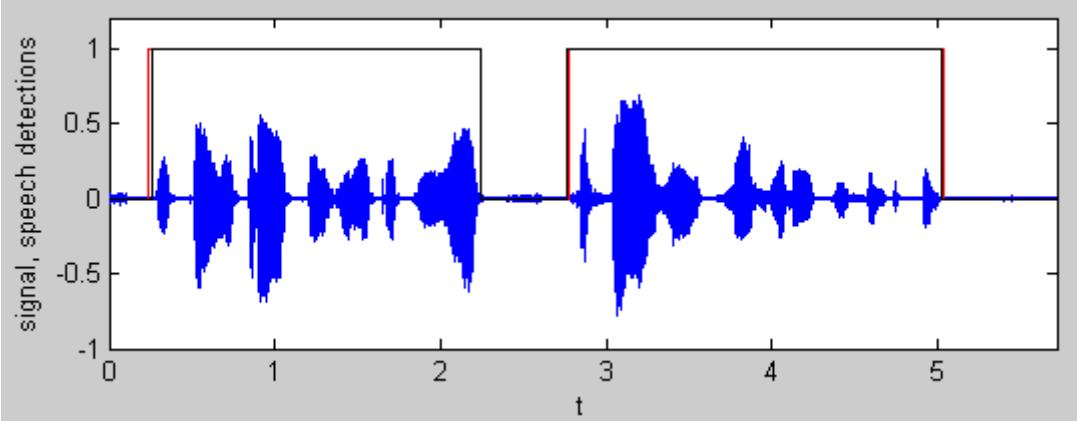


Obr. Detekce řeči na SIGNÁLU 2 – Fáze 4

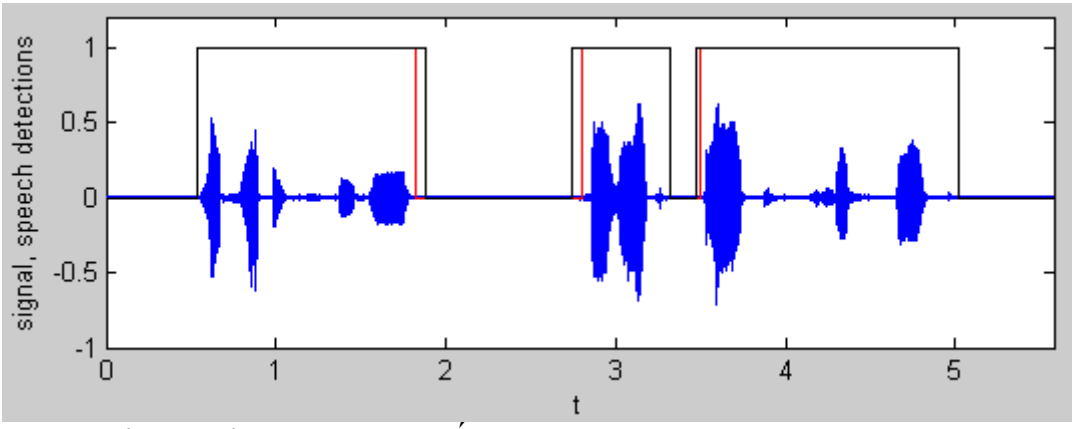


Obr. Detekce řeči na SIGNÁLU 3 – Fáze 4

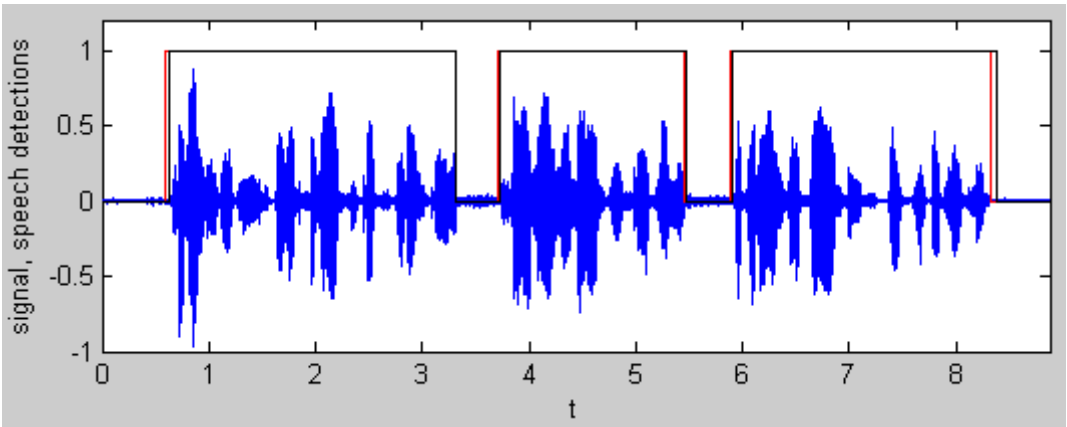
5) Fáze 5 - minimalizování nesprávné detekce řeči u nádechů



Obr. Detekce řeči na SIGNÁLU 1 – Fáze 5

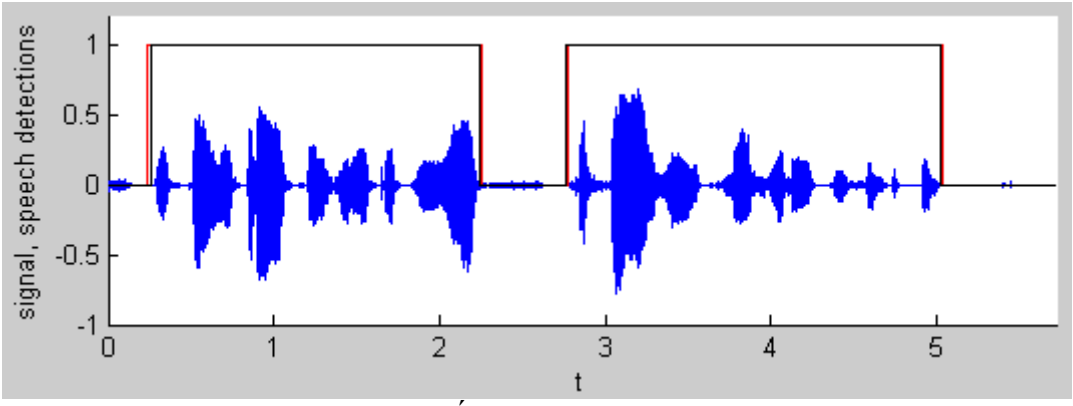


Obr. Detekce řeči na SIGNÁLU 2 – Fáze 5

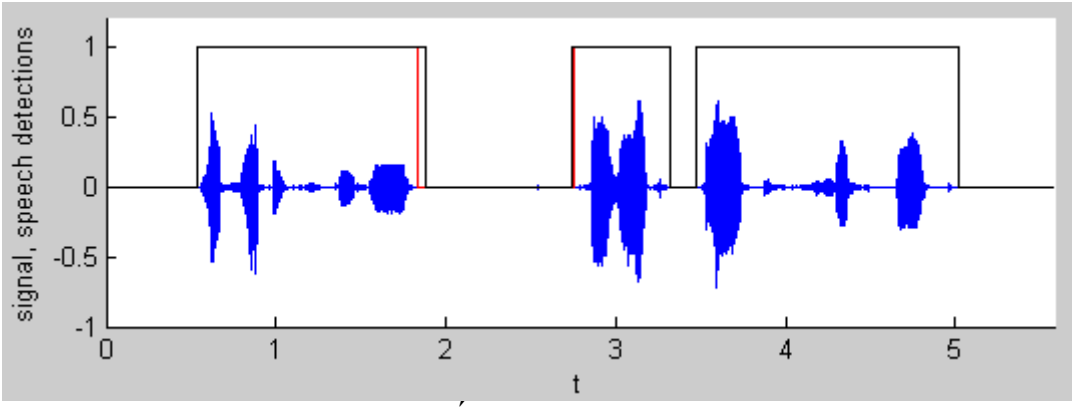


Obr. Detekce řeči na SIGNÁLU 3 – Fáze 5

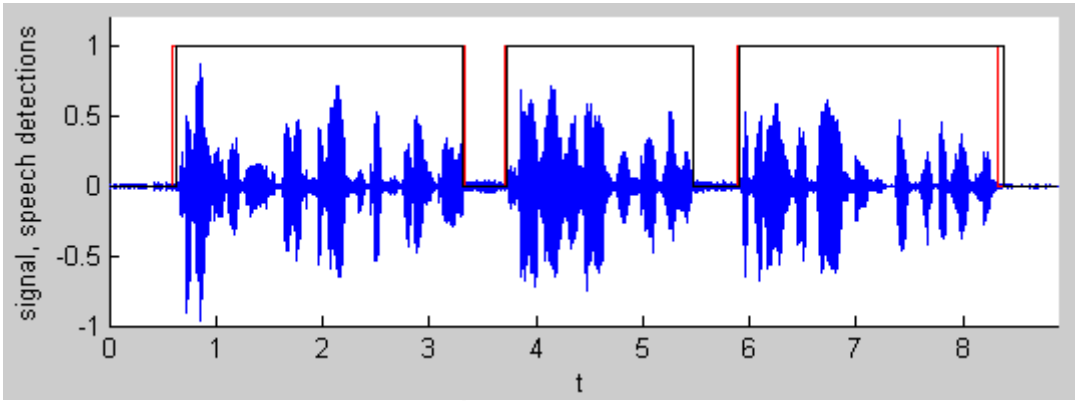
6) Fáze 6 - minimalizace nesprávné detekce na začátcích a koncích promluv



Obr. Detekce řeči na SIGNÁLU 1 – Fáze 6 (konečná detekce)

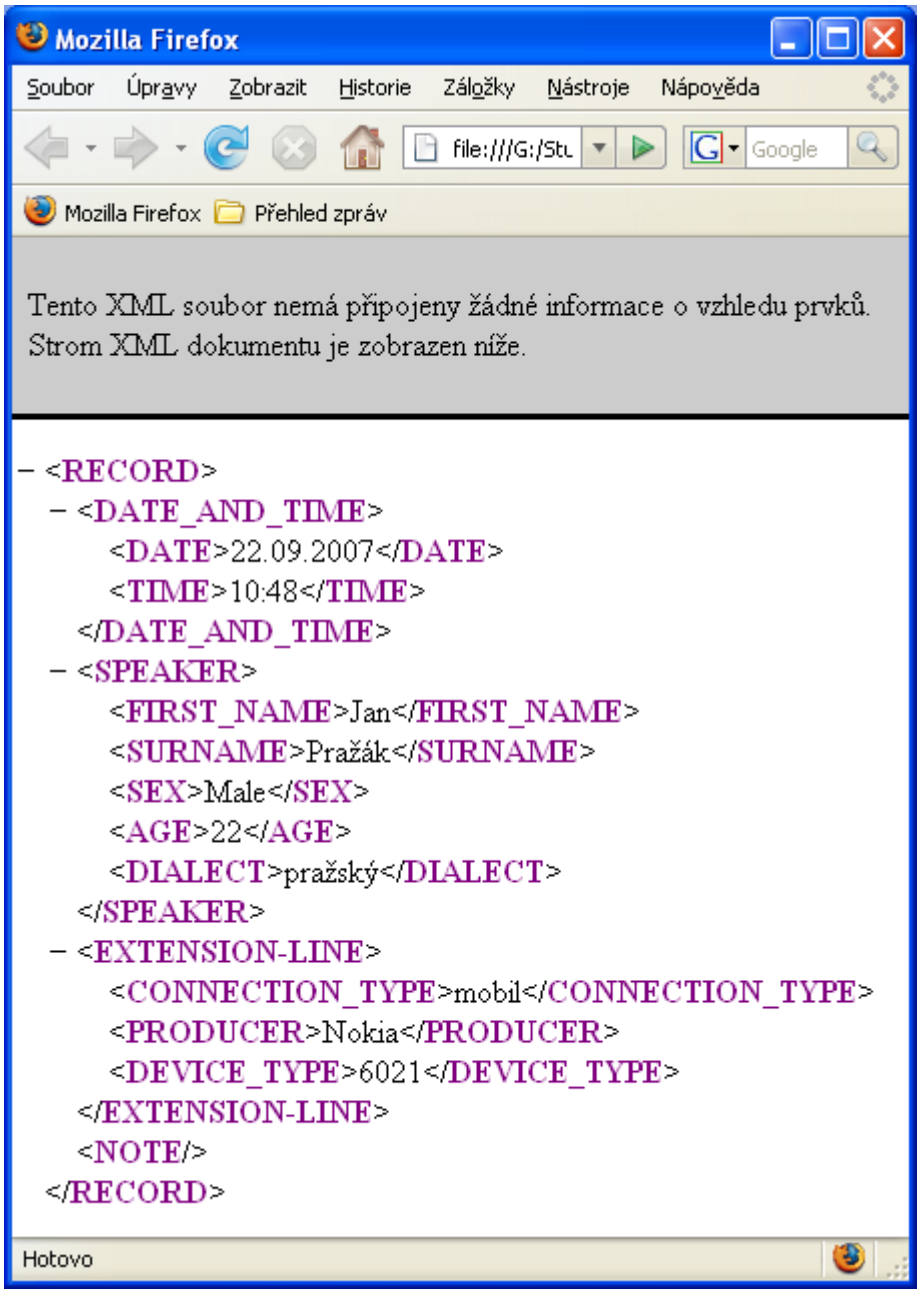


Obr. Detekce řeči na SIGNÁLU 2 – Fáze 6 (konečná detekce)



Obr. Detekce řeči na SIGNÁLU 3 - Fáze 6 (konečná detekce)

Příloha C - výstup programu Dotazník



Seznam použité literatury

[Java2008] Sun Microsystems: The Java™ Tutorials [online], 2008
< http://java.sun.com/docs/books/tutorial/index.html >

[Nouza2001] Nouza, J. (editor): Počítačové zpracování řeči - cíle, problémy, metody a aplikace, 2001, Sborník článků

[Pollák2002] Pollák, P.: Tvorba databází řečových signálů pro účely rozpoznávání a zvýrazňování, červen 2002, Habilitační práce

[Pollák2001] Pollák, P.: Metody odhadu odstupu signálu od šumu v řečovém signálu [online], 2001, Akustické listy 7(3)
< http://noel.feld.cvut.cz/speechlab/publications/019_al01.pdf >

[Radová2004] Radová, V.: Rozpoznávání řečníka, prosinec 2004, Habilitační práce

[SpeechLabTUL2008] Laboratoř počítačového zpracování řeči na Fakultě mechatroniky Technické univerzity v Liberci [online], 2008
< http://itakura.kes.tul.cz/kes/index.html >