VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

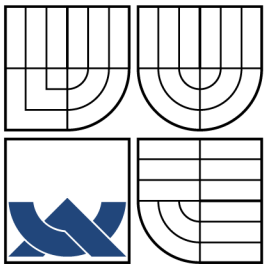
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

MODERNÍ ROZPOZNÁVAČE ŘEČOVÉ AKTIVITY

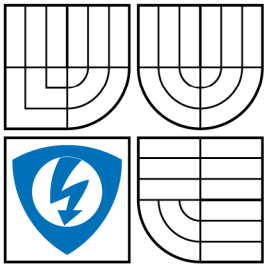
DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE BC. MICHAL ADAMEC
AUTHOR

BRNO 2008



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND
COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

MODERNÍ ROZPOZNÁVAČE ŘEČOVÉ AKTIVITY
MODERN SPEECH/PAUSE DETECTORS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — BC. MICHAL ADAMEC
AUTHOR

VEDOUCÍ PRÁCE — MGR. PAVEL RAJMIC, PH.D.
SUPERVISOR

BRNO 2008

ZDE VLOŽIT LIST ZADÁNÍ

Z důvodu správného číslování stránek

ZDE VLOŽIT PRVNÍ LIST LICENČNÍ SMOUVY

Z důvodu správného číslování stránek

ZDE VLOŽIT DRUHÝ LIST LICENČNÍ SMOUVY

Z důvodu správného číslování stránek

ABSTRAKT

Tato diplomová práce pojednává o standardních metodách detekce řeč/pauza – detektory řečové aktivity pracující na principech krátkodobé energie, reálného kepstra, krátkodobé intenzity a o kombinaci těchto tří detektorů. Dále popisuje detektory založené na skrytých Markovových modelech a detektor řečové aktivity popsaný ve standardu ITU-T G.729. Popsané detektory byly implementovány ve vývojovém prostředí MATLAB. Dále bylo vytvořeno uživatelské rozhraní pro testování funkce implementovaných detektorů. Na základě testování bylo provedeno vyhodnocení pomocí ROC charakteristik.

KLÍČOVÁ SLOVA

Detektory řečové aktivity, krátkodobá energie, krátkodobá intenzita, reálné kepstrum, ITU-T G.729

ABSTRACT

This masters theses deals with standard detection methods of speech/pause – voice activity detectors are based on the principles of short-time energy, real spectrum, short-time intensity and on a combinations of these three detectors. In the next parts, there are mentioned other voice activity detectors based on hidden Markovov's models and a detector described in the ITU-T G.729 standard. All the detectors, mentioned above, were implemented in research environment MATLAB. Further there was created an user interface for testing functions of the implemented detectors. Finally, there was done an evaluation by ROC characteristics according to the results of the testing.

KEYWORDS

Voice activity detectors, short-time energy, short-time intensity, real cepstrum, ITU-T G.729

ADAMEC M. *Moderní rozpoznávače řečové aktivity*. Brno: Vysoké učení technické v Brně. Fakulta elektrotechniky a komunikačních technologií. Ústav telekomunikací, 2008. 75 s. Diplomová práce. Vedoucí práce byl Mgr. Pavel Rajmic, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Moderní rozpoznávače řečové aktivity“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne ....

(podpis autora)

PODĚKOVÁNÍ

Na tomto místě bych chtěl poděkovat vedoucímu mé diplomové práce panu Mgr. Pavlu Rajmicovi, Ph.D. za podmětné připomínky, účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.
Také bych chtěl poděkovat panu Ing. Vojtěchu Stejskalovi, Ph.D. za zájem, věnovaný čas a pomoc při konzultacích.

V Brně dne .....

.....
(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Mnoho systémů, které potlačují nežádoucí aditivní hluk z řečové nahrávky, vyžaduje přesné určení úseků řečové aktivity a úseků řečové pauzy v signálu. O to se starají detektory řečové aktivity (Voice Activity Detectors – VAD). VAD se využívají při zpracování řeči a to např. v algoritmech zvýrazňování řeči a při rozpoznávání řeči. Dále jsou důležitou součástí některých zařízení v oblasti komunikací, kde slouží ke snížení přenosové kapacity tím, že se nepřenášejí pakety neobsahující řeč, např. VoIP (Voice over Internet Protokol).

V této diplomové práci se budeme v kapitole o mluvené řeči zabývat stručným popisem teorie řeči, sem patří například tvorba řeči a akustické vnímání řeči. Dále potom v kapitole 2 probereme úpravu řečového signálu , kde budou popsány úkony, které je nutno provést před samotným zpracováním řeči. Stručně zde bude popsán proces digitalizace, preemfáze a segmentace řečového signálu pomocí oken. V další kapitole se zaměříme na popis parametrů řečového signálu, jako je energie signálu, počet průchodů nulou a kepstrální analýza řečového signálu. Jádrem práce je kapitola 3 pojednávající o detektorech řečové aktivity. Ukážeme si základní rozdělení některých detektorů řečové aktivity a popis činnosti základních principů (energetický detektor, kepstrální detektor, detektor intenzity, detektory založené na prahování HMM pravděpodobností, VAD definován v ITU-T G.729).

V praktické části této diplomové práce budou v MATLABu implementovány tyto rozpoznávače řečové aktivity: detektor podle standardu ITU-T G.729, energetický detektor, detektor intenzity, kepstrální detektor, a kombinovaný detektor vytvořený kombinací tří detektorů (energetického, kepstrálního a detektoru intenzity). Vlastní přínos této diplomové práce začíná od kapitoly 6 popisující kombinovaný detektor.

Pro testování účinnosti rozpoznávačů bude vytvořeno grafické rozhraní v MATLABu. Výsledky testování budou vyhodnoceny pomocí křivek ROC.

1 MLUVENÁ ŘEČ

1.1 Informační obsah řeči

Za nejmenší jednotku řeči, která může tvořit jednotlivá slova, lze považovat foném. Fonémy lze od sebe odlišit například podle způsobu a místa tvoření, podle artikulujícího orgánu nebo sluchového dojmu. Fonologické výzkumy ukázaly, že v existujících světových jazycích je aktivně využíváno jen asi dvanáct jakýchsi univerzálních diferenciálních příznaků [2]. Tento fakt lze objasnit fyziologickými zvláštnostmi hlasového ústrojí člověka, který je schopen svými artikulačními orgány vytvořit kolem dvanácti odlišných poloh při vytváření různých mluvených výrazů. Spojením současně vytvořených diferenciálních příznaků se dosahuje fonému.

Počet fonémů v existujících světových jazycích se pohybuje od 12 do 60 (v českém jazyce 36). Spojením fonémů získáváme slabiku. Promluvy jsou potom již libovolným skládáním různých posloupností slabik.

1.2 Informace v mluvené řeči

Uvažujme zdroj, který může produkovat s nezávislých elementů x_i, které jsou prvky souboru $X = [x_1, \ldots, x_s]$ a jejichž pravděpodobnost výskytu je $p(x_i)$. Abeceda prvků X tvoří úplnou soustavu, takže platí:

$$\sum_{i=1}^{s} p(x_i) = 1. \tag{1.1}$$

Potom informační obsah jednoho prvku x_i bude:

$$l(x_i) = -\log_2 p(x_i) \qquad [\text{bit}] \tag{1.2}$$

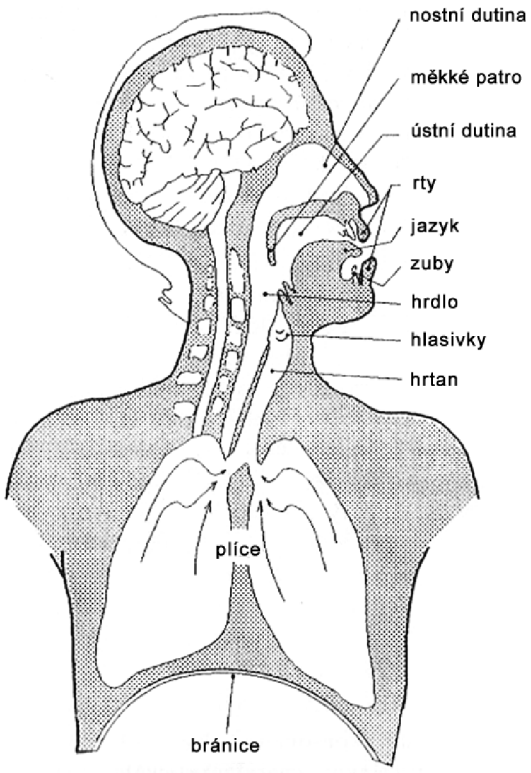
a entropie zdroje:

$$H(x_i) = -\sum_{i=1}^{s} p(x_i) \log_2 p(x_i) \qquad [\text{bit}]. \tag{1.3}$$

Budeme-li předpokládat, že pravděpodobnost výskytu všech fonémů je stejná, potom platí:

$$p(x_i) = \frac{1}{\text{počet všech fonémů}}. \tag{1.4}$$

Výsledky psychoakustických testů udávají, že člověk je schopen zpracovat mluvenou informaci o rychlosti maximálně 50 bit/s [2]. Při běžném hovoru člověk vysloví asi 80 – 130 slov za minutu, což představuje frekvenci výskytu asi 10 fonémů za sekundu. Jestliže uvážíme průměrnou informaci na jeden foném H = 3 – 4 bit, dostaneme pro mluvenou řeč rychlost přenosu informace asi 30 – 40 bit/s. To je tedy informační obsah řeči objevující se v její fonetické struktuře.



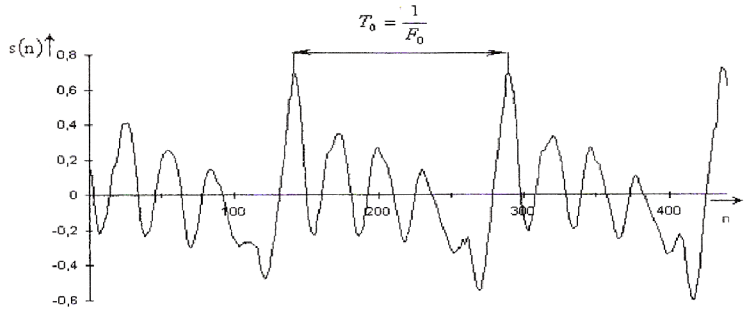
Obr. 1.1: Řečové orgány člověka. Obrázek byl převzat z [2].

1.3 Tvorba řeči

Lidské řečové orgány se skládají z hlasivek, dutiny hrdelní, ústní a nosní, měkkého a tvrdého patra, zubů a jazyka. Zdrojem hlasové energie jsou plíce a s nimi jsou funkčně spjaty dýchací svaly obr. 1.1.

Zdrojem všech znělých zvuků jsou kmitající hlasivky, které jsou umístěny v horní části hrtanu. Vzduch dodávaný plícemi prochází prostorem mezi hlasivkami, tzv. hlasivkovou štěrbinou (latinsky glottis). Hlasivky se rozkmitávají (uzavírají a otevírají hlasivkovou štěrbinu), a tím přeměňují proud vzduchu na pravidelný budící signál. Při kmitání hlasivek vznikají vzduchové rázy v intervalech přibližně 10 ms [5]. Frekvence kmitů závisí jednak na tlaku vzduchu a jednak na svalovém napětí hlasivek. Frekvence kmitů hlasivek F_0 charakterizuje **základní tón lidského hlasu** (část 1.4).

Jestliže člověk mlčí, pak chrupavky drží hlasivkovou štěrbinu odkrytou, takže jí může bez odporu procházet vzduch k dýchání [2].



Obr. 1.2: Ukázka periodické části signálu znělého úseku řeči (f_{vz}= 8 kHz). Obrázek byl převzat z [5].

1.4 Kmitočet základního tónu řeči

Základní tón řečového signálu vyjadřuje kmitání hlasivek při vytváření znělých úseků řeči. Kmitočet základního tónu řeči T_0 (pitch period) můžeme ve vybraných segmentech řeči snadno pozorovat již při pohledu na detailní časový průběh zaznamenané řeči obr. 1.2.

Základní tón řeči můžeme považovat za hlavní parametr, který určuje charakteristiku mluvčího. Jeho střední hodnotu vnímáme jako výšku hlasu, jeho změny v promluvě spojujeme s dynamikou osobnosti mluvčího. Základní tón je důležitou součástí v procesu rozpoznávání mluvčího poslechem.

Hodnoty F_0 leží obvykle v intervalu 50 – 400 Hz:

- dětský hlas má $F_0 = 300 - 400$ Hz
- mužský hlas má průměrný základní kmitočet $F_0 = 120$ Hz
- ženský hlas má průměrný základní kmitočet $F_0 = 210$ Hz

Střední hodnota základního tónu klidné řeči (při vyrovnaném pnutí v hlasivkách) je v nepřímé úměře s délkou membrány hlasivek L_m.

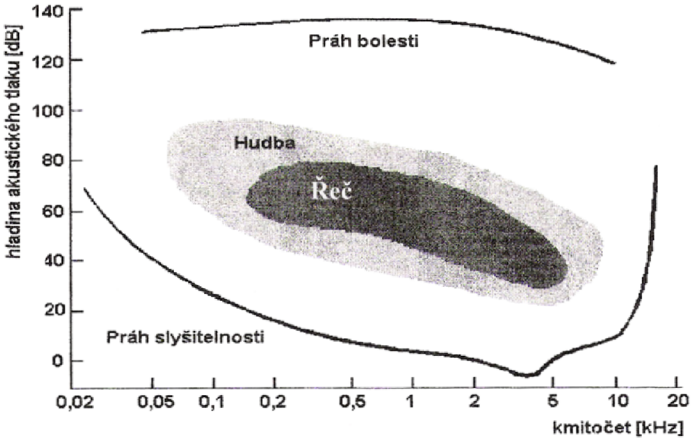
$$F_0 = 1700/L_m \qquad [\text{Hz}]. \tag{1.5}$$

Znamená to například, že základní tón u dospělé ženy $F_0 = 170$ Hz odpovídá délce membrány $L_m = 10$ mm [5].

Výška tónu plní v některých jazycích důležitou funkci. Ve většině základních jazyků nepřímo přispívá k rozpoznávání slov. Například přenáší dodatečné užitečné informace o tom, která slova jsou ve větě zdůrazněna, jestli je věta otázkou, zjištěním nebo příkazem, pocity mluvčího atd. . V některých východních jazycích, jako je například čínština, slouží změna výšky tónu k rozlišení významu foneticky podobných slabik.

1.5 Akustické vnímání řeči

Člověk je schopen vnímat zvuky v rozsahu kmitočtů od cca 20 Hz do 20 kHz a v rozsahu intenzity cca 0 až 130 dB (tzv. plocha slyšení). Normální řeč však leží v podstatně užším pásmu - přibližně mezi 180 Hz až 6 kHz a 30 až 80 dB [5]. Hranice vymezení plochy řeči tvoří nepravidelný obrazec, jak je vidět na obr. 1.3. Pro věrnou reprodukci hlasu se počítá s pásmem do 12 kHz. Dobrá srozumitelnost (i cizích slov) se dosáhne při přenosu řeči v pásmu 300 - 5000 Hz. Pro nejběžnější konverzaci stačí dokonce přenos v pásmu 300 - 2500 Hz.



Obr. 1.3: Plocha slyšení akustických signálů. Obrázek byl převzat z [5].

Intenzita některých typických zvuků, které se mohou vyskytovat jako hluk pozadí:

140 dB	start letadla
120 dB	rockový koncert
100 dB	křik z blízka
80 dB	rušná ulice
70 dB	normální řeč
50 dB	tichá řeč
30 dB	jemné šeptání
20 dB	venkovská krajina v noci
6 dB	práh slyšitelnosti při 1 kHz
0 dB	referenční úroveň

Člověk je schopen vnímat pouze intenzitu a kmitočet, zatímco změny fáze nepostřehne, neboť lidské ucho je necitlivé na fázové poměry.

Hlasitost je míra subjektivního vjemu intenzity zvuku. Vztah mezi hlasitostí H a intenzitou zvuku I na kmitočtu a pro tóny o kmitočtu 1000 Hz je dán vztahem:

$$H = 10 \log \frac{I}{I_0} \qquad [\text{dB}]. \tag{1.6}$$

Referenční prahová intenzita je mezinárodně stanovená na hodnotu $I_0 = 10^{-12} Wm^{-2}$. Hlasitost tónů jiných kmitočtů se stanovuje subjektivním porovnáváním s tónem 1000 Hz.

2 ÚPRAVA SIGNÁLU PŘED ZPRACOVÁNÍM

Řečový signál se ve všech oblastech zpracování řeči zpracovává v číslicové podobě.

2.1 Předzpracování analogového signálu

V prvním kroku je třeba převést signál z akustické podoby do elektrické. Znamená to, že převádíme změny akustického tlaku na elektrický signál. Vhodným zařízením pro převod je mikrofon, který zaručí velmi dobrý poměr signál/šum (řeč/zvuk v pozadí).

Signál, který tímto způsobem získáme, je zpravidla v rozsahu několika milivoltů a musí být, pokud možno bez šumu a s lineární kmitočtovou závislostí v pracovním pásmu kmitočtů, zesílen [5]. Analogový řečový signál $s(t)$ je nutné omezit dolní propustí s ohledem na následné vzorkování. Musí být splněn Shannonův vzorkovací teorém, který říká, že vzorkovací kmitočet musí být minimálně dvakrát větší, než maximální kmitočtová složka spektra původního spojitého signálu.

$$f_{vz} \geq 2f_m. \tag{2.1}$$

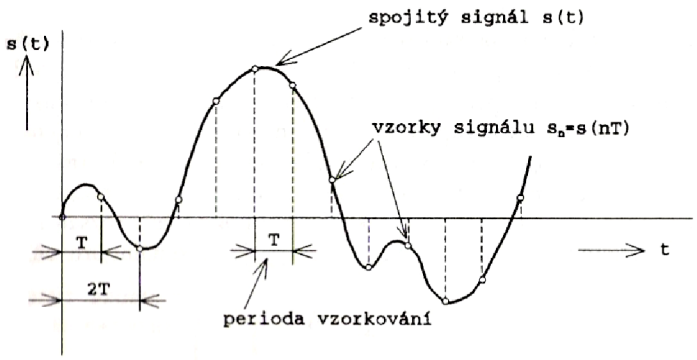
Z toho plyne, že vzorkovací kmitočet musí být minimálně dvakrát vyšší, než je mezní kmitočet dolní propusti f_{DP}.

Mezní kmitočet se často doporučuje volit podle některého standardu CCITT (*Comité Consultatif International Télégraphique et Téléphonique*), např.

$f_{\mathrm{DP}} = 3{,}4$ kHz (CCITT G.711)
$f_{\mathrm{DP}} = 7{,}0$ kHz (CCITT G.722)

Nízké kmitočty signálu pod 200 Hz obsahují málo důležitých informací (řeč je také bez těchto kmitočtů dobře srozumitelná). V těchto nízkých kmitočtech se také nachází velmi mnoho rušení, např. síťové rušení 50 Hz, rušení od mluvčích (dýchání) a rušení od mikrofonu. Filtrováním horní propustí s mezním kmitočtem cca 200 Hz můžeme tato rušení odstranit, aniž by se významně změnil užitečný signál [5].

Střední úroveň řečového signálu se při normální řeči mění obvykle o několik decibelů v časovém rozmezí několika sekund. Změnou polohy mikrofonu a úst mluvčího lze způsobit opět rozdíl několika decibelů. Mnoho parametrů signálu je závislých na kolísání hlasitosti. Protože tyto efekty nemají fonetický význam, je žádoucí vyrovnávat celkovou intenzitu řečového signálu hned na počátku zpracování ještě v analogové podobě.



Obr. 2.1: Znázornění procesu vzorkování. Obrázek byl převzat z [2].

2.2 Digitalizace

Takto analogově předzpracovaný řečový signál budeme nyní digitalizovat. Proces digitalizace zahrnuje provedení těchto kroků: vzorkování, kvantizaci a kódování.

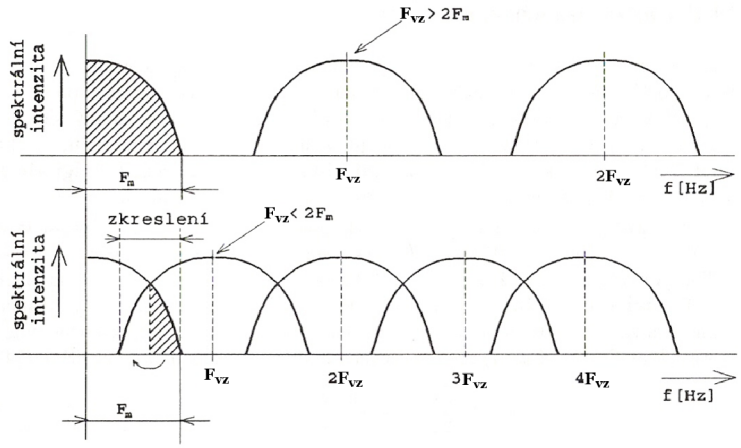
2.2.1 Vzorkování

Vzorkování je transformace signálu $s(t)$ spojitého v čase, na posloupnost vzorků $s_n = s(nT)$ diskrétních v čase. Toto vzorkování probíhá v časových okamžicích $t_n = nT$, kde T je perioda vzorkování a $n = 0, ..., \infty$ [2]. Proces vzorkování je znázorněn na obr. 2.1. Při vzorkování nesmíme zapomenout na Shannonův vzorkovací teorém (2.1). Je-li analogový signál $s(t)$ frekvenčně omezen na pásmo 0 až F_m [Hz], lze $s(t)$ rekonstruovat z hodnot vzorků $s(nT)$ podle vztahu

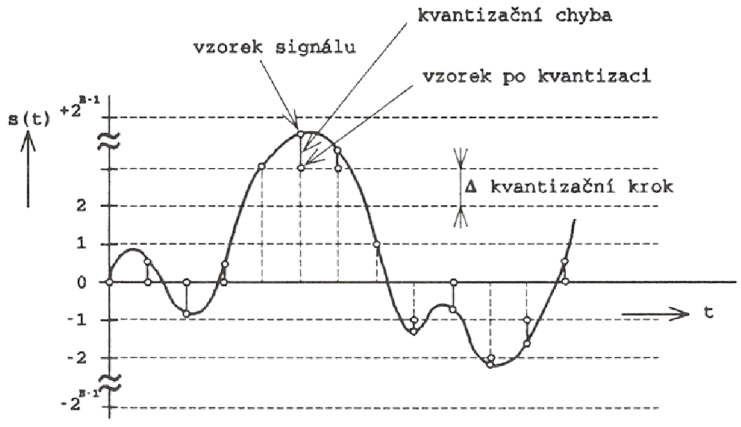
$$s(t) = \sum_{n=-\infty}^{\infty} s(nT)[\frac{\sin \pi(t/T - n)}{\pi(t/T - n)}], \tag{2.2}$$

přitom $s(nT)$ jsou vzorky signálu $s(t)$ v čase $t = nT$, $n = 0, ..., \infty$ a T je perioda vzorkování. Je nutné, aby na hranici frekvenčního pásma F_m již byla energie signálu o více než 30 až 40 dB nižší než střední úroveň signálu v jeho hlavním frekvenčním pásmu. Není-li hodnota F_m jasně definována (v signálu se nacházejí ne dostatečně potlačené složky s vyšší frekvencí), je třeba před vzorkováním signál filtrovat dolní propustí s dostatečnou strmostí [2].

Při porušení vzorkovacího teorému dochází ke zkreslení složek vyšších frekvencí. Důsledek porušení vzorkovacího teorému je znázorněn na obr. 2.2.



Obr. 2.2: Znázornění vlivu frekvence vzorkování na rozložení spektrální intenzity. Obrázek byl převzat z [2].



Obr. 2.3: Proces kvantizace. Obrázek byl převzat z [2].

2.2.2 Kvantizace a kódování

Kvantizace s následujícím kódováním je aproximace analogové hodnoty vzorku signálu jednou z konečného počtu číselných hodnot. Většinou je prováděna převodníkem A/D, který přijímá vstupní analogové napětí (vytvořené vzorkovačem) a jako výstup generuje odpovídající kódovou reprezentaci.

Počet kvantizačních úrovní se obvykle volí ve tvaru 2^B (B je počet kvantizačních bitů v binárním kódu). Pro návrh kvantizéru je postačující udat počet úrovní kvantování a kvantizační krok Δ. Parametry Δ a B se vybírají tak, aby byl pokryt celý rozsah signálu [2]. Za předpokladu, že $|s(nT)| \leq S_{\max}$, kde $S_{\max}$ je maximální úroveň vzorkovaného signálu, dostaneme

$$2S_{\max} = \Delta 2^B. \tag{2.3}$$

Protože dynamický rozsah řečového signálu je asi 60 dB, je pro jeho kvalitní převod zapotřebí $B = 11{-}12$ bitů [5].

Při kvantizačním procesu dochází k určité ztrátě informace vlivem zaokrouhlování okamžitých velikostí signálu. Tato ztráta se nazývá **kvantizační zkreslení** nebo též kvantizační šum. Odstup signálu od kvantizačního šumu je pro B-bitový převod dán vztahem

$$\mathrm{SNR} = 6B - 7{,}24. \tag{2.4}$$

2.3 Preemfáze

Podstatná část celkové energie řečového signálu (u některých mluvčích více než polovina) leží v kmitočtovém pásmu pod hranicí 300 Hz, ačkoli užitečné informace v signálu jsou téměř kompletně obsaženy v pásmu nad 300 Hz. Jelikož kvantizační šum vykazuje rovnoměrné spektrum, je jeho negativní vliv podstatně větší na energeticky slabší, ale důležitější vyšší složky spektra řečového signálu. U znělých zvuků navíc obvykle první formant energeticky silně převyšuje ostatní formanty.

Uvedené efekty lze částečně zmírnit filtrací řečového signálu číslicovým filtrem s charakteristikou horní propusti:

$$H(z) = 1 - \lambda z^{-1}. \tag{2.5}$$

Tato filtrace se provádí před vážením rámce. Vede ke zdůraznění vyšších kmitočtů a nazývá se preemfáze. V časové oblasti vše vyjadřuje vztah:

$$s''(n) = s'(n) - \lambda s'(n-1), \tag{2.6}$$

kde s' je hodnota řečového signálu před preemfází a s'' je hodnota řečového signálu po preemfázi.

Koeficient preemfáze λ leží obvykle v intervalu (0,9 až 1,0). Někdy je vhodné použít adaptivní preemfázi, při které se λ mění s časem podle podílu prvních dvou autokorelačních koeficientů

$$\lambda = R(1)/R(0). \tag{2.7}$$

2.4 Segmentace pomocí oken

Řečový signál je téměř výhradně zpracováván metodami tzv. krátkodobé analýzy. Tyto metody vycházejí z předpokladu, že vlastnosti signálu se v čase mění „pomalu“. Signál je za tím účelem rozdělen na rámce (segmenty) o délce N vzorků a každý rámec je potom popsán vektorem příznaků.

Délka rámce musí být na jedné straně dostatečně malá, aby bylo možné naměřené parametry uvnitř rámce aproximovat konstantními hodnotami a na druhé straně dostatečně velká, aby bylo zaručeno, že požadované parametry budou bezchybně změřeny. Protože změny nastavení lidského hlasového ústrojí probíhají v nejkratším intervalu 10 až 25 ms, jsou oba požadavky splněny pro úseky právě tak dlouhé.

Při segmentaci je slovo rozděleno na n rámců (segmentů), přičemž všechny rámce mají stejnou délku odpovídající N vzorkům. Dva sousední rámce se mohou překrývat. Částečným překrýváním rámců se dosáhne většího vyhlazení časových průběhů parametrů signálu, ale zpomalí se časový posun a částečně se zvýší výpočetní nároky.

Řečový segment $s(n)$ o N vzorcích může být vytvořen z řečového signálu po preemfázi $s''(n)$ pomocí váhové posloupnosti tzv. okna $w(n)$, kterým se vybírají (váží) vzorky $s''(n)$. Provedeme tak násobením:

$$s(n) = s''(n)w(n). \tag{2.8}$$

Pomocí okna vybíráme příslušné vzorky signálu a přidělíme jim určitou váhu. Váhovou funkci $w(n)$ určuje typ okna. Při zpracování řečového signálu používáme nejčastěji okno typu: „pravoúhlé“ (část 11.1) a „Hammingovo“ (část 11.1).

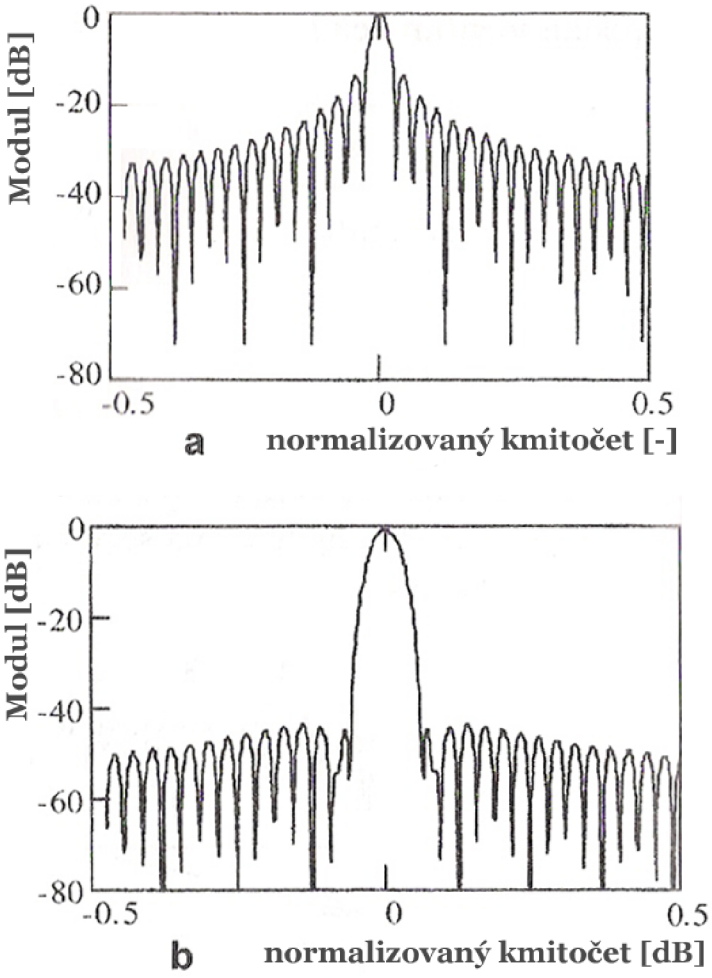
Délka okna N a tím současně délka vybraného rámce řeči je vyjádřena v počtu vzorků. Zvolené okno se pohybuje po časové ose s krokem N vzorků v případě, že rámce na sebe navazují nebo s krokem menším než N vzorků, pokud se rámce překrývají.

Okénkové funkce v podstatě představují filtr typu dolní propust. Spektrum vybraného rámce získané Fourierovou transformací reprezentuje výsledek konvoluce

skutečného spektra daného úseku řečového signálu se spektrem použité okénkové funkce. Při násobení řečového signálu pravoúhlým oknem vznikají dva nežádoucí efekty – rozmazání a rozptyl spektra. Jsou způsobeny tím, že spektrum pravoúhlého okna (obr. 11.2 a) je tvořeno jedním hlavním lalokem a větším množstvím vedlejších laloků. Konvolucí spektra okna se spektrem signálu se jediná spektrální čára ve spektru signálu rozšíří (rozmaže) na tvar hlavního laloku. Šířka hlavního laloku tak určuje kmitočtové rozlišení DFT a pro délku okna NT_{vz} je dána vztahem $2/NT_{vz}$, kde T_{vz} je vzorkovací perioda. Z toho plyne, že chceme-li dosáhnout velkého spektrálního rozlišení (při stejném vzorkování), musíme volit N co největší. To může mít za následek špatnou detekci, jelikož při dlouhém analyzovaném úseku budou rychlé spektrální změny průměrovány a nemohou být detekovány [5].

Druhý nežádoucí efekt (rozptyl spektra) je způsoben vedlejšími laloky ve spektru okna a projevuje se tím,že ve spektru navzorkovaného řečového signálu se objeví nové spektrální čáry vně hlavního laloku. Tento efekt nelze potlačit změnou délky okna, můžeme ho ovlivnit pouze tvarem okna. U pravoúhlého okna je výška prvního vedlejšího laloku 13 dB pod maximem hlavního laloku [5].

U řečového signálu (zejména v jeho znělých úsecích) se vyskytují rozdíly mezi nejsilnějšími a nejslabšími kmitočtovými komponenty více než 40 dB. Použijeme-li typ pravoúhlého okna, nemůžeme slabé komponenty ve spektru vůbec postihnout. Tento problém vyřešíme použitím vhodnějšího typu okna, obvykle Hammingova. Toto okna má ve spektru sice dvojnásobný hlavní lalok, ovšem útlum vedlejších laloků 43 dB je podstatně lepší (obr. 11.2 b).



Obr. 2.4: Modulová charakteristika pravoúhlého okna (a) a Hammingova okna (b). Obrázek byl převzat z [5].

3 DETEKTORY ŘEČOVÉ AKTIVITY – STANDARDNÍ METODY DETEKCE ŘEČ/PAUZA

Detektor řečové aktivity je systém, do kterého vstupuje čistý řečový signál, nebo směs řeči a hluku. Výstupem je pak „1“, je-li v i-tém segmentu přítomna řeč a „0“ pro segment bez řečové aktivity.

3.1 Požadavky na detektory

Detektory řečové aktivity obvykle mají splňovat následující požadavky [3]:

- detektor by měl být zrealizovaný ve frekvenční oblasti
- informaci o přítomnosti řeči by měl získávat pouze v krátkých aktuálních signálových segmentech a pracovat v reálném čase
- detektor by měl dosahovat velké pravděpodobnosti správné detekce řeči, pauzy i při relativně malém poměru signál/šum

3.2 Princip stanovení řečové aktivity

Většina detektorů využívá obdobný princip stanovení řečové aktivity. Ten lze shrnout do několika kroků:

1. Vstupní signál se rozdělí na časové segmenty (rámce).
2. Stanovení potřebné charakteristiky signálu (dle detektoru, např. energie, kepstrum ...).
3. Vypočtená charakteristika se v každém rámci porovná s prahovou hodnotou. Prahová hodnota může být v průběhu detekce přepočítávána a aktualizována (detektory s adaptivně nastavovaným prahem), nebo též stanovena pevně.
4. Je-li daná charakteristika signálu v rámci větší než prahová, je detekována řeč, jinak pauza.

3.2.1 Činnost detektorů s adaptivně nastavovaným prahem

1. Inicializační fáze: v signálu nesmí být přítomna řeč. Nastaví se počáteční hodnota prahu. Z několika počátečních rámců se vypočte střední hodnota *E(X)*

a rozptyl *D(X)* (obecné charakteristiky *X*). Prahová hodnota t se nastaví podle následujícího vztahu:

$$t = E(X) + \alpha\sqrt{(D(X))}, \tag{3.1}$$

kde α je konstanta, která se pohybuje v rozmezí 1 až 3, závisí na typu použitého detektoru a odstupu signál/šum. Definice:

$$E(X) = \sum_{x_i=Q} (x_i)p(x_i) \tag{3.2}$$

$$D(X) = \sum_{x_i=Q} [x_i - E(X)]^2 p(x_i), \tag{3.3}$$

přičemž náhodná veličina X se nazývá diskrétní, když existuje posloupnost reálných čísel x_i a posloupnost nezáporných reálných čísel p_i taková, že platí:

$$p(X = x_i) = p_i \qquad \text{a} \qquad \textstyle\sum_i p(X = x_i) = \sum_i p_i = 1 \tag{3.4}$$

2. Fáze detekce: určí se hodnota vybrané charakteristiky signálu v aktuálním rámci *X*. Je-li větší než prahová hodnota t, je detekována řeč, jinak pauza. Zároveň se upraví střední hodnota *E(X)* a rozptyl *D(X)* vybrané charakteristiky a určí se nová prahová hodnota.

3.3 Typy detektorů

3.3.1 Ideální detektor

Ideální detekce se realizuje tak, že ručně označíme úseky signálu s řečovou aktivitou. Pro svou pracnost se však tato detekce nedá použít pro větší množství dat a taky jí nelze detekovat signál v reálném čase. V praxi se realizuje aplikací kepstrálního šumu na řečový signál bez šumového pozadí [4].

3.3.2 Energetický detektor

Tento detektor výpočtem zjišťuje energii signálu pro každý rámec [3]. Využívá se funkce krátkodobé energie, tu lze definovat vztahem:

$$E_n = \sum_{k=-\infty}^{\infty} [s(k)w(n-k)]^2, \tag{3.5}$$

kde $s(k)$ je vzorek signálu v čase k a $w(n)$ je příslušný typ váhovacího okénka (kapitola 11.1). Při měření krátkodobé energie je doporučená délka rámců 10 až

20 ms při frekvenci vzorkování 8 – 10kHz. Funkce je citlivá na velké změny úrovně signálu, proto se často využívá krátkodobá intenzita, která tento nedostatek nemá:

$$M_n = \sum_{k=-\infty}^{\infty} |s(k)|w(n-k). \tag{3.6}$$

Vypočtený segment energie se porovná s prahovou hodnotou energie E_p definovanou:

$$E_p = 1{,}5.E_d, \tag{3.7}$$

kde E_d je úroveň energie hluku pozadí aktualizována podle vztahu:

$$E_d = (1-p)E_d^o + pE, \tag{3.8}$$

kde E_d^o je poslední známá hodnota energie.

Aktualizace se provádí v řečových pauzách. E je energie signálu v segmentu a parametr p se volí např.: 0,5.

Je-li $E_n > E_p$ je detekována řeč, v opačném případě je energie šumu aktualizována (3.8).

3.3.3 Kepstrální integrální detektor

Vstupní signál značíme $x_l[n]$ (index „l“ znamená l-tý rámec vstupního signálu). Používá-li se při výpočtu reálná část, hovoří se o reálném kepstru:

$$c_l[k] = \mathrm{Re}\{IDFT\{\log|DFT\{x_l[n]\}|\}\}. \tag{3.9}$$

Vybraná charakteristika signálu $x_l[n]$ se u kepstrálního integrálního detektoru řeč/pauza stanovuje z kepstra signálu v l-tém rámci $c_l[k]$ a průměrného kepstra pozadí $\overline{c}_l[k]$. Hovoří se o tzv. vzdálenosti kepstra signálu v rámci [1], značíme Δc_l:

$$\Delta c_l = 4{,}3429\sqrt{(c_1[0]-\overline{c}_l[0])^2 + 2\sum_{k=1}^{p}(c_1[k]-\overline{c}_l[k])^2}. \tag{3.10}$$

Výpočet Δc_l:

Používá se reálné kepstrum (3.8) signálu $x_l[n]$ v aktuálním rámci l o délce N vzorků.

Detekce řečového úseku se provádí porovnáním kepstrální vzdálenosti Δc_l:

1 – pokud $\Delta c_l \geq t_l$
0 – pokud $\Delta c_l < t_l$

V řečové pauze se aktualizuje prahová hodnota t_l. Využívá se matematické statistiky:

$$t_l = E(\Delta c[n]) + \alpha\sqrt{D(\Delta c[n])}, \tag{3.11}$$

také dochází k aktualizaci průměrného kepstra pozadí $\overline{c}_{l+1}[k]$

$$\overline{c}_{l+1}[k] = (1-\beta)\overline{c}_l[k] + \beta c_l[k]. \tag{3.12}$$

Poznámky: α – je vhodné volit v intervalu 1,8 až 3,5, β – časová konstanta exponenciálního průměrování (nabývá hodnot 0 až 1).

Detektor pro správnou funkci potřebuje počáteční prodlevu (inicializační fáze detektoru (část 3.2.1)).

4 DETEKTORY ŘEČOVÉ AKTIVITY ZALOŽENÉ NA HMM

Tento typ detektoru řečové aktivity je založen na statistickém modelování – skrytých Markovových modelech. Skryté Markovovy modely (HMM) jsou statistické stavové automaty, které se dají dobře použít k rozpoznávání řeči, ale stejně tak i k detekci řeči. HMM modelují prvky řeči a na základě tohoto statistického modelování lze rozlišovat řeč od ticha (šumu pozadí). Detektory založené na HMM umožňují detekci řeči i na dynamicky se měnícím šumovém pozadí a dokážou modelovat časový vývoj šumového pozadí, čímž můžeme zvětšit citlivost detektoru na některé specifické situace, např. rozjezd auta [6]. Každou situaci můžeme modelovat jiným typem modelu. Pomocí jednostavových modelů modelujeme stacionární pozadí, pomocí vícestavových potom nestacionární procesy.

Podle [6] tyto detektory dosahují lepších výsledků, než tradiční detektory, a to hlavně u silně zarušených signálů.

Tato kapitola je v práci uvedena pro doplnění možností detekce řečové aktivity. Detektor tohoto typu není v praktické části implementován, protože téma HMM je komplikované a je mimo rozsah této práce.

4.1 Detektory založené na prahování HMM pravděpodobností

Detekce probíhá na základě modelování částí řeči pomocí jednoho nebo i více HMM. Jeden model většinou odpovídá úsekům řeči a druhý úsekům šumového pozadí. Bude použit pojem dlouhodobý a krátkodobý segment. Z krátkodobého segmentu řeči je vypočítán jeden vektor parametrů. Pojmem dlouhodobý segment se myslí posloupnost vektorů parametrů. V dlouhodobých segmentech stejné nebo rostoucí délky se provádí výpočet pravděpodobnosti, zda HMM generuje příslušný úsek šumu nebo řeči. Výstupem jsou vektory pravděpodobností příslušné ke každému řečovému rámci. Je-li použit pouze jeden model – model šumu, může být prahována přímo tato pravděpodobnost. Jsou-li použity dva modely – model šumu i model řeči, jsou pravděpodobnosti nejprve vzájemně odečteny a výsledný rozdíl je prahován. Tyto vektory, které jsou prahovány, jsou označeny jako kriteriální funkce [6].

4.1.1 Kriteriální funkce

Je-li použit jeden model šumu, který se označí λ_N, potom je kriteriální funkce $f[n]$ tvaru:

$$f[n] = \log\left(P(\mathrm{O}|\lambda_{\mathrm{N}})\right), \tag{4.1}$$

kde O značí dlouhodobý segment řeči. Jsou-li použity dva modely: model šumu λ_N a model řeči λ_S, potom má kriteriální funkce $f[n]$ tvar:

$$f[n] = \log\left(P(\mathrm{O}|\lambda_{\mathrm{S}})\right) - \log\left(P(\mathrm{O}|\lambda_{\mathrm{N}})\right). \tag{4.2}$$

4.1.2 Algoritmus detektoru využívající segmenty stejné délky

U tohoto typu detektoru dochází k analýze HMM pravděpodobnosti (získané z jednoho nebo dvou modelů) z posunovaných dlouhodobých segmentů stejné délky. Algoritmus detekce má 4 fáze:

1. Inicializační fáze: Zde se nastavují parametry modelů řeči λ_S a šumu λ_N.

2. Trénování: Modely se trénují pomocí Baum-Welchova algoritmu [6].

3. Kriteriální funkce: V této fázi se vypočte kriteriální funkce, podle vztahů 4.1 a 4.2.

4. Fáze klasifikace: V poslední fázi dochází k prahování, detekci a vyhlazení mediánovým filtrem [6].

5 DETEKTOR ITU-T G.729

Detektor pracuje s délkou rámce 240 vzorků. Detekce se provádí každých 10 ms. Detekce řečové aktivity, detektorem definovaným v ITU-T G.729 je založená na extrakci čtyř parametrů [9]:

- úzkopásmová energie
- širokopásmové energie
- středního počtu průchodů signálu nulou
- kmitočtem spektrálních párů – LSF (Line Spectral Frequency)

Během řečové pauzy je z těchto parametrů získáván dlouhodobý průměr, který se mění v závislosti na šumovém pozadí. V každém rámci se počítají rozdílové parametry, jako rozdíl všech parametrů a jejich dlouhodobého průměru. Na základě množiny rozdílových parametrů a rozhodovacích podmínek dojde k označení rámce jako „řečový“ nebo „šumový“.

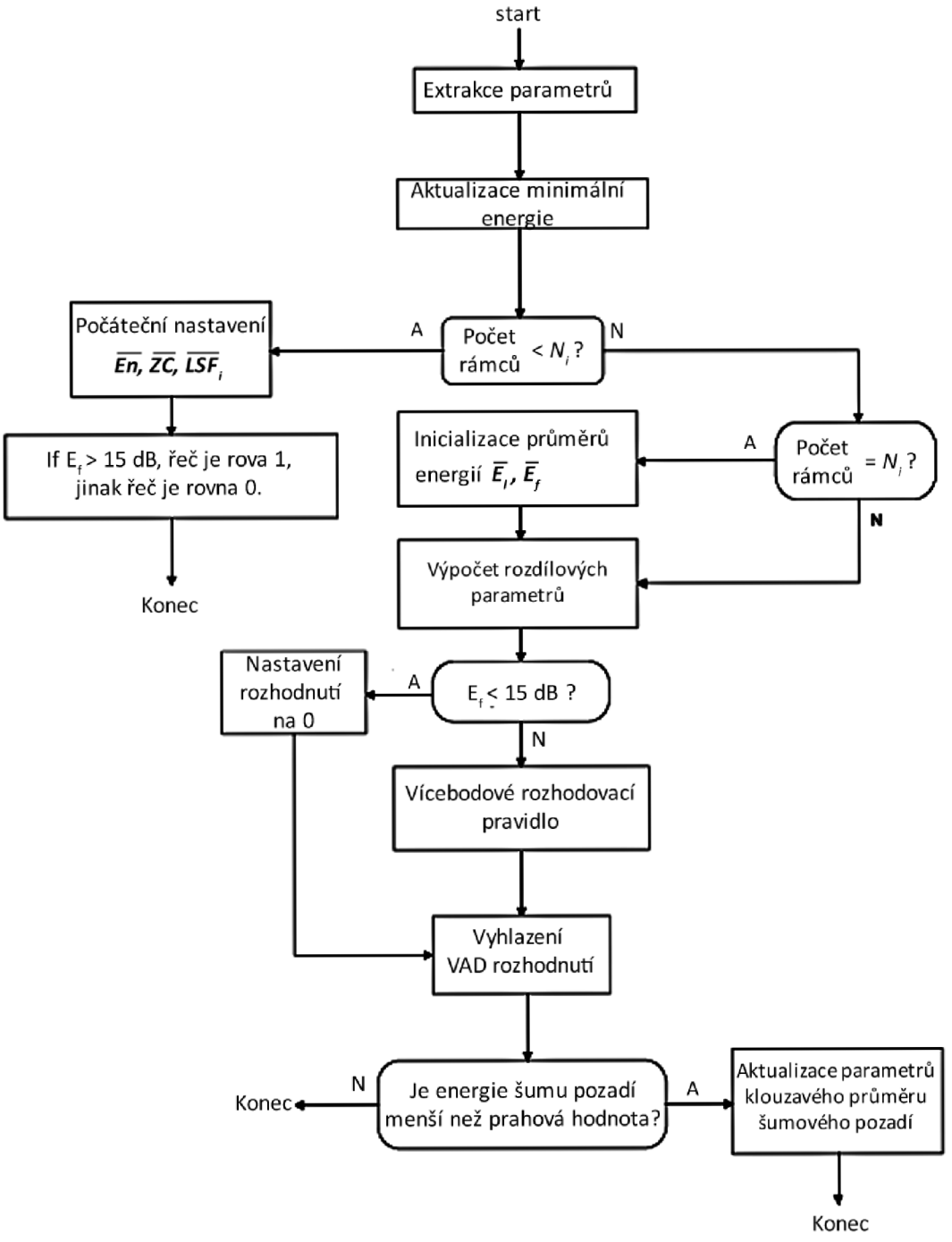
5.1 Popis VAD algoritmu

Prvním krokem algoritmu je předzpracování vstupního řečového signálu (část 5.2).

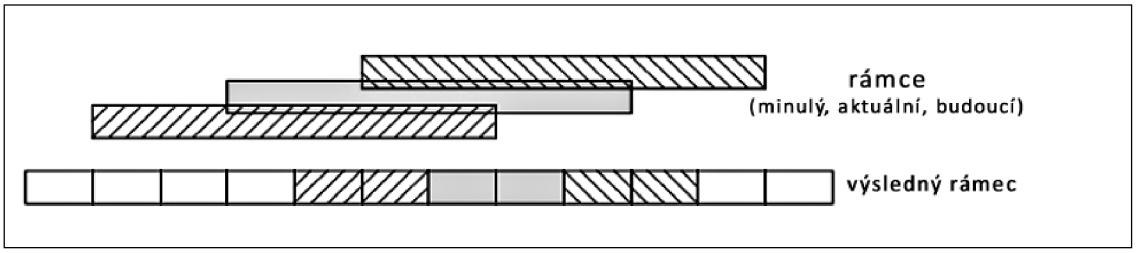
Na obr. 5.1 je znázorněný vývojový diagram VAD. Algoritmus VAD se provádí pro každý řečový rámec. Řečové rámce jsou od počátku každého hovoru/nahrávky nepřetržitě získávány pomocí posouvající se okénkové funkce (5.2). V první řadě jsou ze vstupního signálu extrahovány čtyři parametricky, což jsou širokopásmová a úzkopásmová rámcová energie, LSF a počet průchodů nulou v rámci.

Jestliže je číslo rámce menší než N_i, vypočtou se počáteční hodnoty dlouhodobého průměru parametrů a pokud je energie rámce po LPC analýze větší než 15 dB, bude rámec označený jako řečový – 1 (viz. 5.5). Jinak je rozhodnutí o řečové aktivitě rovno 0. Jestliže je číslo rámce rovno N_i, inicializují se počáteční průměry charakteristických energií (E_f, E_l) v závislosti na šumovém pozadí.

V další části se vypočítají rozdílové parametry. Množina rozdílových parametrů se počítá jako rozdíl mezi aktuálními parametry rámce a klouzavého průměru charakteristik šumového pozadí. V následující části se provede počáteční rozhodnutí o řečové aktivitě, používá se vícebodové rozhodovací pravidlo [9]. V posledním kroku se zvažuje, jestli je energie šumu pozadí rámce menší než hraniční hodnota, když ano, dochází k aktualizaci parametrů klouzavého průměru šumového pozadí, jestliže ne, je algoritmus ukončen.



Obr. 5.1: Vývojový diagram algoritmu G.729. Obrázek byl převzat z [9].



Obr. 5.2: Ukázka skladby řečového rámce. Obrázek byl převzat z [9].

5.2 Předzpracování vstupního řečového signálu

Vstupní signál je filtrován horní propustí s mezním kmitočtem $f_m = 140$ Hz. Účelem filtrace je odstranit nežádoucí složky, které se vyskytují na nízkých kmitočtech. K filtrování je použit filtr s následující přenosovou funkcí:

$$H_{hl}(z) = \frac{0.4663718 - 0.92724705z^{-1} + 0.46363718z^{-2}}{1 - 1.9059465z^{-1} + 0.9114024z^{-1}} \tag{5.1}$$

Takto připravený signál bude dále označován jako $s[n]$, [9].

5.3 Extrakce parametrů

Signál $s[n]$ je dále váhován nesymetrickým oknem. Okénková funkce se skládá ze dvou částí: první část tvoří polovina Hammingova okna, část druhá je čtvrtina periody kosinové funkce. Funkce je dána následovně:

$$w_{lp}(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{399}\right) & n = 0, ..., 199 \\ \cos\left(\frac{2\pi(n-200)}{159}\right) & n = 200, ..., 239 \end{cases} \tag{5.2}$$

Okno lineární predikční analýzy je složeno ze 120 vzorků minulého řečového rámce, 80 vzorků aktuálního rámce a 40 vzorků budoucího rámce. Jelikož pracujeme se 40 vzorky budoucího rámce, je třeba vnést do algoritmu 40 vzorků zpoždění, což odpovídá času 5 ms.

Pro signál vážený oknem platí:

$$s'(n) = w_{lp}(n)s(n) \qquad n = 0, ..., 239 \tag{5.3}$$

Takto upravený signál se použije k výpočtu autkokorelačních koeficientů:

$$r(k) = \sum_{n=k}^{239} s'(n)s'(n-k) \qquad k \ = 0, ..., 10 \tag{5.4}$$

5.3.1 Kmitočet spektrálních párů (LSF – Line Spectral Frequencies)

Množina lineárních predikčních koeficientů je odvozen z autokorelace a množina $\{LSF\}_{i=1}^{p}$, kde p=10, je získána z LP koeficinentů, jak je popsáno v části 3.2.3/G.729 [9].

5.3.2 Širokopásmová energie

Širokopásmová energie odpovídá logaritmu normalizované hodnoty prvního autokorelačního koeficientu [9].

$$E_f = 10 \log_{10} \left[\frac{1}{N} R(0) \right], \tag{5.5}$$

kde $N = 240$ je velikost řečového rámce, získána vážením okénkovou funkcí.

5.3.3 Úzkopásmová energie

Úzkopásmová energie E_l naměřená v pásmu od 0 do F_l Hz, se vypočítá následovně:

$$E_l = 10 \log_{10} \left[\frac{1}{N} h^T \mathbf{R} \mathbf{h} \right], \tag{5.6}$$

kde $\mathbf{h}$ je impulsní odezva FIR filtru s mezním kmitočtem F_l a $\mathbf{R}$ je Toeplitzova matice s autokorelačními koeficienty na každé diagonále [9].

5.3.4 Střední počet průchodů signálu nulou

Následující vztah definuje hodnotu středního počtu průchodu signálu nulou [9]:

$$ZC = \frac{1}{2M} \sum_{i=0}^{M-1} \left[|\mathrm{sgn}[s(i)] - \mathrm{sgn}[s(i-1)]| \right], \tag{5.7}$$

kde $M = 80$.

5.4 Inicializace klouzavého průměru parametrů z charakteristik šumového pozadí

Pro prvních N_i rámců označíme klouzavý průměr spektrálních parametrů jako $\{\overline{LSF}\}_{i=1}^{p}$, který získáme jako průměr $\{LSF\}_{i=1}^{p}$ v p rámcích. Klouzavý průměr středního počtu průchodů nulou označíme jako $\overline{ZC}$, který získáme jako průměr středního počtu průchodů nulou ZC v rámcích [9].

Klouzavý průměr šumového pozadí širokopásmové energie označíme $\overline{E}_f$ a úzkopásmové energie $\overline{E}_l$. Jejich inicializace je následující: V prvním kroku inicializační procedura použije průměrnou energii $\overline{E}_n$, která je vypočtena jako průměr širokopásmové energie E_f v prvních N_i rámcích. Tyto tři průměry ($\overline{E}_n$, $\overline{ZC}$ a $\{\overline{LSF}\}_{i=1}^{p}$) zahrnují jenom rámce, jejichž energie E_f je větší než 15 dB. V druhém kroku je průběh inicializační procedoru následující:

Když $\overline{E}_n \leq T_1$, pak
$\overline{E}_f = \overline{E}_n + K_0$
$\overline{E}_l = \overline{E}_n + K_1$.

Jinak, když $T_1 \leq \overline{E}_n \leq T_2$, pak
$\overline{E}_f = \overline{E}_n + K_2$
$\overline{E}_l = \overline{E}_n + K_3$.

V ostatních případech:
$\overline{E}_f = \overline{E}_n + K_4$
$\overline{E}_l = \overline{E}_n + K_5$.

Konstanty jsou uvedy v Tabulce konstant, viz. příloha A.1.

5.5 Výpočet dlouhodobé minimální energie

Parametr dlouhodobá minimální energie E_{min} se vypočítá jako minimální hodnota širokopásmové energie E_f z N_0 předchozích rámců. Jakmile je N_0 poměrně rozsáhlé, použije se jako E_{min} minimální hodnota E_f získaná v přechozích výpočtech.

5.6 Výpočet rozdílových parametrů

Čtyři rozdílové parametry jsou vypočteny z hodnot parametrů a klouzavého průměru šumového pozadí daného parametru [9].

Spektrální rozložení ΔS

Hodnota změny spektrálního rozložení je vypočtena jako suma kvadrátů rozdílů mezi vektorem $\{LSF\}_{i=1}^{p}$ aktuálního rámce a vektorem klouzavého průměru $\{\overline{LSF}\}_{i=1}^{p}$.

$$\Delta S \ = \sum_{i=1}^{p} \left(LSF_i - \overline{LSF_i} \right)^2. \tag{5.8}$$

Rozdíl širokopásmové energie ΔE_f

Rozdíl širokopásmové energie je vypočten jako rozdíl mezi energií E_f aktuálního rámce a klouzavým průměrem širokopásmové energie $\overline{E}_f$.

$$\Delta E_f = \overline{E}_f - E_f \tag{5.9}$$

Rozdíl úzkopásmové energie ΔE_l

Rozdíl úzkopásmové energie je vypočten jako rozdíl mezi energií E_l aktuálního rámce a klouzavým průměrem úzkopásmové energie $\overline{E}_l$.

$$\Delta E_l = \overline{E}_l - E_l \tag{5.10}$$

Rozdíl středního počtu průchodů nulou ΔZC

Rozdíl středního počtu průchodů signálu nulou je vypočten jako rozdíl mezi středním počtem průchodů signálu nulou ZC aktuálního rámce a klouzavým průměrem středního počtu průchodů signálu nulou $\overline{ZC}$.

$$\Delta ZC = \overline{ZC} - ZC \tag{5.11}$$

5.7 Vícebodové rozhodovací pravidlo

Rozhodnutí o přítomnosti řečové aktivity je označeno jako I_{VD}. Při inicializaci a pokud vektor rozdílových parametrů leží uvnitř oblasti bez řečové aktivity je nastaveno na 0 („FALSE“). V opačném případě je rozhodnutí nastaveno na 1 („TRUE“). Čtrnáctibodové rozhodovací pravidlo ve čtyřrozměrném prostoru je definováno následovně [9]:

1. pokud $\Delta S > a_1 \cdot \Delta ZC + b_1$, potom $I_{VD} = 1$
2. pokud $\Delta S > a_2 \cdot \Delta ZC + b_2$, potom $I_{VD} = 1$
3. pokud $\Delta E_f < a_3 \cdot \Delta ZC + b_3$, potom $I_{VD} = 1$
4. pokud $\Delta E_f < a_4 \cdot \Delta ZC + b_4$, potom $I_{VD} = 1$
5. pokud $\Delta E_f < b_5$, potom $I_{VD} = 1$
6. pokud $\Delta E_f < a_6 \cdot \Delta S + b_6$, potom $I_{VD} = 1$
7. pokud $\Delta S > b_7$, potom $I_{VD} = 1$
8. pokud $\Delta E_l < a_8 \cdot \Delta ZC + b_8$, potom $I_{VD} = 1$
9. pokud $\Delta E_l < a_9 \cdot \Delta ZC + b_9$, potom $I_{VD} = 1$

10. pokud $\Delta E_l < b_{10}$, potom $I_{VD} = 1$

11. pokud $\Delta E_l < a_{11} \cdot \Delta S + b_{11}$, potom $I_{VD} = 1$

12. pokud $\Delta E_l > a_{12} \cdot \Delta E_f + b_{12}$, potom $I_{VD} = 1$

13. pokud $\Delta E_l < a_{13} \cdot \Delta E_f + b_{13}$, potom $I_{VD} = 1$

14. pokud $\Delta E_l < a_{14} \cdot \Delta E_f + b_{14}$, potom $I_{VD} = 1$

Pokud nevyhovuje žádná podmínka, potom $I_{VD} = 0$. Konstanty jsou uvedy v Tabulce konstant, viz. příloha A.1.

5.8 Vyhlazení rozhodnutí VAD

Funkce VAD rozhodnutí se vyhlazuje s ohledem na dlouhodobou stacionární povahu řečového signálu. Vyhlazení se provádí ve čtyřech krocích.

Aktivita vyhlazovacího procesu je indikována návěštím, které je označeno jako v_flag [7]. Vždy před vykonáním vyhlazovacího procesu je návěští nastaveno na nulu. Vyhlazené rozhodnutí o řečové aktivitě je označeno jako S_{VD}^{0} v aktuálním rámci, S_{VD}^{-1} v předchozím rámci a S_{VD}^{-2} v předminulém rámci. S_{VD}^{-1} a S_{VD}^{-2} je na počátku nastaveno na 1, $S_{VD}^{0} = I_{VD}$. První krok vyhlazovacího procesu:

$$\text{if } \Big[(I_{VD} = 0) \text{ and } (S_{VD}^{-1} = 1) \text{ and } (E > \overline{E_f} + T_3)\Big] \text{ then } \Big[S_{VD}^{0} = 1 \text{ and } v\_flag = 1\Big].$$

V druhém kroku vyhlazování se definuje logický parametr F_{VD}^{-1} a počítadlo vyhlazovacího procesu C_e. Parametr F_{VD}^{-1} je při inicializaci nastaven na 1 a C_e na 0. Širokopásmová energie předchozího rámce je označena jako E_{-1}. Druhý krok potom vypadá následovně:

if $\Big[(F_{VD}^{-1} = 1)$ and $(I_{VD} = 0)$ and $(S_{VD}^{-1} = 1)$ and $(S_{VD}^{-2} = 1)$ and $(|E_f - E_{-1}| \leq T_4)\Big]$ then

{

$S_{VD}^{0} = 1$

$v\_flag = 1$

$C_e = C_e + 1$

if$(C_e \leq N_1)$\{

$F_{VD}^{-1} = 1$

}

else {

$F_{VD}^{-1} = 0$
$C_e = 0$
}
}
else
$F_{VD}^{-1} = 1$

Ve třetím kroku vyhlazování se definuje počítadlo spojitosti šumu C_S, které se při inicializaci nastaví na nulu. Pokud bude $S_{VD}^0 = 0$, C_S se inkrementuje. Třetí krok vypadá následovně:

if $\Big[(S_{VD}^0 = 1)$ and $(C_S > N_2)$ and $(E_f - E_{-1}| \leq T_5)\Big]$ then
{
$S_{VD}^0 = 0$
$C_S = 0$
}
if $(S_{VD}^0 = 1)$ then $C_S = 0$

Ve čtvrtém kroku vyhlazení se kontroluje splnění následující podmínky:

if $\Big[(E_f < \overline{E}_f + T_6)$ and $(frm\_count > N_0)$ and $(v\_flag = 0)\Big]$, potom $S_{VD}^0 = 0$, kde frm_count je číslo aktuálního rámce.

5.9 Aktualizace klouzavého průměru charakteristik šumového pozadí

Klouzavý průměr charakteristik šumového pozadí je aktualizován v poslení části VAD algoritmu. Je testována následující podmínka: $(E_f < \overline{E}_f + T_6)$, při jejím splnění dojde k aktualizaci průměru [9].

Klouzavý průměr charakteristik šumového pozadí je aktualizován užitím autoregresivního (AR) schématu prvního řádu. Rozdílové AR koeficienty jsou využity pro různé parametry a množina koeficientů se používá na začátku promluvy nebo pokud je detekována velká změna šumových charakteristik [7].

Nechť β_{E_f} je AR koeficient pro aktualizaci $\overline{E}_f$, β_{E_l} je AR koeficient pro aktualizaci $\overline{E}_l$, β_{ZC} je AR koeficient pro aktualizaci $\overline{ZC}$ a β_{LSF} je AR koeficient pro aktualizaci $\{\overline{LSF}\}_{i=1}^{p}$. Součet všech rámců, v nichž se provádí aktualizace je označen jako C_n.

AR aktualizace se provádí následovně:

$$
\begin{aligned}
\overline{E}_f &= \beta_{E_f} \cdot \overline{E}_f + \left(1 - \beta_{E_f}\right) \cdot E_f \\
\overline{E}_l &= \beta_{E_l} \cdot \overline{E}_l + \left(1 - \beta_{E_l}\right) \cdot E_l \\
\overline{ZC} &= \beta_{ZC} \cdot \overline{ZC} + \left(1 - \beta_{ZC}\right) \cdot ZC \\
\overline{LSF}_i &= \beta_{LSF} \cdot \overline{LSF}_i + \left(1 - \beta_{LSF}\right) \cdot LSF_i \quad i = 1, \ldots, p
\end{aligned} \tag{5.12}
$$

Poté budou aktualizovány E_f a C_n:

if $\left[(frame\_count > N_0) \text{ and } (\overline{E}_f < E_{min})\right]$ then
{
$\overline{E}_f = E_{min}$
$C_n = 0$
}

6 KOMBINOVANÝ DETEKTOR

Detektor byl vytvořen v rámci této diplomové práce. Je zkonstruován na základě kombinace výstupu detekce tří detektorů řečové aktivity (energetický detektor, detektor intenzity a kepstrální detektor), které byly popány výše.

Každý z těchto tří detektorů má nějakou slabinu a naopak nějakou přednost, v které lépe vyniká při detekci řečové aktivity. Motivací pro tvorbu kombinovaného detektoru je tedy nalézt přednosti zmiňovaných detektorů a tím zvýšit účinnost detekce.

Zda bude rozpoznávání takto vytvořeným detektorem účinnější, než rozpoznávání samotnými detektory se prokáže až při testování a vyhodnocení pomocí ROC křivek. Musíme však brát na zřetel také fakt, že oproti výpočetní náročnosti jednotlivých detektorů, výpočetní náročnost kombinovaného detektoru značně vzroste.

6.1 Popis algoritmu kombinovaného detektoru

V prvním kroku algoritmu se načte vstupní řečový signál. Dále se provede rozhodnutí o řečové aktivitě zmiňovanými třemi pomocnými detektory. Následuje podmínkové rozhodovací pravidlo, na jehož základě se rozhoduje o řečové aktivitě. Posledním krokem je vyhlazení rozhodnutí detektoru.

6.1.1 Podmínkové rozhodovací pravidlo

Rozhodnutí o přítomnosti řečové aktivity se provádí pro každý rámec a je označeno jako *vad*. Na počátku je nastaven ukazatel *flag*, který značí, zda již byla v daném rámci provedena detekce (*flag = 1*), nebo nebyla (*flag = 0*).

V první podmínce se klade důraz na výstup kepstrálního detektoru. Konkrétně se zjišťuje, jestli u tohoto detektoru nastává sestupná hrana detekce, tj. označení předchozích dvou rámců jako řečové a následujícího jako neřečový. Pokud je podmínka splněna, nastane sestupná hrana v daném úseku i u kombinovaného detektoru a ukazatel *flag = 1*.

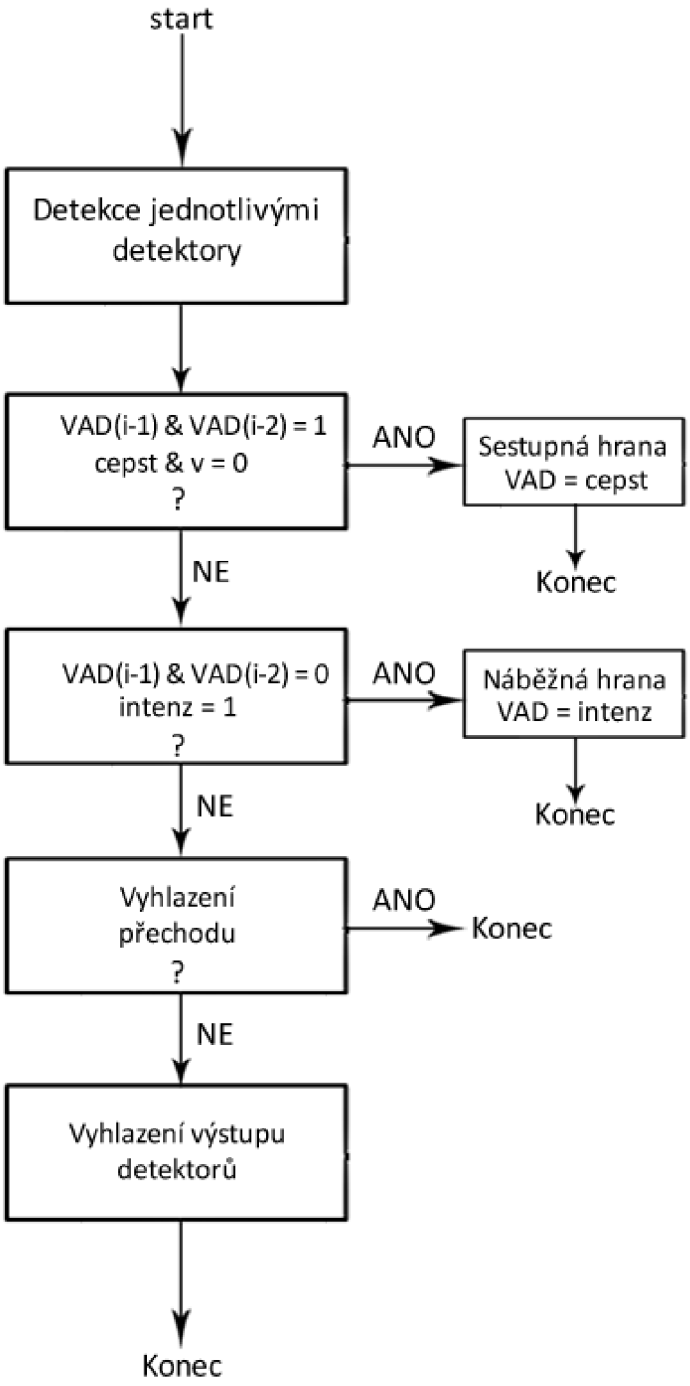
Na počátku každé z následujících podmínek se testuje, zda ještě nebyla provedena detekce, tzn. je-li *flag = 0*. Pokud *flag = 1*, podmínky se neprovádí a přeskakují se.

Druhá podmínka analogicky zjišťuje náběžnou hranu, přičemž klade důraz na detektor intenzity. Jestliže u tohoto detektoru nastává náběžná hrana detekce, tj. označení předchozích dvou rámců jako neřečové a následujícího jako řečový, pak nastane náběžná hrana i u kombinovaného detektoru a ukazatel *flag = 1*.

Třetí podmínka zajišťuje, aby po přechodu ze stavu 0 do stavu 1, byl následující rámec opět označen jako řečový (stav 1).

Čtvrtá podmínka zajišťuje, aby po přechodu ze stavu 1 do stavu 0, byl následující rámec opět označen jako neřečový (stav 0).

Poslední pátá podmínka ošetřuje stav, kdy nebyla splněna žádná z předchozích podmínek. Rozhodnutí o řečové aktivitě se provede na základě vyhodnocení výstupů všech tří detektorů (převládá-li stav 1, je rámec označen za řečový, převládá-li stav 0, je rámec označen za neřečový). Tato podmínka by se dala stejně dobře realizovat pomocí mediánového filtru, který by byl aplikován na výstupy všech tří pomocných detektorů v daném rámci.

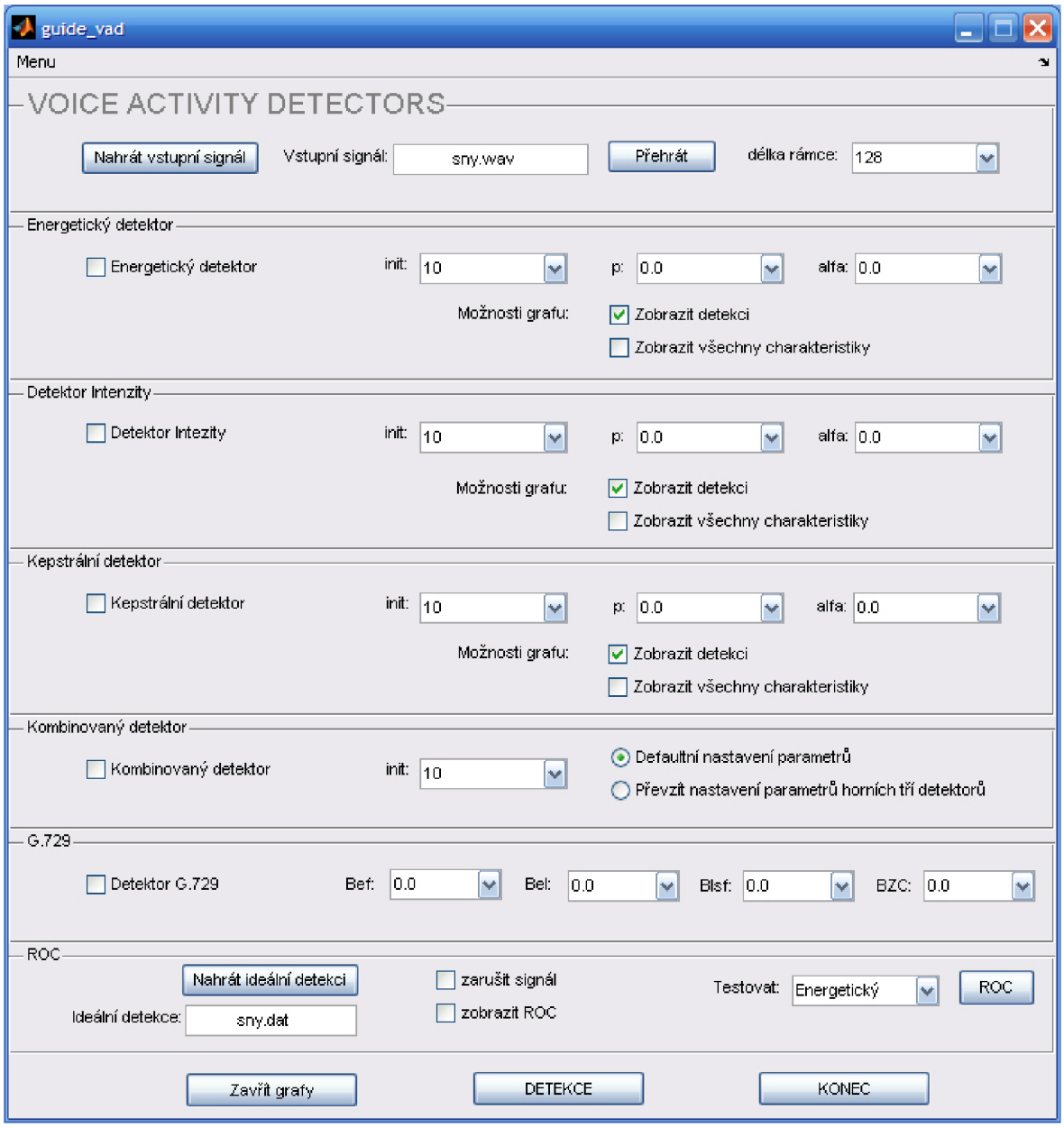


Obr. 6.1: Vývojový diagram kombinovaného detektoru.

7 IMPLEMENTACE V MATLABU

7.1 Uživatelské rozhraní pro testování metod

Na obr. 7.1 je zobrazeno uživatelské rozhraní pro Matlab, které slouží k testování a srovnávání realizovaných metod detekce řečové aktivity.



Obr. 7.1: Uživatelské rozhraní pro testování metod.

7.1.1 Popis vstupních volitelných parametrů

- **Vstupní signál** – zvukový soubor ve formátu *.wav, který chceme analyzovat. Nahrávka se vybírá tlačítkem „Nahrát vstupní signál“.
- **Délka rámce** – počet vzorků vstupního signálu, který se použije pro tvorbu jednoho rámce (segmentu). Volitelné hodnoty jsou 128 nebo 256.
- **Init** – počet inicializačních rámcu potřebných k počátečnímu nastavení prahové hodnoty. Volitelné hodnoty jsou 10 až 30 rámců.
- **p** – časová konstanta exponenciálního průměrování. Volí se hodnoty 0 až 1.
- **alfa** – parametr, který se uplatňuje při stanovování počáteční prahové hodnoty a při obnově prahové hodnoty.
- **Bef, Bel, Blsf, BZC** – Bef je AR koeficient pro aktualizaci $\overline{E}_f$, Bel je AR koeficient pro aktualizaci $\overline{E}_l$, Blsf je AR koeficient pro aktualizaci $\{\overline{LSF}\}_{i=1}^{p}$ a BZC je AR koeficient pro aktualizaci $\overline{ZC}$.
- **Možnosti grafu**

  Při zaškrtnutí volby „Zobrazit detekci“ se zobrazí graf, který zobrazuje amplitudovou charakteristiku zvukového signálu s provedenou detekcí řečové aktivity. Při zaškrtnutí volby „Zobrazit všechny charakteristiky“ se zobrazí graf obsahující dva podgrafy. Horní podgraf zobrazuje časový průběh zvukového signálu s provedenou detekcí řečové aktivity. Dolní podgraf zobrazuje všechny průběhy, typické pro daný detektor řečové aktivity.
- **Výchozí nastavení parametrů**

  Volba pro kombinovaný detektor. Vybráním této možnosti při detekci kombinovaným detektorem bude vybráno výchozí nastavení vstupních parametrů pro energetický detektor, detektor intenzity a kepstrální detektor.
- **Převzít nastavení parametrů horních tří detektorů**

  Volba pro kombinovaný detektor. Vybráním této možnosti při detekci kombinovaným detektorem budou jako vstupní parametry vzaty hodnoty nastavené ve volbách „init“, „p“ a „alfa“ pro energetický detektor, detektor intenzity a kepstrální detektor.
- **dat** – textový soubor ve formátu *.dat
- **roc** – volí se logická hodnota 1 nebo 0. Při volbě 1 se vykreslí křivky ROC. Při volbě 0 se křivky nevykreslí.

- **zarušit signál** – volí se logická hodnota 1 nebo 0. Při volbě 1 proběhne celá funkce v šesti cyklech a v každém cyklu se přidává k řečové nahrávce bílý šum. Výsledné testovací signály mají odstup signálu od šumu v hodnotách SNR $= \{30, 20, 10, 5, 0, -5\}$ dB.

7.1.2 Funkce uživatelského rozhraní pro testování metod

Načtení testovaného signálu se provede tlačítkem „Nahrát vstupní signál“. Nahrávka musí být ve formátu *.wav. Po stisku tlačítka „Přehrát“ bude nahrávka přehrána. Dále se zvolí délka rámce. Zaškrtnutím příslušného políčka se volí, který detektor bude použit k detekci řečové aktivity v načtené nahrávce. Poté se zvolí vstupní parametry detektoru. Tlačítkem „DETEKCE“ se spouští proces detekce řečové aktivity vybranými detektory. Zobrazí se grafy, které lze zavřít tlačítkem „Zavřít grafy“. Uživatelské rozhraní se ukončuje tlačítkem „KONEC“.

7.1.3 Funkce uživatelského rozhraní pro vyhodnocování metod

Tlačítkem „Nahrát ideální detekci“ se provede načtení souboru ve formátu *.dat. Tento soubor obsahuje testovaný signál, opatřený značkami začátků a konců řečových pauz (tzv. labely), který slouží k porovnání správnosti detekce vstupního signálu. Dále se vybere detektor, jehož účinnost má být vyhodnocena. Vyhodnocení se spouští tlačítkem „ROC“. Zaškrtnutím volby „zobrazit ROC“ se zobrazí závislost HR0 na FAR0. Volba „zarušit signál“ umožňuje přidat do řečové nahrávky bílý šum a tím postupně měnit odstup signálu od šumu v hodnotách SNR $= \{30, 20, 10, 5, 0, -5\}$ dB.

7.2 Popis funkcí

7.2.1 Popis funkce `vadenerg.m`

Jedná se o funkci realizující detekci řečové aktivity na principu krátkodobé energie.

Funkce nejprve provede načtení signálu pomocí příkazu `wavread` . Dále následuje rozhodovací podmínka pro volbu vhodné okénkové funkce. Standardně je nastaveno *Hammingovo okénko*, druhou možností je *pravoúhlé okénko*.

Další částí skriptu je inicializace. Zde se z počátečních rámců zjistí prahová hodnota energie. V poslední fázi se stanovuje řeč nebo pauza pomocí vhodné detekční podmínky. V rámci každého cyklu detekce probíhá obnova prahové hodnoty. Výsledek detekce je vyjádřen graficky.

Ke spuštění funkce se používá následujícího zápisu:
`vadenerg(vst_sig,delka_ramce,init,p,alfa)`
Např.: `vadenerg('Psenice',256,30,0,0.2)`

7.2.2 Popis funkce `vadcepstralni.m`

Tato funkce má podobnou strukturu jako energetický detektor. Pracuje však na odlišném principu. V inicializační fázi se stanovuje průměrné reálné kepstrum pozadí. Výpočtem vzdálenosti kepstra pozadí a reálných kepster v inicializačních rámcích se stanoví prahová hodnota detekce. Následuje výpočet kepstrální vzdálenosti reálného kepstra od kepstra pozadí i v ostatních rámcích. Posledním krokem je samotná detekce řečové aktivity porovnáváním kepstrální vzdálenosti reálného kepstra a prahové hodnoty. V rámci každého cyklu detekce probíhá obnova prahové hodnoty.

Ke spuštění funkce se používá následujícího zápisu:
`vadcepstralni(vst_sig,delka_ramce,init,p,alfa)`
Např.: `vadcepstralni('Psenice',256,30,1,0.5)`

7.2.3 Popis funkce `vadintenz.m`

Tento detektor je založený na funkci krátkodobé intenzity. Jedná se o obdobu energetického detektoru. Tato funkce však není, na rozdíl od energetického detektoru, citlivá na velké změny úrovně signálu.

Ke spuštění funkce se používá následujícího zápisu:
`vadinenz(vst_sig,delka_ramce,init,p,alfa)`
Např.: `vadintenz('Psenice',256,30,0,0.1)`

7.2.4 Popis funkce `g729.m`

Jedná se o funkci, která realizuje detektor řečové aktivity definován v ITU-T G.729. Funkce je založena na extrakci čtyř parametrů (úzkopásmová energie, širokopásmová energie, střední počet průchodů nulou a kmitočet spektrálních párů).

Algoritmus detekce řečové aktivity pro detektor g729.m je detailně popsán v kapitole 5.1.

Ke spuštění funkce se používá následujícího zápisu:
`g729(vst_sig,Bef,Bel,BZC,Blsf)`
Např.: `g729('franek',0.9,0.9,0.9,0.9)`

7.2.5 Popis funkce `detektor.m`

Funkce detektor.m realizuje tzv. „kombinovaný detektor“. Tento detektor byl vytvořen na základě výstupu detekce tří detektorů řečové aktivity (energetický detektor, detektor kntenzity a kepstrální detektor), které byly popány výše.

Funkci lze spouštět s defaultním nastavením vstupních parametrů pro pomocné detektory, nebo lze tyto parametry nastavit pro každý detektor zvlášť. O režimu ve kterém se detekce spustí rozhoduje vstupní proměnná „rezim“. Pokud „rezim = 0“, funkce se spustí s defaultně nastavenými parametry, pokud „rezim = 1“, pak se všechny vstupní parametry musí při spuštění zadat.

Ke spuštění funkce se používá následujícího zápisu:
`detektor(vst_signal,delka_ramce,rezim,init)`
Např.: `detektor('franek',256,0,10)`
nebo
`detektor(vst_signal,delka_ramce,rezim,init,pc,alfac,pi,alfai,pe,alfae)`
Např.: `detektor('franek',256,1,10,0.9,0.5,0,0.1,0.1,0.2)`

7.2.6 Popis funkce `idealni.m`

Funkce idealni.m má za úkol načíst soubor typu *.dat, který obsahuje labely testované zvukové nahrávky. Návratovou hodnotou funkce je vektor ideální detekce testovaného signálu. Délka vektoru je sjednocena s délkou výstupního vektoru testovaného detektoru. Vstupní parametry funkce:

- **dat** – textový soubor ve formátu *.dat
- **lv** – délka výstupního vektoru testovaného detektoru

Ke spuštění funkce se používá následujícího zápisu:
`vektor = idealni(dat,lv)`
Např.: `vektor = idealni('franek.dat',144)`

7.2.7 Popis funkce `vyhodnotROC.m`

Funkce je určená k vyhodnocení ROC křivek testovaného detektoru řečové aktivity. Křivky jsou tvořeny na základě srovnání vektoru ideální (ruční) detekce a výstupního vektoru testovaného detektoru. Funkce také umožňuje přidat do řečové nahrávky bílý šum a tím postupně měnit odstup signálu od šumu v hodnotách SNR $= \{30, 20, 10, 5, 0, -5\}$ dB. Pro získání vektoru ideální detekce, je uvnitř funkce volána funkce idealni.m a pro získání výstupního vektoru testovaného detektoru je volána funkce příslušného detektoru.

Ke spuštění funkce se používá následujícího zápisu:
`[HR1,HR0]=vyhodnotROC(vst_sig,roc,zasum)`
Např.: `[HR1,HR0]=vyhodnotROC('franek',1,1)`

Všechny výše popsané funkce jsou na CD přiloženém k diplomové práci.

7.3 Charakteristika testovacích signálů

Pro testování implementovaných detektorů bylo použito šest nahrávek s různým charakterem rušení a odstupem signálu od šumu, které byly namluveny třemi různými mluvčími. Nahrávky jsou ve formátu *.wav, f_{vz} = 8 kHz. Jde o následující nahrávky:

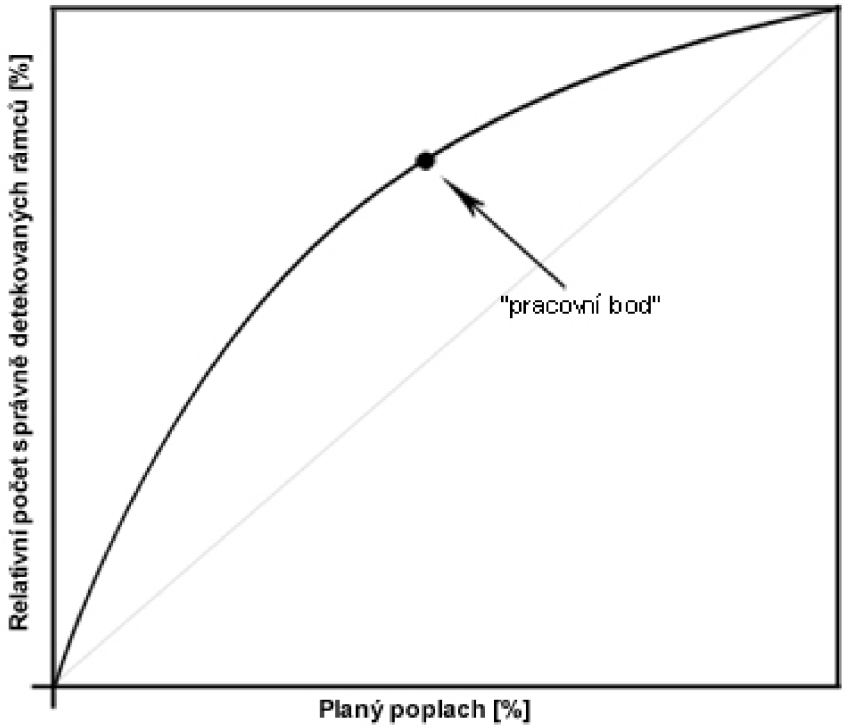
`sny.wav` – délka 44 s, mluvčí žena, bez rušení.
`sny2.wav` – délka 47 s, mluvčí muž, bez rušení.
`sny_sprcha.wav` – délka 47 s, mluvčí žena, rušení – na pozadí hluk dopadající vody ze sprchy.
`sny_pracka.wav` – délka 41 s, mluvčí muž, rušení – na pozadí hluk pračky.
`mixer.wav` – délka 23 s, mluvčí muž, rušení – nestacionární zvuk mixeru.
`vysavac.wav` – délka 21 s, mluvčí muž, rušení – hluk vysavače.

8 ROC

V této části budou popsány ROC křivky (receiver operating characteristic), které jsou použity k vyhodnocenní implementovaných detektorů řečové aktivity. Tyto grafy kompletně popisují chybovost detektorů a ukazují kompromis mezi pravděpodobností chybného určení řeči a řečové pauzy. Pro každý testovaný signál je určen relativní počet správně detekovaných řečových rámců (HR1 – hit rates 1) a relativní počet správně detekovaných rámců řečových pauz (HR0 – hit rates 0) [12]. Tyto hodnoty jsou definovány následovně:

$$HR1 = \frac{N_{1|1}}{N_1^{\text{ref}}}, HR0 = \frac{N_{0|0}}{N_0^{\text{ref}}}, \tag{8.1}$$

kde N_1^{ref} a N_0^{ref} jsou referenční hodnoty (celkový počet řečových nebo neřečových rámců) zjištěné „ruční detekcí“ a $N_{1|1}$ a $N_{0|0}$ je aktuální počet správně klasifikovaných řečových nebo neřečových rámců. „Planý poplach“ (FAR – false alarm rate) je v každém stavu definován jako FAR0 = 1 − HR1[12].



Obr. 8.1: Typický průběh ROC křivky.

Na obrázku 8 je znázorněn typický průběh ROC křivky použité pro naše účely. Pohyb „pracovního bodu“ po křivce je reakcí na změnu vybraného parametru testovaného detektoru. Polohou „pracovního bodu“ volíme „kompromis“ např. mezi

relativním počtem správně detekovaných řečových pauz (na svislé ose grafu) a relativním počtem „planých poplachů“ (na vodorovné ose), přičemž relativní počet „planých poplachů“ tvoří doplněk k relativnímu počtu správně detekovaných řečových rámců.

8.1 Typy vytvořených ROC křivek

Pro první tři implementované detektory (energetický detektor, detektor založený na krátkodobé intenzitě a kepstrální detektor) je nejprve zhotovena závislost HR0 [%] na FAR0 [%] při konstantní hodnotě parametru p (časová konstanta exponenciálního průměrování používaná při aktualizaci prahové hodnoty), která je volena na základě předchozího experimentálního testování. Proměnným parametrem je v tomto případě koeficient alfa (pomocná konstanta pro stanovení prahové hodnoty a pro obnovnu prahu), který se nastavuje od hodnoty 0 po 1 s krokem $0,1$.

Pro všechny implementované detektory jsou zhotoveny následující dva typy ROC křivek.

Druhým typem křivek je závislost HR1 [%] na SNR [dB]. Tato závislost byla testována na nezarušených signálech, ke kterým byl aditivně přidán bílý šum s odstupem signálu od šumu $[-5, 0, 5, 10, 20, 30]$ dB . Hodnota parametu p byla opět konstantní, stejně tak i hodnota koeficientu alfa, která je zvolena na základě testování prvního typu ROC křivek (kompromis mezi HR0 a FAR0).

Posledním typem křivek je závislost HR0 [%] na FAR0 [%] a to buď při změně SNR $= [-5, 0, 5, 10, 20, 30]$ dB provedené opět aditivním přidáním bílého šumu (pro nezarušené signály), nebo při SNR $= 30$ dB pro signály s jiným charakterem rušení.

9 VYHODNOCENÍ NAMĚŘENÝCH HODNOT

Z prvního typu ROC křivek lze vyčíst, že se zvyšující se hodnotou parametru alfa se zvyšuje taky relativní počet planých poplachů FAR0 a zároveň i relativní počet správně detekovaných rámců řečových pauz HR0.

Na základě těchto křivek jsem jako kompromis mezi HR0 a FAR0 zvolil optimální hodnotu alfa. Například u prvního signálu (graf A.3) jsem volil pro energetický detektor alfa_e $= 0,1$, pro detektor intenzity alfa_i $= 0,2$ a pro kepstrální detektor alfa_k $= 0,6$.

U nezarušených signálů se pro tyto parametry alfa jeví obdobná účinnost detektoru intenzity a kepstrálního detektoru a to kolem 90 % HR0 a 26 % FAR0. Pro energetický detektor je to potom 96 % HR0 a 36 % FAR0. U nezarušených signálů je vhodné volit pro energetický detektor hodnoty koeficientu alfa v rozmezích od 0 do $0,2$.

U zarušených signálů dosáhl největší účinnosti při vhodně nastaveném koeficientu alfa detektor intenzity. Nejhůře se naopak u zarušených signálů projevoval energetický detektor, který dosahoval ve většině případů příliš vysokých hodnot FAR0, což znamená příliš malé hodnoty (pod 50 %) HR1. Při detekci nahrávky „mixer.wav“ (signál zarušený zvukem mixeru) obstál pouze detektor intenzity. Energetický detektor a kepstrální detektor není vhodné pro tento typ rušení používat (graf A.2).

Pro zhodnocení vlastností detektoru G.729 a kombinovaného detektoru je zkonstuována ROC křivka A.2 a A.2 (grafy závislosti HR1 na SNR $= [-5, 0, 5, 10, 20, 30]$ dB) pro nezarušené signály s aditivním bílým šumem a ROC křivka A.2 (graf závislosti HR0 na FAR0) pro signály s různým charakterem rušení. Detektor G.729 dosahoval od odstupu signálu od šumu 5 dB a více nejlepších a nejvyrovnanějších výsledků ze všech implementovaných detektorů. Jeho účinnost dosahuje v průměru hodnot HR1 = 92 %, při HR0 = 83 % (viz. graf A.2). Při vyšších hodnotách rušení však prudce klesají hodnoty HR0 a tento detektor se jeví jako nepoužitelný k detekci řečové aktivity takto zarušených signálů. U signálů s různým charakterem rušení dosáhl nejlepší detekce detektor G.729 s průměrem hodnot HR1 = 94 %, HR0 = 70 % (viz. graf A.2). Dobrých výsledků dosáhl taky detektor intenzity s průměrem hodnot HR1 = 79 %, HR0 = 85 %.

Kombinovaný detektor u těchto signálů dosahuje poměrně vyrovnaných hodnot v průměru HR1 = 80 %, HR0 = 70 %. U nezarušených signálů kombinovaný detektor prokazuje v zásadě lepší vlastnosti, než energetický detektor, ale jeho účinnost je nižší při detekci řečových rámců (HR1), než účinnost kepstrálního detektoru a detektoru intenzity (viz. graf A.2).

10 ZÁVĚR

Cílem diplomové práce, zaměřené na rozpoznávače řečové aktivity, bylo seznámit se s vybranými metodami a algoritmy detekce řečové aktivity.

V této práci byly krátce popsány standardní metody detekce řečové aktivity (energetický detektor, detektor založený na krátkodobé intenzitě, kepstrální detektor). Kombinací těchto tří detektorů byl v rámci této diplomové práce vytvořen kombinovaný detektor. Dále byl popsán detektor podle standardu ITU-T G.729. Všechny tyto metody detekce řečové aktivity byly implementovány v MATLABu.

Pro testování implementovaných metod detekce bylo v MATLABu vytvořeno uživatelské prostředí s možností volby vstupního signálu a parametrů detekce pro jednotlivé detektory. Toto uživatelské prostředí slouží k testování a porovnání funkce všech implementovaných detektorů řečové aktivity. Vyhodnocení je provedeno pomocí ROC křivek. Před samotným testováním detektorů se nastavují vstupní parametry tak, aby bylo dosaženo nejlepších výsledků detekce jednotlivých detektorů.

Po implementaci bylo provedeno odzkoušení vytvořených detektorů na šesti nahrávkách, s různým charakterem rušení a odstupem signálu od šumu, které byly namluveny třemi různými mluvčími.

K vyhodnocení testování byly použity ROC křivky, popsány v kapitole 8, výsledné grafy jsou ukázány v přílohách diplomové práce (viz. A.3). Vytvořil jsem tři typy těchto křivek (viz. kapitola 8.1).

Na základě testování a vyhodnocení funkcí detektorů (kapitola 9) lze konstatovat následující.

Účinnosti energetického detektoru, kepstrálního detektoru a detektoru intenzity, jsou u nezarušených signálů (při vhodném nastavení vstupních parametrů) obdobné, přičemž detektor intenzity dosáhl poměrně dobrých výsledků i u signálů zarušených různým charakterem rušení. Naopak energetický detektor nedoporučuji pro detekci řeči u zarušených signálů používat.

Implementovaný kombinovaný detektor je výpočetně náročnější, než samostatné detektory, ale žádný výrazný přínos pro detekci řečové aktivity nepřináší.

Po stránce výpočetní náročnosti a složitosti implementace je nejnáročnější detektor definovaný standardem ITU-T G.729, tento detektor však z implementovaných detektorů přináší celkově nejlepší a nejvyrovnanější výsledky detekce řečové aktivity. Není však vhodný k detekci řeči s odstupem signálu od šumu nižším, než 5 dB.

LITERATURA

[1] PORUBA, J. *Separace řečového signálu ze šumového prostředí.* Dizertační práce. VUT, Brno 2003.

[2] PSUTKA, J. *Komunikace s počítačem mluvenou řečí* 1. vydání. Praha: Academia, 1995. 287 s. ISBN 80-200-0203-0.

[3] PORUBA, J., MATĚJÍČEK, L. *Odfiltrování rušivých signálů ze zašumělé řeči.* Článek Elektrorevue, VUT Brno 2002 [cit. 2007-12-10]. Dostupné z: (http://www.elektrorevue.cz/clanky/02047/index.html)

[4] VONDRÁŠEK, M. *Odhad SNR řečového signálu snímaného v hlučném prostředí.* Diplomová práce. ČVUT, Praha 2004.

[5] SIGMUND, M. *Analýza řečových signálů.* 1. vydání. VUT Brno 2000. ISBN 80-214-1783-8

[6] TATARINOV, J. *Detektory řečové aktivity založené na skrytých Markovových modelech.* Analýza a zpracování řečových a biologických signálů, SBORNÍK PRACÍ 2005. ČVUT Praha 2005. ISBN 80-01-03412-7

[7] BÁRTA, P. *Detektor hlasové aktivity.* Diplomová práce, VUT, Brno 2007.

[8] ADAMEC, M. *Detektory řečové aktivity a jejich spolehlivost u zarušených signálů.* Bakalářská práce, VUT, Brno 2006

[9] ITU-T. *Annex B: A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70* November 1996 [cit. 2007-23-11]. Dostupné z: (http://www.itu.int/rec/T-REC-G.729-199610-I!AnnB/en)

[10] ITU-T. *Coding of speech at 8 kbit/s using conjugate-structure algebraic-cod-excited linear-prediction (CS-ACELP)* Marz 1996 [cit. 2007-23-11]. Dostupné z: (http://www.itu.int/home/index.html)

[11] DELLER, J.R., HANSEN, J.H.L., PROAKIS, J.G. *Discrete-Time Processing of Speech Signals (IEEE Press Classic Reissue)* Wiley-IEEE Press; Reprint edition, 1999. ISBN 0780353862

[12] GÓRRIZ J.M., RAMÍREZ J., LANG E.W., PUNTONET C.G. *Hard C-means clustering for voice activity detection* Speech Communication 48, July 2006.

11 DODATKY

11.1 Váhovací funkce – vážení okénkem

Řečový signál má tu vlastnost, že se v čase mění „pomalu“. Hlasové ústrojí člověka je schopno změnit své parametry minimálně za 10ms. Řečový signál je tedy zpracováván metodami tzv. krátkodobé analýzy.

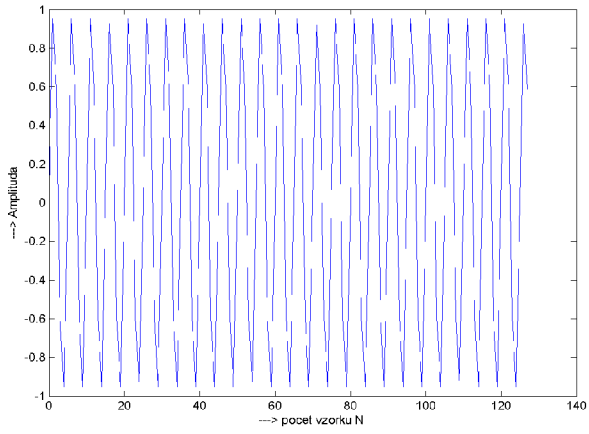
Signál je rozdělen na segmenty (frame) o délce N vzorků a každý segment se poté zpracovává samostatně. Sousední segmenty se mohou překrývat. Délka segmentu se pohybuje v intervalu 10 až 25ms, což souvisí schopností změny lidského hlasového ústrojí.

Úkolem okénka je vybrat příslušné vzorky signálu a přidělit jim při zpracování určitou váhu.

Nejčastěji používaná okénka, při zpracování signálu v časové oblasti, jsou pravoúhlé a Hammingovo okénko [3]:

- pravoúhlé okénko: aplikuje stejnou váhu na všechny vzorky signálu.

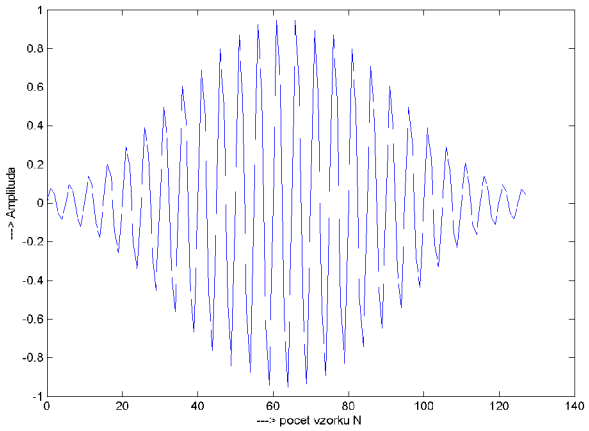
$$\begin{aligned} w(n) &= 1 \quad \text{pro } n = 1, 2, \ldots, N \\ w(n) &= 0 \quad \text{pro ostatní } n \end{aligned}$$



Obr. 11.1: Sinusový signál vážený pravoúhlým okénkem pro N=128.

- Hammingovo okénko: potlačí vzorky na okrajích segmentu, tím zvyšuje stabilitu některých výpočtů. Má schopnost potlačit postranní laloky ve spektru, které nám u obdélníkového okna zkreslují skutečnou podobu spektra.

$$w(n) = 0{,}54 - 0{,}46\cos(2\pi n/N) \quad \text{pro } n = 1, 2, \ldots, N$$
$$w(n) = 0 \quad \text{pro ostatní } n$$



Obr. 11.2: Sinusový signál vážený Hammigovým okénkem pro N=128.

Přesto, že je pravoúhlé okno jednodušší, často se upřednostňuje použití Hammingova okna. Hammingovo okno totiž potlačuje vzorky na okrajích segmentů, čímž se zvyšuje stabilita některých výpočtů.

11.2 Odstup signálu od šumu (SNR)

Základní definice odstupu signálu od šumu je dána vztahem:

$$SNR = 10\log\frac{P_s}{P_n} = 10\log\frac{\sigma_s^2}{\sigma_n^2}, \tag{11.1}$$

kde P_s je výkon užitečného signálu a P_n je výkon šumu [4].

Za předpokladu, že řečový signál i šumové pozadí jsou náhodné stacionární signály s nulovou střední hodnotou, jsou výkony signálů dány jejich rozptyly σ_s^2 a σ_n^2.

11.2.1 Odhad SNR s výkonovým odečítáním

V praktických aplikacích měření SNR máme k dispozici jen směs x řečového signálu s a šumu n:

$$x = s + n. \tag{11.2}$$

Výkony řeči i šumu je tak nutno odhadnout z jediného signálu. Budou-li řeč a šum nekorelované signály, pak pro výkon směsi platí:

$$\sigma_x^2 = \sigma_s^2 + \sigma_n^2. \tag{11.3}$$

Vztah pro výpočet SNR (11.1) lze potom přepsat:

$$\overline{\mathrm{SNR}} = 10 \log \frac{\overline{\sigma}_s^2}{\overline{\sigma}_n^2} = 10 \log \frac{\sigma_s^2 - \overline{\sigma}_n^2}{\overline{\sigma}_n^2}, \tag{11.4}$$

kde pruh nad znakem značí, že se jedná o „odhad“, v našem případě odhad výkonu šumu.

Úloha je tak zjednodušena jen na nalezení odhadu výkonu šumu. Z principu je však obtížné rozlišit ve výkonové oblasti výkon řeči a výkon šumu z jediného signálu. Existují dva přístupy k nalezení odhadu šumu. První skupina využívá informace o řečové aktivitě. Výkon šumu se získává průměrováním v řečových pauzách. Druhá skupina odhaduje výkon šumu sledováním minima krátkodobého výkonu.

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

B	počet kvantizačních bitů v binárním kódu
CCITT	Comité Consultatif International Télégraphique et Téléphonique
c_l	reálné kepstrum
E_d	energie hluku pozadí
E_f	úzkopásmová energie
E_l	širokopásmová energie
E_n	krátkodobá energie
E_p	prahová energie
HR0	„Planý poplach“ – false alarm rate
f_vz	vzorkovací kmitočet
F_0	základní tón řeči
F_m	maximální kmitočet
H	hlasitost
HMM	skrytý Markovovův model – Hidden Markovov‘s Model
HR1	relativní počet správně detekovaných rámců řeči – hit rates 1
HR0	relativní počet správně detekovaných rámců řečových pauz – hit rates 0
I	intenzita zvuku
I_0	referenční prahová intenzita zvuku
ITU	Mezinárodní telekomunikační unie – International Telecommunication Union
ITU-T	telekomunikační sektor ITU – Telecommunication Standardization Sector of ITU
I_VD	výstup detekce hlasové aktivity ve VAD G.729
L_m	délka membrány hlasivek
LPC	lineární predikční kódování – Line Predictive Coding

LSF	kmitočet spektrálních párů – Line Spectral Frequency
LSF	spektrální páry – Line Spectral Pars
M_{n}	krátkodobá intenzita
ROC	Receiver Operating Characteristic
SNR	odstup signálu od šumu – Signal Noise Ratio
T_0	základní perioda řeči
T	perioda vzorkování
VAD	detektor řečové aktivity – Voice Activity Detector
ZC	střední počet průchodů signálu nulou

SEZNAM PŘÍLOH

A PŘÍLOHY

A.1 Tabulka konstant pro detektor G.729

Tab. A.1: Tabulka konstant

Název	**Konstanta**	**Název**	**Konstanta**
N_i	32	N_1	4
N_0	128	N_2	10
K_0	0	T_1	671088640
K_1	−53687091	T_2	738197504
K_2	−671088640	T_3	26843546
K_3	−93952410	T_4	40265318
K_4	−134217728	T_5	40265318
K_5	−161061274	T_6	40265318
a_1	23488	b_1	28521
a_2	−30504	b_2	19446
a_3	−32768	b_3	−32768
a_4	26214	b_4	−19661
a_5	0	b_5	−30802
a_6	28160	b_6	−19661
a_7	0	b_7	30199
a_8	16384	b_8	−22938
a_9	−19065	b_9	−31576
a_{10}	0	b_{10}	−173671
a_{11}	22400	b_{11}	−27034
a_{12}	30427	b_{12}	29959
a_{13}	−24576	b_{13}	−29491
a_{14}	23406	b_{14}	−28087

A.2 Obsah přiloženého CD

Přiložené CD obsahuje tři adresáře:

`/nahravky`
obsahuje nahrávky použité pro testování detektorů. Charakter nahrávek je popsán v kapitole 7.3.

`/text`
obsahuje text diplomové práce ve formátu *.pdf

`/funkce`
obsahuje tři podadresáře:

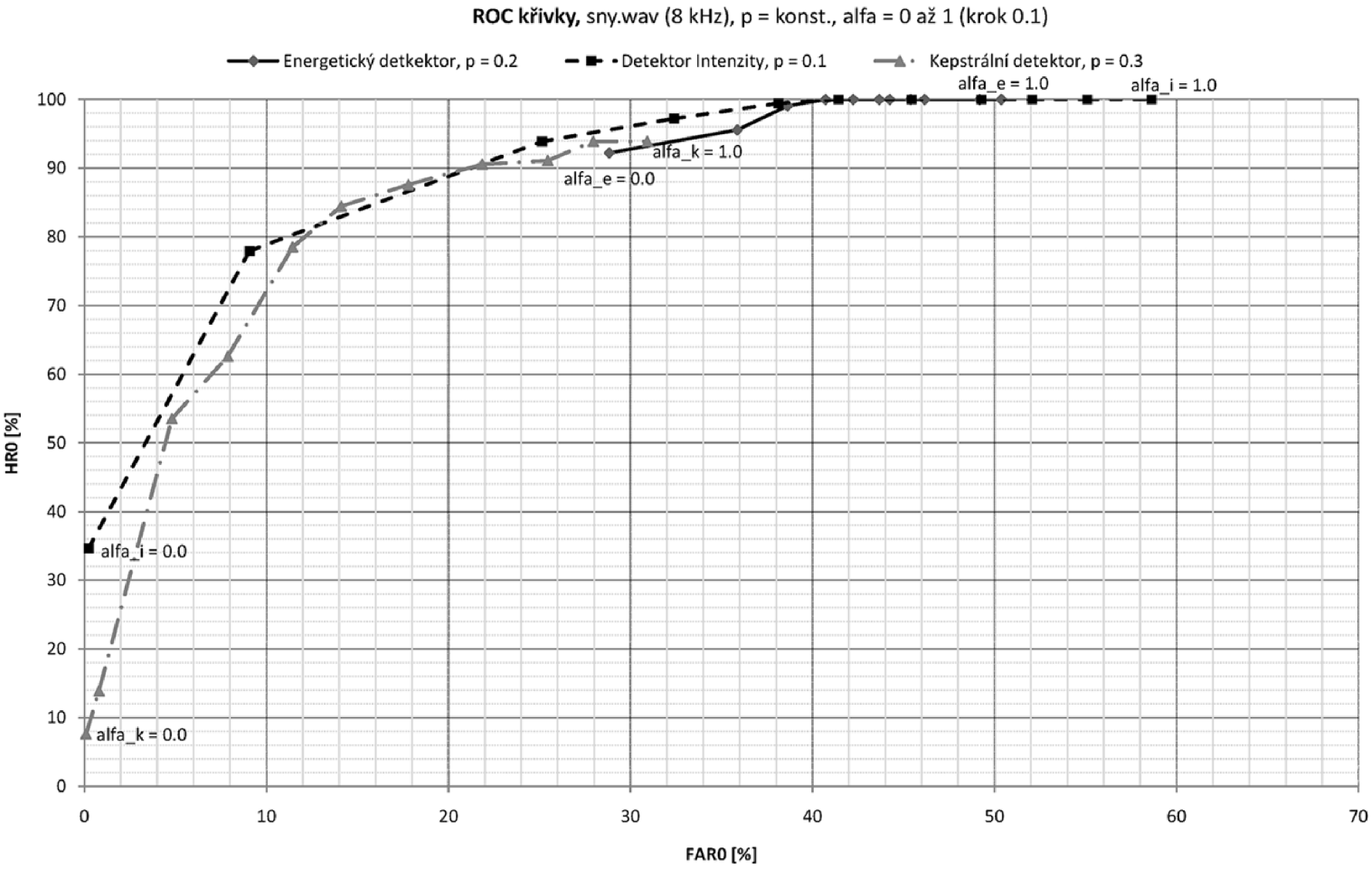
1. podadresář: `/jednotlive_detektory`
Tento podadresář obsahuje adresáře s výslednými m-funkcemi MATLABu pro detekci řečové aktivity.

2. podadresář: `/roc_testovani`
Tento podadresář obsahuje m-funkce MATLABu, které byly použity k testování vytvořených detektorů a k sestrojení ROC charakteristik.

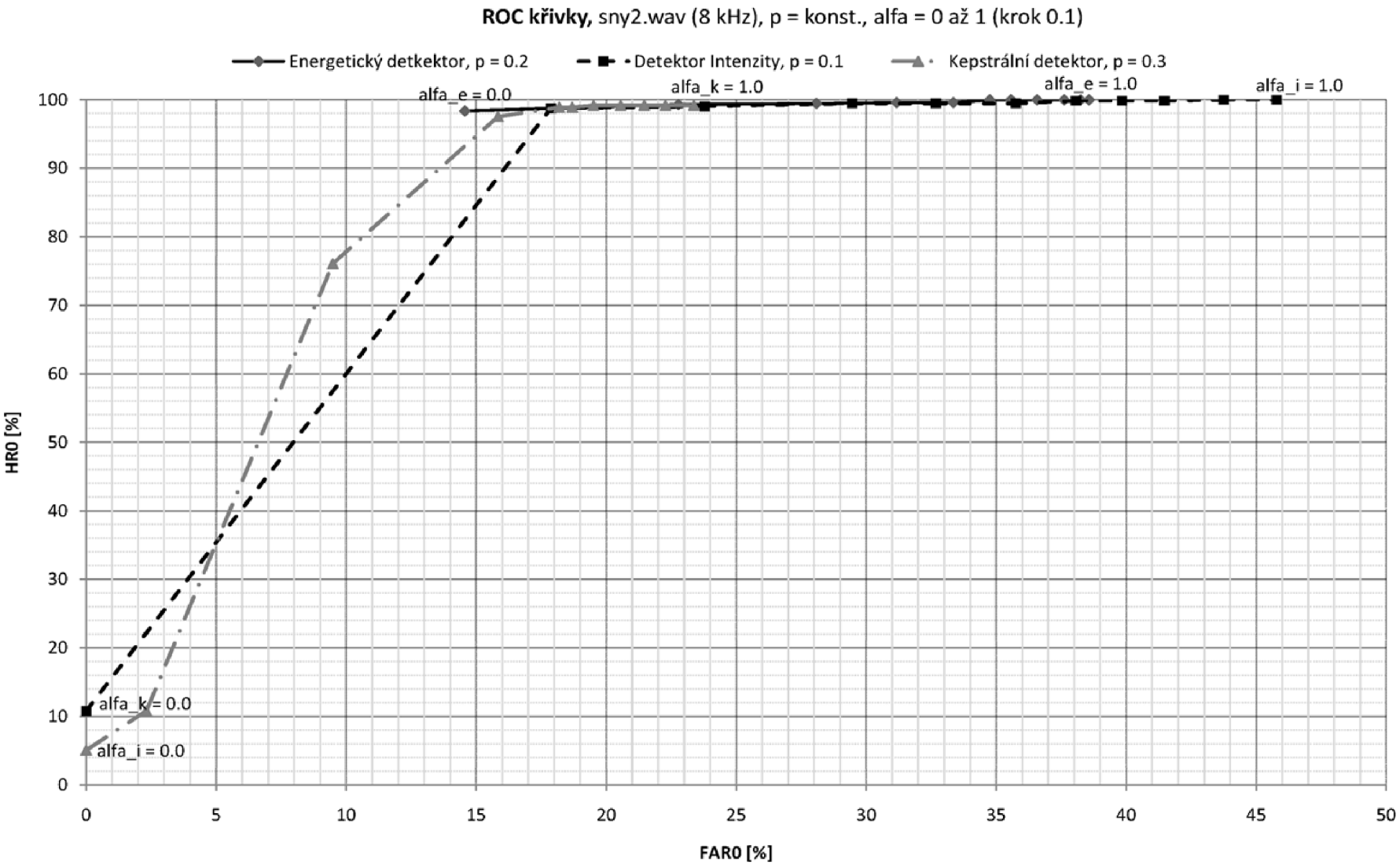
3. podadresář: `/uzivatelske_rozhrani`
Tento podadresář obsahuje m-funkce MATLABu pro uživatelské rozhraní. Uživatelské rozhraní se spouští zapsáním příkazu „GUIDE“ do příkazového řádku MATLABu. Poté se v dialogovém okně vybere záložka „Open Existing GUI“ a zvolí se soubor „guide_vad.fig“.
Pozn. k vytvoření uživatelského rozhraní byl použit MATLAB verze 7.2.0.232 (R2006a).

Podadresář `/uzivatelske_rozhrani` ještě obsahuje dva podadresáře: /dat a /wav.
`/dat` – zde jsou uloženy soubory typu *.dat, které obsahují značky začátků a konců řečových pauz v jednotlivých řečových nahrávkách.
`/wav` – obsahuje řečové nahrávky.

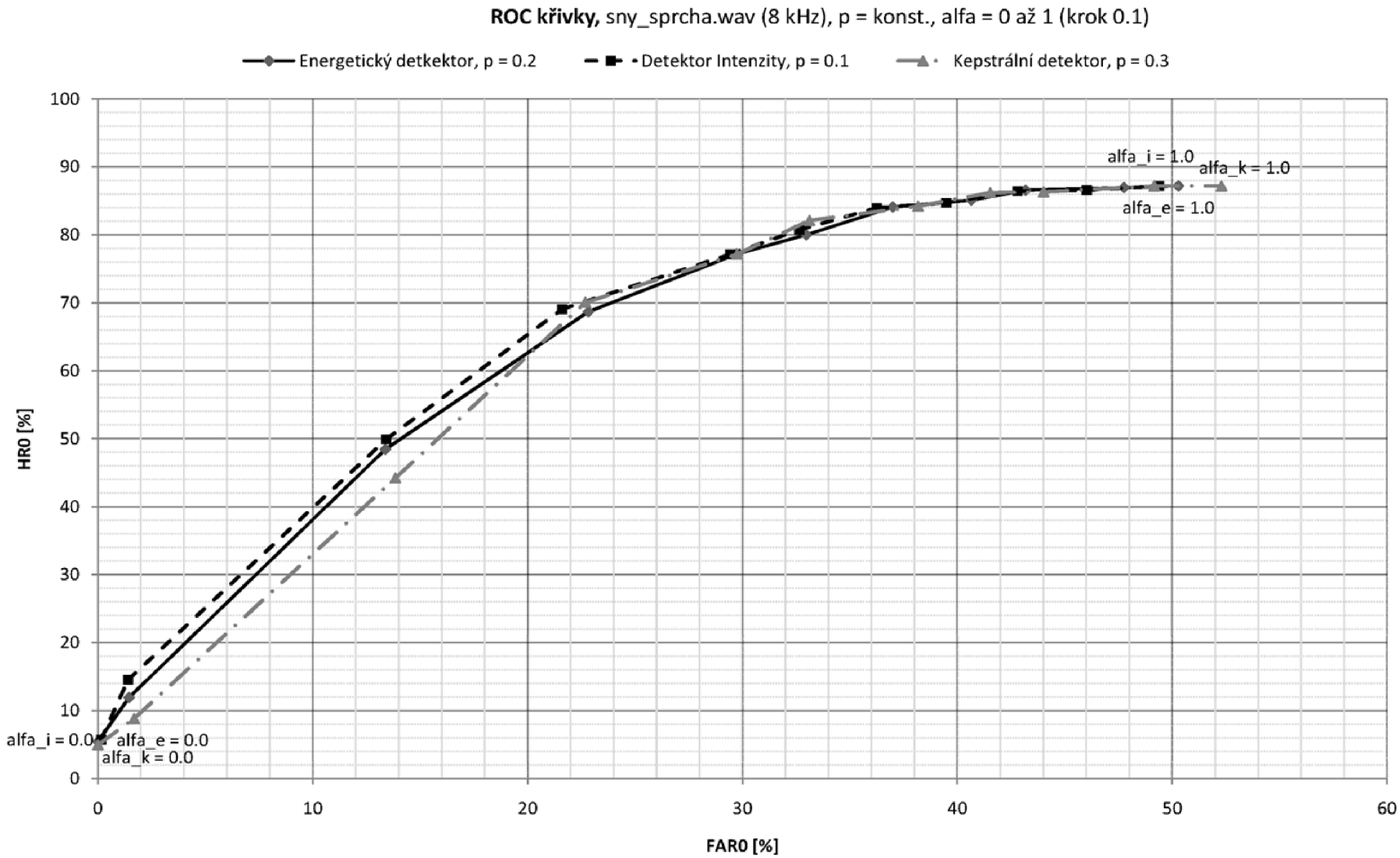
A.3 ROC křivky



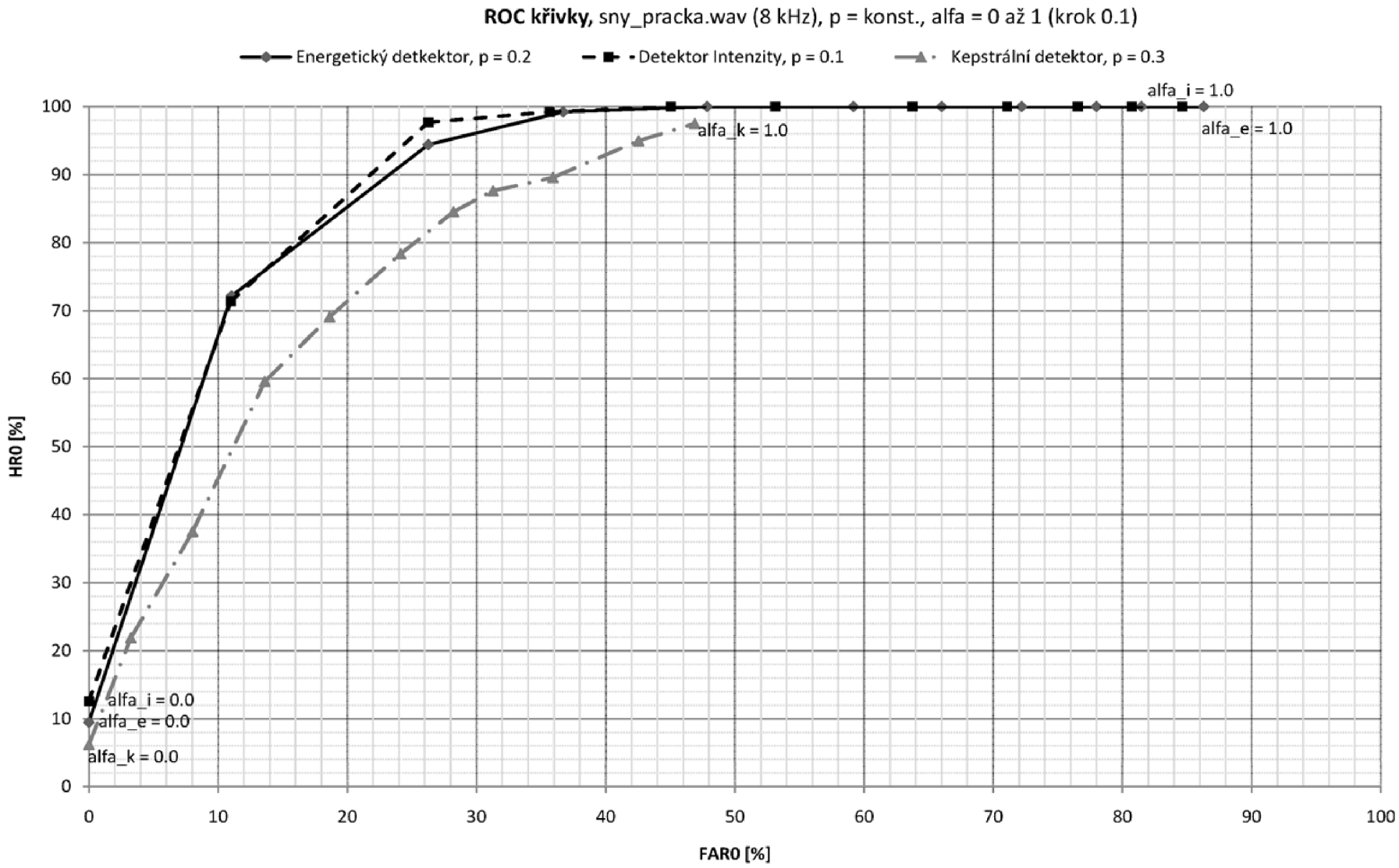
Obr. A.1: ROC křivky. Řečový signál sny.wav, p = konst., alfa = 0 až 1 (krok 0,1).



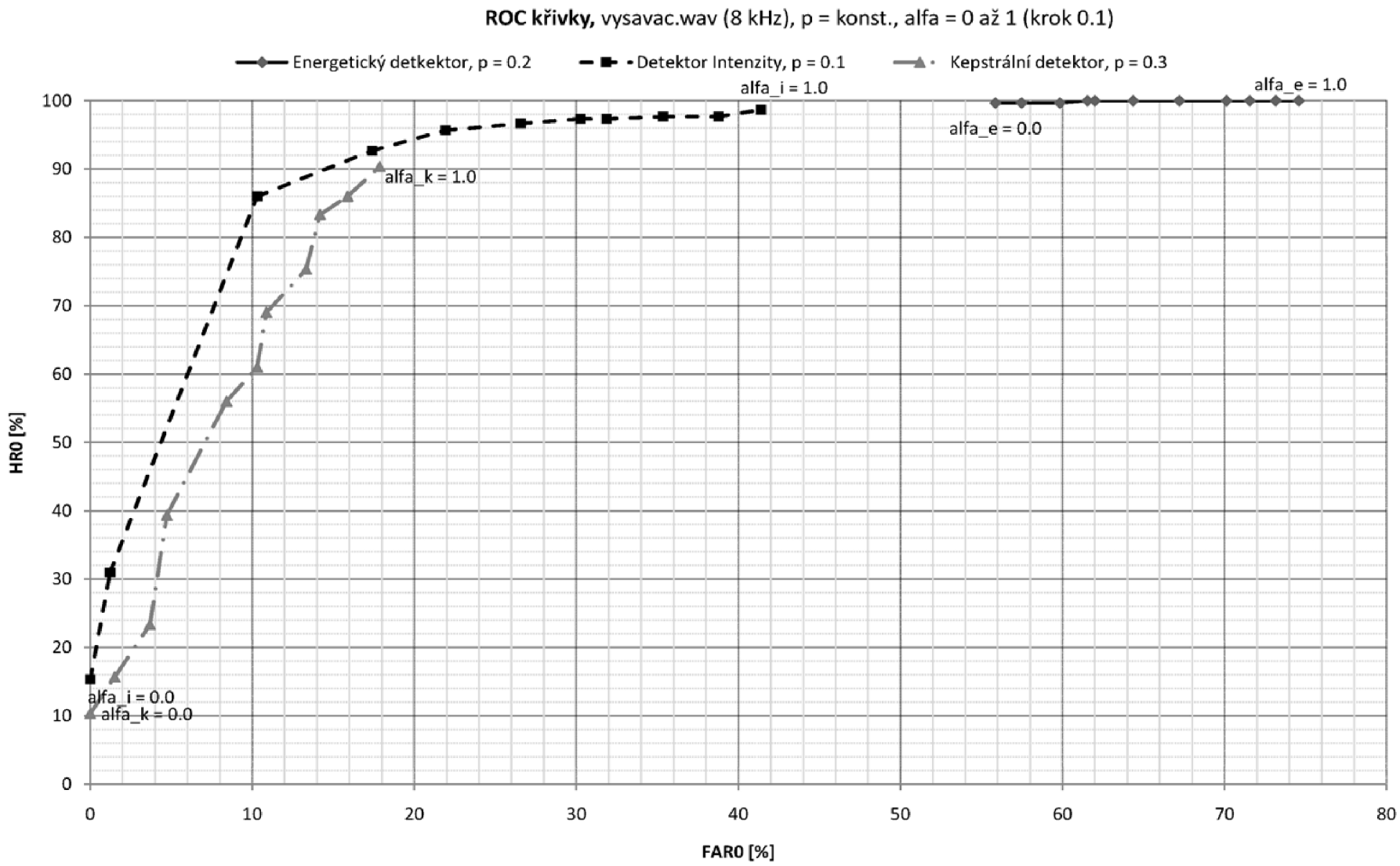
Obr. A.2: ROC křivky. Řečový signál sny2.wav, p = konst., alfa = 0 až 1 (krok 0,1).



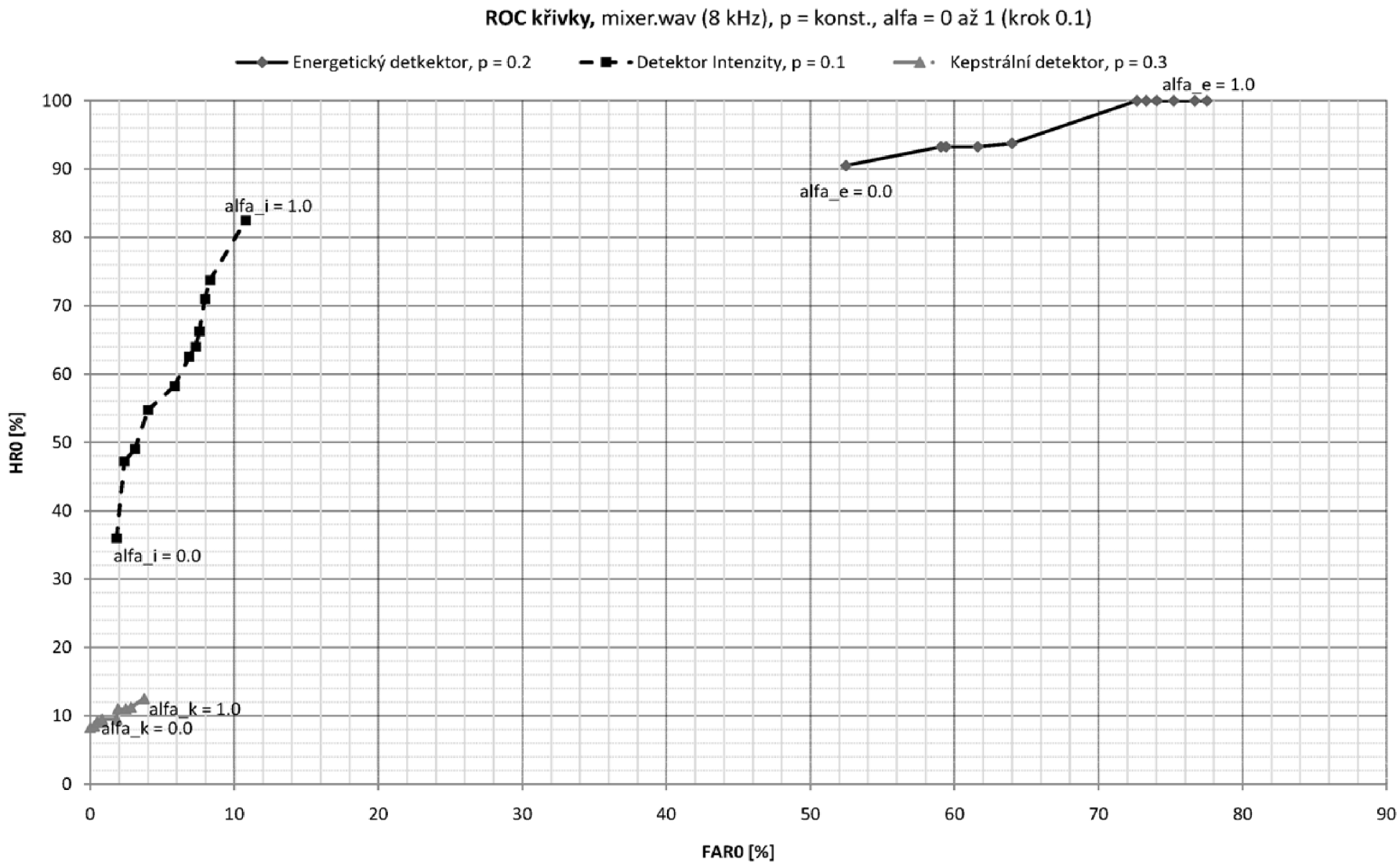
Obr. A.3: ROC křivky. Řečový signál sny_sprcha.wav, p = konst., alfa = 0 až 1 (krok 0,1).



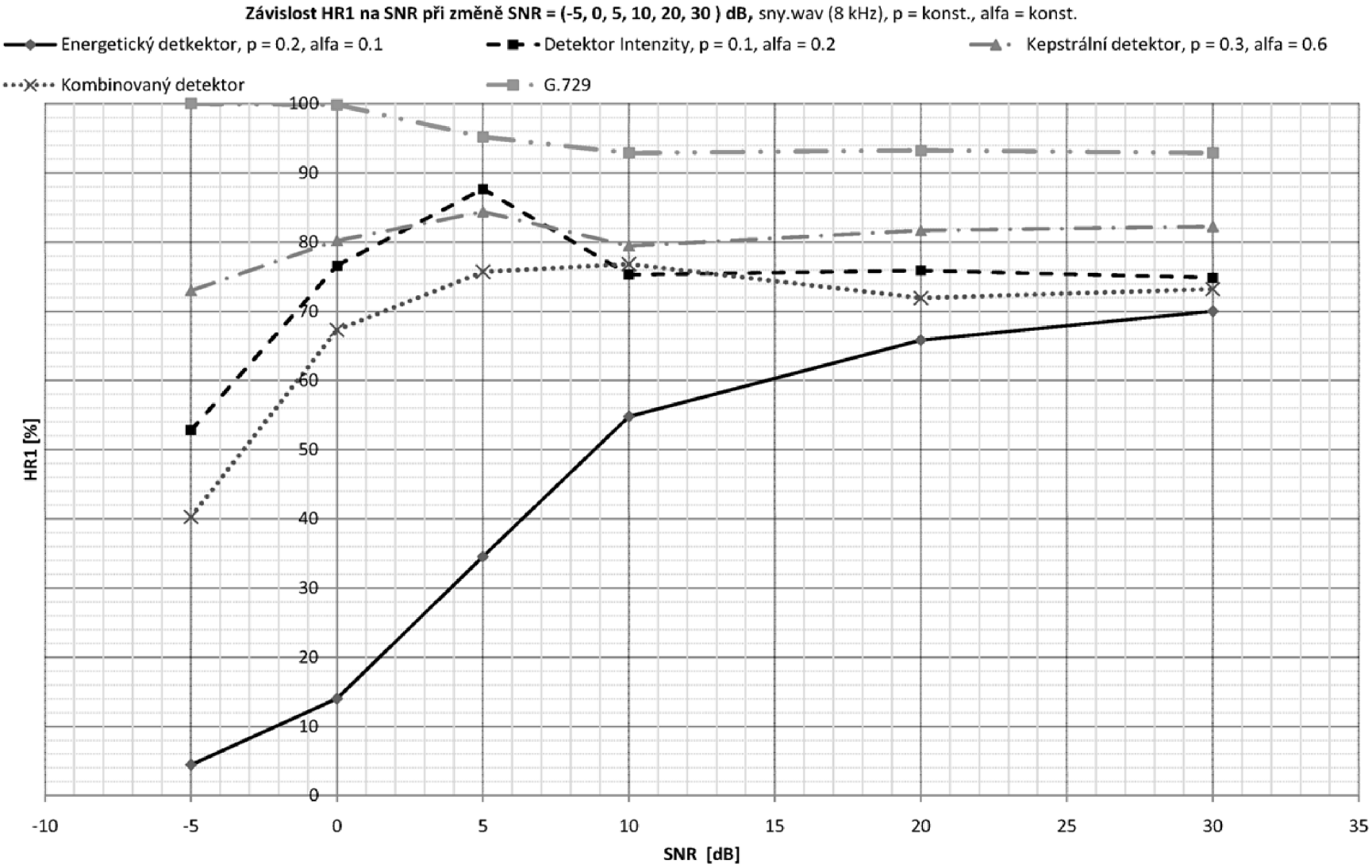
Obr. A.4: ROC křivky. Řečový signál sny_pracka.wav, p = konst., alfa = 0 až 1 (krok 0,1).



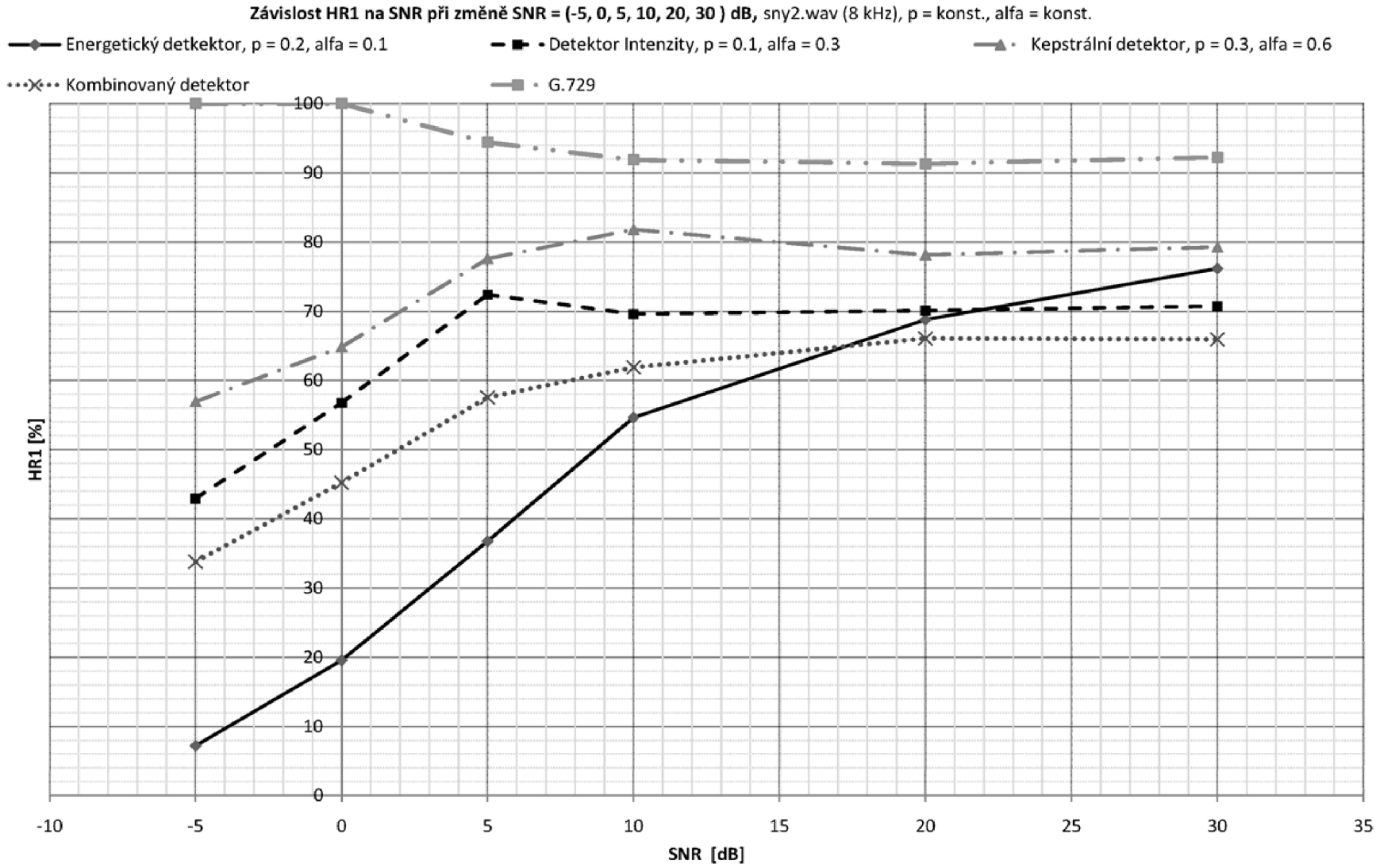
Obr. A.5: ROC křivky. Řečový signál vysavac.wav, p = konst., alfa = 0 až 1 (krok 0,1).



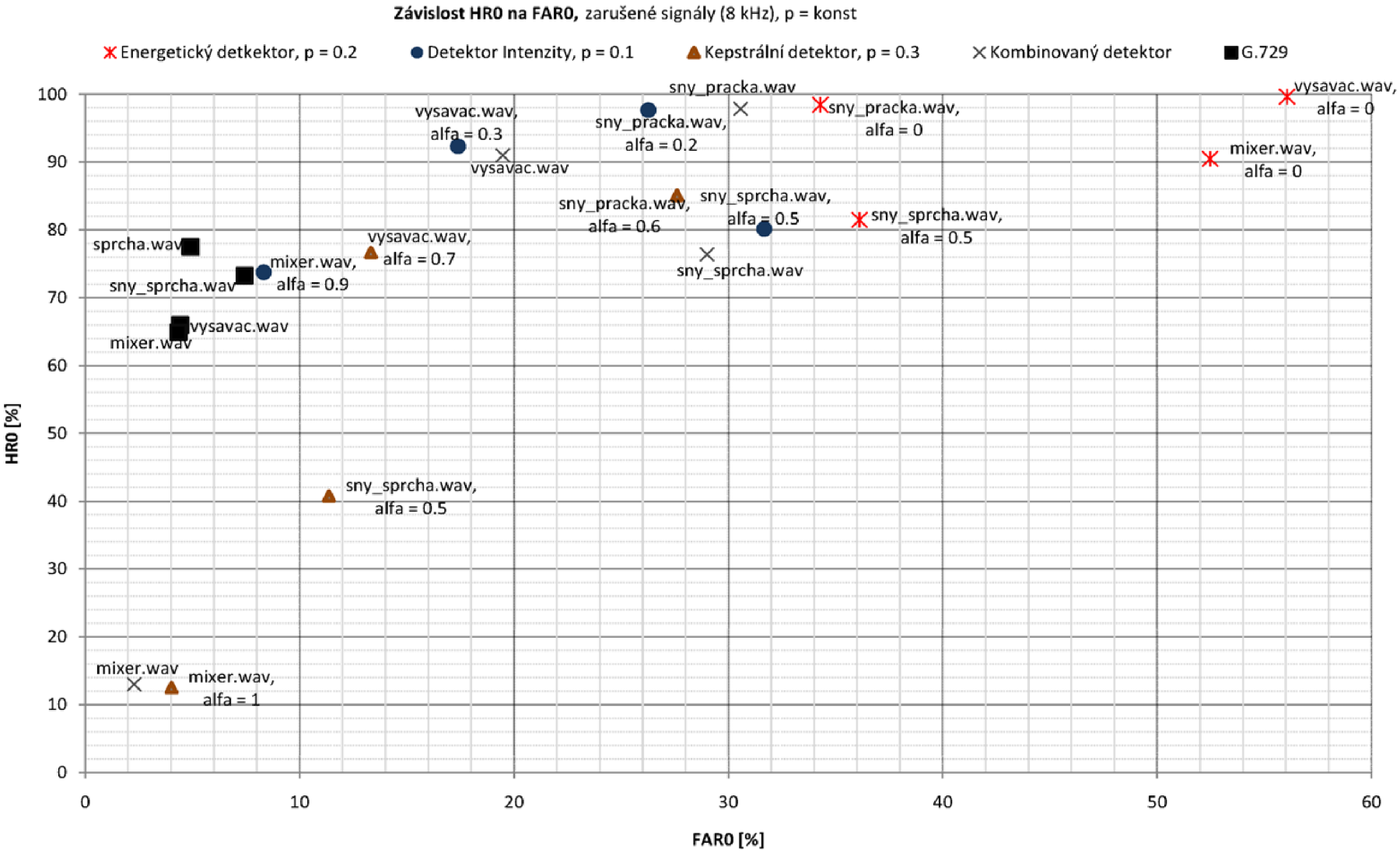
Obr. A.6: ROC křivky. Řečový signál mixer.wav, p = konst., alfa = 0 až 1 (krok 0,1).



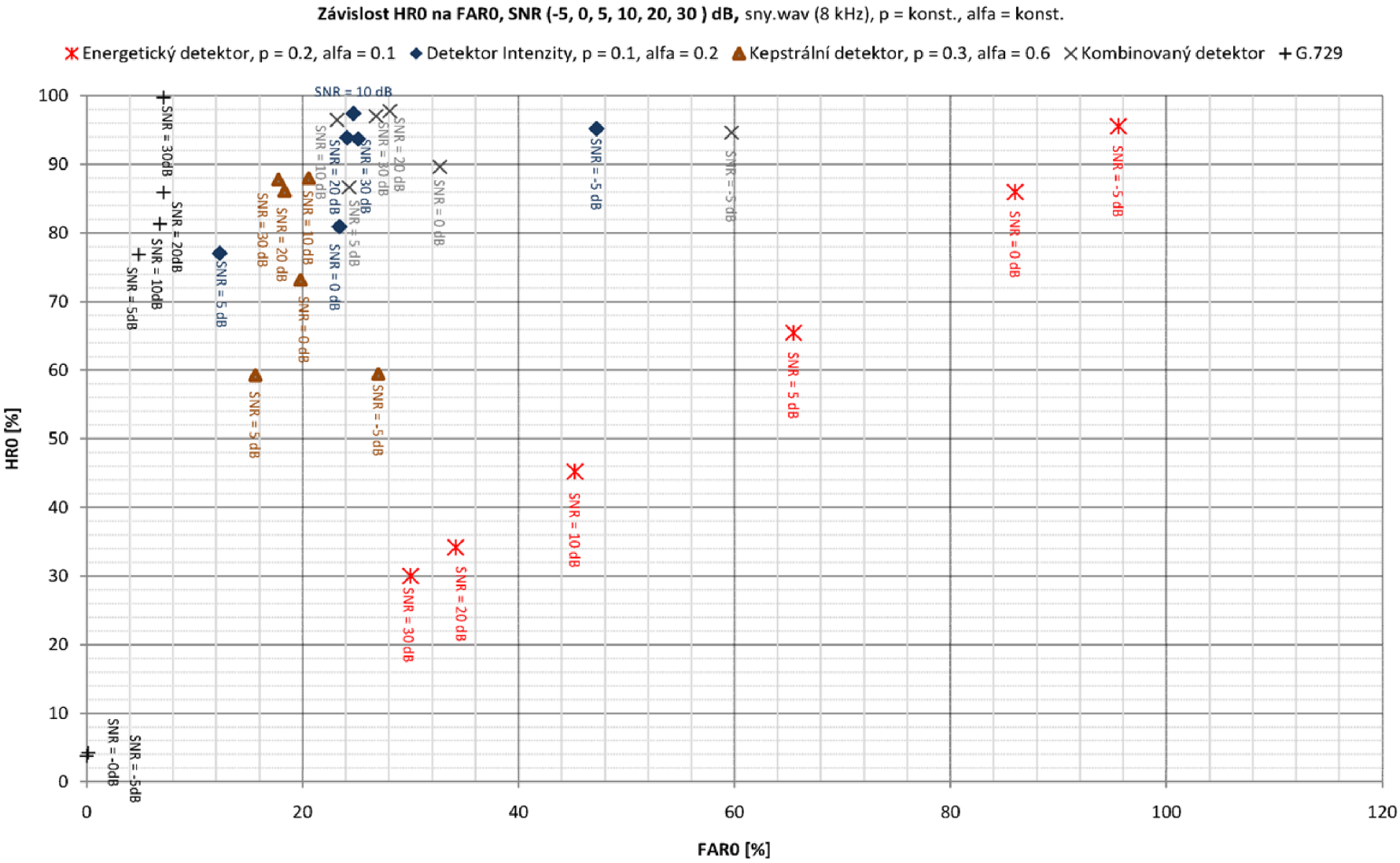
Obr. A.7: Závislost HR1 na SNR při změně SNR = $(-5, 0, 5, 10, 20, 30)$ dB. Řečový signál sny.wav, p = konst., alfa = konst.



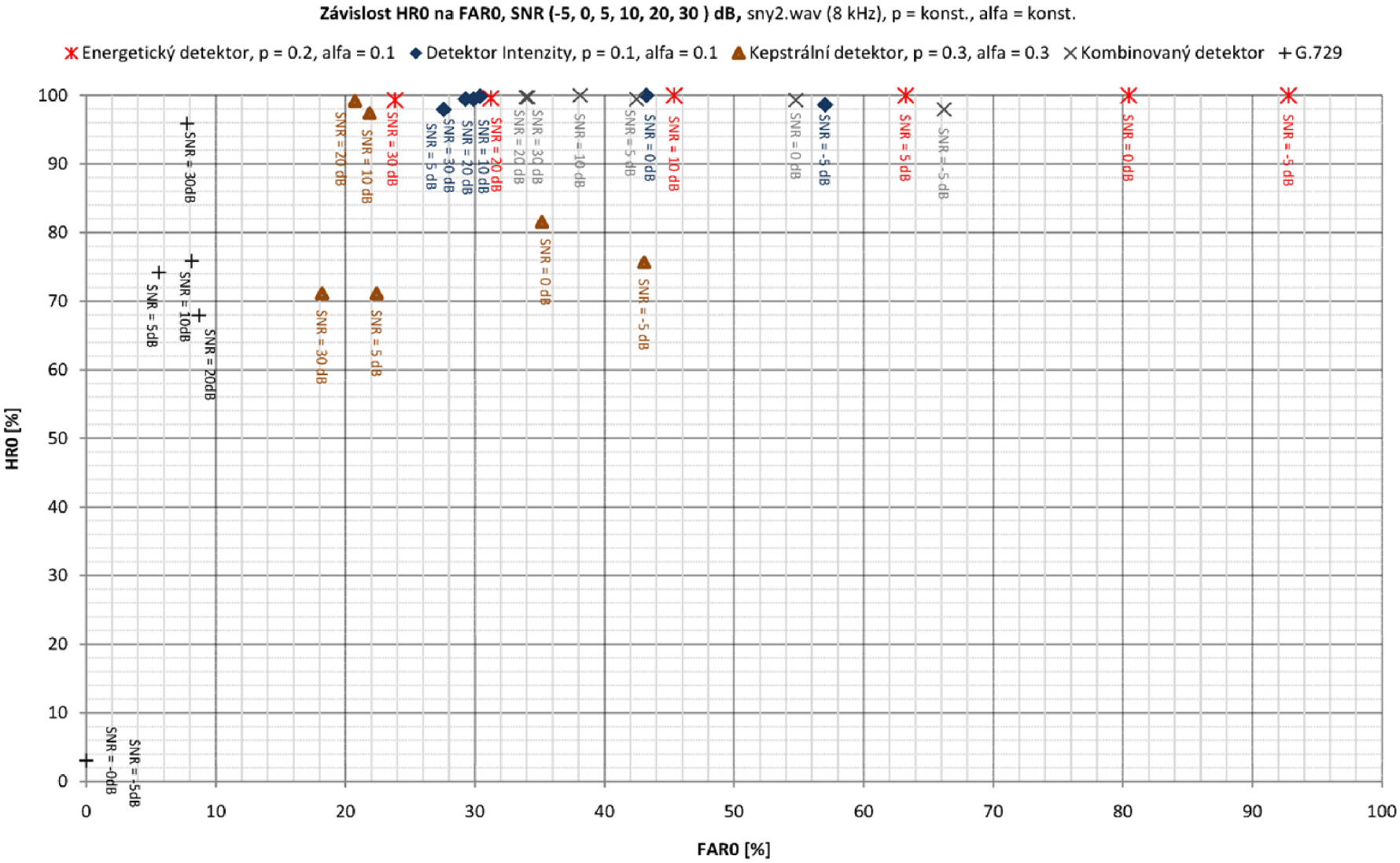
Obr. A.8: Závislost HR1 na SNR při změně SNR = $(-5, 0, 5, 10, 20, 30)$ dB. Řečový signál sny2.wav, p = konst., alfa = konst.



Obr. A.9: Závislost HR0 na FAR0 pro zarušené signály.



Obr. A.10: Závislost HR0 na FAR0, při změně SNR = $(-5, 0, 5, 10, 20, 30)$ dB. Řečový signál sny.wav, p = konst., alfa = konst.



Obr. A.11: Závislost HR0 na FAR0, při změně SNR = $(-5, 0, 5, 10, 20, 30)$ dB. Řečový signál sny2.wav, p = konst., alfa = konst.