VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

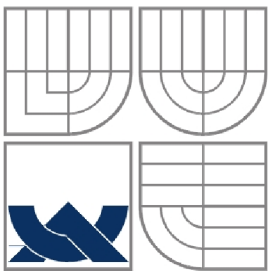
AUTOMATICKÁ TVORBA TITULKŮ K FILMU S ROZPOZNÁVAČEM ŘEČI

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

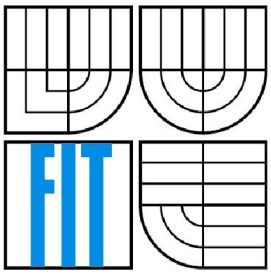
AUTOR PRÁCE
AUTHOR

GYÖRGY CSINTALAN

BRNO 2011



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

AUTOMATICKÁ TVORBA TITULKŮ K FILMU S ROZPOZNÁVAČEM ŘEČI

AUTOMATIC GENERATING OF SUBTITLES BY SPEECH RECOGNIZER

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

GYÖRGY CSINTALAN

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. PETR SCHWARZ, Ph.D.

BRNO 2011

Abstrakt

Tato bakalářská práce popisuje vývoj aplikace pro automatickou tvorbu titulků k filmům pomocí BSAPI (Brno Speech Application Interface). V úvodní části, je čtenář uveden do problematiky a jsou naznačeny cíle práce. Následuje popis teoretického základu zpracování řeči a popis rozpoznávače a jeho API (BSAPI). Dále je uveden způsob extrakce zvukové stopy z videa z teoretického hlediska. V nasledující části jsou popsány principy zpracování výstupu rozpoznávače a vytvoření titulků. V další kapitole je popsána tvorba aplikáce z implementačního hlediska. Dále následuje popis experimentů vykonaných v různych situacích a jsou navrženy další možné způsoby zlepšení kvůli dosáhnutí kvalitnejšího výstupu, například filtrování zvuku pomocí Wienerova filtru. V závěrečné části jsou rozebrány dosáhnuté výsledky a získané zkušenosti.

Abstract

This bachelor thesis describes developing an application for automatic generation of subtitles for films using BSAPI (Brno Speech Application Interface). At first the reader is guided to the problematics of this task and the aim of this work is described. In the next section the speech recognition API (BSAPI) is described from a theoretical point of wiev. After this, the next section is about extracting voice channel from video. The next chapter describes the implementation of the application. Further, experiments in different situations are described and possible improvements are discussed in order to achieve better outputs, for example speech enhancement by Wiener filter. In the conclusion are discussed achieved results and experiments.

Klíčová slová

automatická tvorba titulků, rozpoznávání řeči, BSAPI, directshow, AVI, SRT, redukování šumu, Wienerův filter

Keywords

automatic generating of subtitles, speech recognition, BSAPI, directshow, AVI, SRT, noise reduction, Wiener filter

Citace

György Csintalan: Automatická tvorba titulků k filmu s rozpoznávačem řeči, bakalářska práce, Brno, FIT VUT v Brne, 2011

Automatická tvorba titulků k filmu s rozpoznávačem řeči

Prohlášení

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Ing. Petra Schwarza Ph.D.

Uviedol som všetky literárne zdroje a publikacie, z ktorých som čerpal.

.....
György Csintalan
18.5.2011

Poděkování

Túto cestou by som chcel poďakovať pánovi Ing. Petrovi Schwarzovi Ph.D. za odbornú pomoc v celom priebehu písania tejto práce, za užitočné rady a za poskytnutie kvalitných informačných zdrojov.

Obsah

1 Úvod

V poslednom čase význam rozpoznávania reči postupne stúpa. Ponúka prirodzené riešenie pre zjednodušenie užívateľských rozhraní, jednoduchý spôsob získania informácií na základe vysloveného textu. Rozpoznávanie a spracovanie reči určite bude mať kľúčový význam v technike budúcnosti.

Automatizované vytvorenie titulkov k filmom je jeden z možných nekonečne veľa oblastí aplikácie rozpoznávania reči. Taký program môže byť veľmi úžitočným pre sluchovo postihnutých, ktorí potrebujú získať väčšinu informácií z okolia iným spôsobom ako ostatní. Pokročilé aplikácie takého typu môžu im poskytnúť pomôcť pri zábave.

1.1 Ciele práce

Okrem hlavného cieľa tejto práce, ktorým je vývoj aplikácie pre automatickú tvorbu titulkov k filmom treba vyzdvihnúť aj význam experimentovania s daným rozpoznávačom v takých reálnych situáciach, keď vstupné podmienky sa menia tak často, ako filmy. V bežných filmoch pomer reči a šumu nie je konštantný, zvuková stopa môže obsahovať okrem reči aj muziku a rôzne zvuky reálneho sveta, ktoré znamenajú veľkú výzvu pre rozpoznávača.

Požiadavky pre výslednú aplikáciu sú nasledujúce:

- vytvorenie aplikácie pre platformu Windows
- formát vstupného videa : AVI
- stačí rátať s dvojkanálovou audio stopou maximálne
- jednoduché grafické užívateľské rozhranie
- výstupný formát titulkov : SRT

1.2 Motivácia

Tento problém bol pre mňa veľmi zaujímavý. Bolo treba riešiť pár otázok. Bolo treba nájsť vhodné nástroje pre extrakciu zvuku vo Windowsu s využitím Windows SDK a ukladať zvukovú stopu do príslušného formátu. Najzaujímavejšou časťou bolo experimentovanie s rozpoznávačom v nových situáciach, keď úlohou bolo skúmanie efektivity pri rozpoznávania zvuku vo filmoch. Získané znalosti a možné spôsoby zlepšenia výstupu boli prediskutované s vedúcim úlohy, s pánom Ing. Petrom Schwarzom Ph.D.

Som hudobníkom, pre mňa hluch je jeden z najdôležitejších zmyslov, preto cítim veľký súcit s hluchovo postihnutými. Z toho dôvodu jednou z motivujúcich činiteľov pri písaní tejto práce bola možnosť pomôcť týmto ľuďom.

2 Rozpoznávanie reči

Reč je prirodzenou formou ľudskej komunikácie. Potrebné znalosti získame už v malom veku. Reč je taká prirodzená, že si ani neuvedomujeme aký je zložitý fenomén. Orgány pre tvorbu ľudského zvuku majú nelineárne vlastnosti. Ich operácie sú kontrolované vedomite pričom rôzne faktory jako napríklad výslovnosť, prízvuk, artikulácia, rýchlosť, výška tónu, hlučnosť atď. majú veľký vplyv. Naša irregulárna rečová charakteristika pri prenášaní môže byť skreslená vďaka šumu a iných zvukov pozadia, elektrickej charakteristiky (v prípade použitia telefónov alebo iných elektrických zariadení) . Vďaka všetkým zdrojom náhodnosti je rozpoznanie reči veľmi zložitá záležitosť.

2.1 BSAPI (Brno Speech Application Interface)

Brno Speech Core poskytuje kolekciu základov pre jednoduché a rýchle prototypy rozpoznávačov reči. Implementuje širokú škálu algoritmov od čítania vstupných rečových súborov alebo čítania vstupu z mikrofónu, spracovania zoznamových súborov cez parametrizáciu, klasifikáciu, dekódovanie, fonémové rozpoznávanie, plynulé rozpoznanie reči, keyword spotting, identifikácia jazyka, k identifikácii rečníka na základe zvukového vzoru. Brno Speech Application Interface (BSAPI) je interface medzi BSCORE a iné software. Tento software vyrába firma Phonexia [1].

Balík BSAPI-1.0.32 obsahuje:

- hlavičkový súbor C/C++ bsapi.h
- súbory atlas.dll a bsapi.dll
- Windows batch file na vytvorenie statickej knihovne vcimplib.bat, spustením tohoto súboru je vytvorený bsapi.lib, čo je potrebné pri programovaní vo Visual C++
- dokumentácia vo forme HTML a v PDF
- demonštračné príklady pre ukážku s rôznymi častiami BSAPI

2.1.1 Prepis hovoreného slova na text (LVCSR)

Pre túto prácu bola využitý modul pre prepis hovoreného slova na text pre český jazyk (LVCSR – Large-Vocabulary Continous Speak Recognition). Používaná verzia je určená pre český jazyk, ale firma Phonexia vyrába tento software aj pre iné jazyky (napríklad: anglickú verziu). Česká verzia disponuje slovníkom, ktorý má viac ako 1 milión slov. Vyžadovaný formát vstupného súboru je jednokanálový (MONO) WAV, vzorkovacia frekvencia musí byť 8 kHz PCM (Pulse Code Modulation) .

Cieľom koncepcie LVCSR je prepisovanie vstupnej reči na skupiny slov. Prepis reči na text funguje využitím moderných technik ako sú neurónové sieťe a diskriminatívne trénovanie.

2.1.2 Licencia

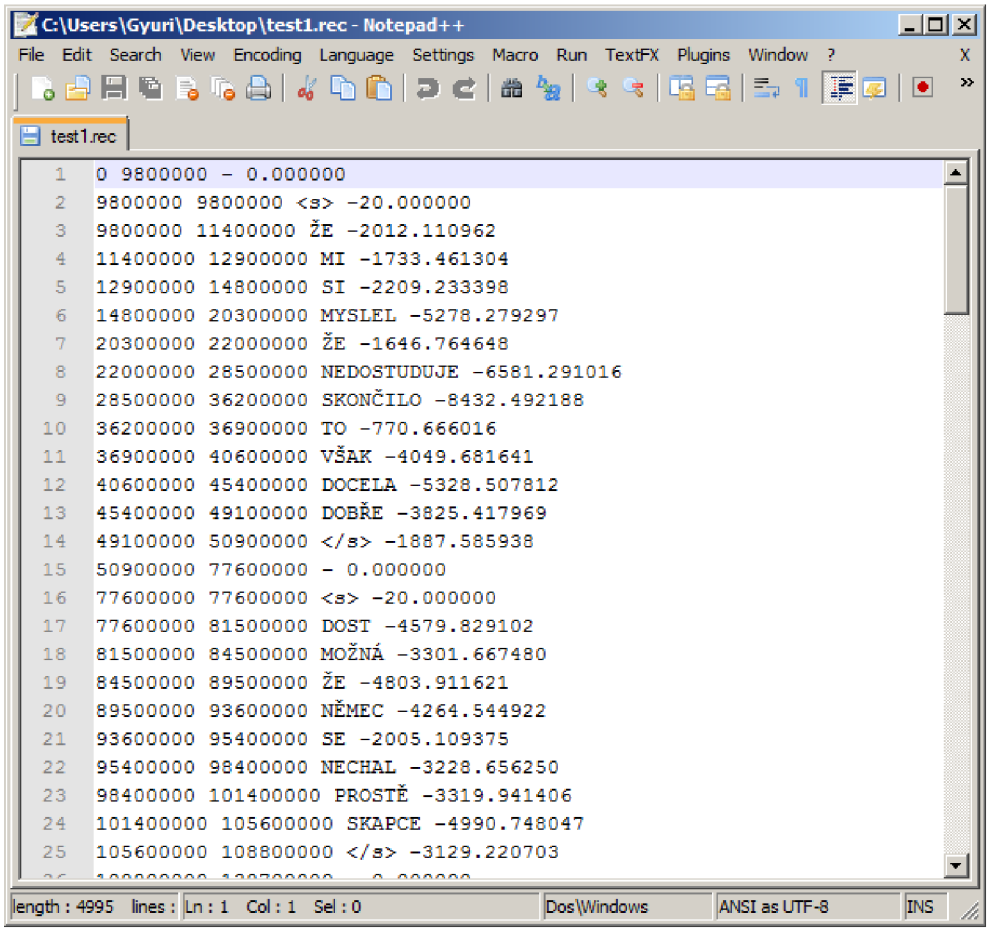
Pre použitie rozpoznávača je potrebná licencia firmy Phonexia. Na overenie validity licencie BSAPI poskytuje licenčný manager. Manager komunikuje s licenčným serverom a skontroluje, či je licencia validná. Licenčný súbor (v tomto prípade `license.dat`) pre tento účel, má byť priložené k aplikácii .

2.1.3 Konfigurácia rozpoznávača

Pri inicializácii rozpoznávača treba zadať pripravený konfiguračný súbor. Súbor obsahuje informácie o segmentácií, o použitej znakovej sade, rôzne inicializačné údaje jednotlivých modulov BSCORE a cesty prípadných iných konfiguračných súborov.

2.1.4 Výstup rozpoznávača

Výstup rozpoznávača obsahuje rozpoznanú reč v textovej podobe s príslušnými časovými údajmi uvedených v 100 ns - ových jednotkách. Spoznané skupiny slov, vo výstupnom súboru sú umiestnené medzi XML tagmi `<s>` a `</s>` .



obr. 1 výstup z rozpoznávača

3 Extrakcia zvukovej stopy

V tejto kapilole bude popísaný spôsob extrakcie zvukovej stopy z videa na platforme Windows. Pred uvedením prostriedkov o spôsobe riešenia je teoretický úvod, kde sú vysvetlené používané pojmy.

3.1 WAV (Waveform Audio File Format)

WAV [3] je jeden z bežných zvukových kontejnerov. Je to štandard na ukladanie bitového prúdu audia na počítači v kúskoch data, v takzvaných „data chunks “. Môže obsahovať komprimované audio, ale vo všeobecnosti obsahuje nekomprimované audio s kódovaním LPCM (Linear Pulse Code Modulation) [4]. Prípona WAV súborov je „.wav“. (Túto príponu môžu dostať aj iné zvukové súbory, napríklad typu NIST Sphere, ktorý je formátom vo väčšine prípadov používaný na ukladanie reči.)

3.2 AVI (Audio Video Interleave)

AVI [5] je formátom kontejneru multimediálnych dát. Môže obsahovať naraz audio aj video dáta v takom kontejnere, ktorý povolí synchronné prehrávanie audia a videa. Podobne ako WAV, aj AVI ukladá dáta v blokoch (data chunks). Prípona AVI súborov je „.avi“.

3.3 COM (Component Object Model)

COM [6] je platformne nezávislý objektovo orientovaný systém pre vývoj binárnych softwareovýh komponentov, ktoré môžu vzájomne pôsobiť. Nespecifikuje štruktúru aplikácie ani implementačné detaily, ale určí objektový model a programátorské požiadavky, ktoré umožnia interakciu medzi COM objekty a iné.

COM definuje všeobecnú podobu COM objektov. Vo všeobecnosti objekt pozostáva zo skupiny dát a funkcií, ktoré tieto dáta manipulujú. Tieto dáta objektov sú prístupné iba cez spomínaných skupín funkcií. Taká skupina funkcií sa nazýva rozhranie (interface) a jednotlivé funkcie rozhrania sú metódy. COM požaduje, aby jediným spôsobom prístupu k metodám rozhrania bolo možné cez ukazateľov na rozhranie. Okrem určenia standardu pre objektov, COM definuje základné rozhrania, ktoré poskytujú potrebné obecné funkcie pre všetky technológie založené na COM, definuje spôsob spolupráce objektov a bezpečnostné vlastnosti pre poskytnutie integrity.

3.4 Microsoft DirectShow

Pretože požadovaná platforma pre aplikáciu je Windows, bolo treba nájsť spôsob a vhodné prostriedky na extrakciu zvuku na tejto platforme. **Windows SDK (Software Developement Kit)** [7] ponúka násjtrojov pre vývoj aplikácií na platforme Windows. Pre vývoj multimediálnych aplikácií je určený Microsoft DirectShow [8]. DirectShow je framework a API na prácu s multimédiou. Je založený na Microsoft Windows COM (Component Object Model). Poskytuje multimediálne rozhranie pre rôzne programovacie jazyky. Základ frameworku tvoria filtre, ktoré môžu byť navrhnuté pre rôzne účele. Je rozšíritelný, čo znamená možnosť pridania vlastných filtrov.

3.5 Programovanie DirectShow aplikácií

V tejto podkapitole budú ukázané základné koncepty tvorby DirectShow aplikácií.

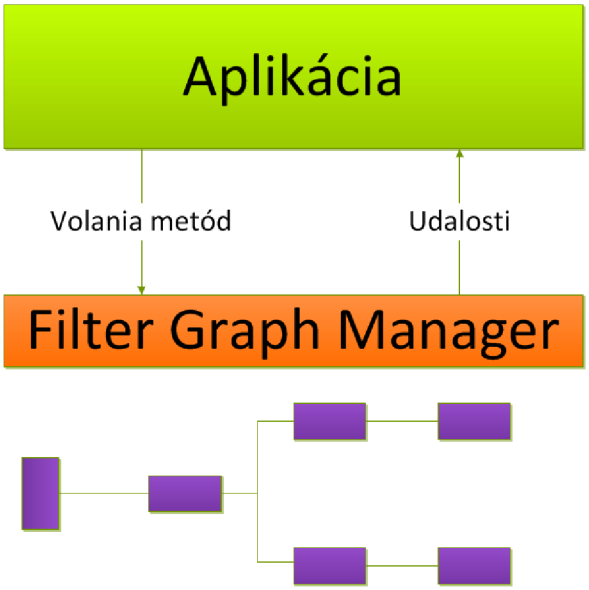
3.5.1 Filtre a filter graph

Základným stavebným blokom v DirectShow je objekt nazývaný **filter**. Filter je softwareovým componentom (COM objekt), ktorý vykoná určíté operácie na multimediálne streamy. Filter získaný vstup spracuje. Výstupom filtra je spracovaný vstupný prúd. Podľa činnosti filtra rozlišujeme rôzne typy filtrov:

- *Source filter* : Poskytuje vstupná data filter grafu z určitého zdroju.
- *Transform filter* : Transformuje vstupný prúd na výstupný.
- *Renderer filter* : Prezentuje jeho vstupná dáta pre užívateľovi (renderovanie).
- *Splitter filter* : Rozdelí vstupný prúd na viac výstupných prúdov, pričom typicky parsuje vstup.
- *Mux filter :* Skladá viac vstupných prúdov do jedného výstupného.

Vzájomným spojením filtrov (výstup jedného filtra sa stane vstupom druhého filtra) vznikne tzv. **filter graf (filter graph)**, určitý na vykonanie úloh filtrov zreťazene. Datový tok v aplikácii spravuje COM objekt tzv. **filter graf manžér (filter graph manager)**. Aplikácia prikazuje iba high-level volania ako *Run*, *Stop*, *Pause*. Filter graf manažér :

- Koordinuje stavové zmeny filtrov. Stavove zmeny filtrov nastanú partikulárne. Stavové zmeny nie sú ovplyvnené priamo aplikáciou, príkazy aplikácie doručí filter graf manažér postupne pre každého filtra.
- Poskytuje *referenčný čas (reference clock)*, čo zaručí synchronizáciu prúdov. Čas prezentácie multimédie sa nazýva *presentačný čas (presentation time)*, ktorý je relativný voči referenčnému času.
- Spravuje udalosti. Používa frontu udalostí pre informovanie aplikácie, keď nastane nejaká udalosť.
- Poskytuje rôzne metódy pre aplikáciu napr. pre pridávanie filtrov do grafu a pre spojenie filtrov.



Obr. 2 Bloková reprezentácia typickej DirectShow aplikácie

3.5.2 Simulačný nástroj GraphEdit

GraphEdit je visuálny nástroj pre vytvorenie a simuláciu činnosti filter grafov. Pomocou tohoto nástroja je možné otestovať filter graf ešte pred implementáciou. Dovoľuje programátorovi vidieť, že aké filtre sú registrované na aktuálnom počítači. GraphEdit je tiež súčasťou Microsoft Windows Software Development Kitu (SDK).

3.6 Extrakcia zvukovej stopy pomocou DirectShow

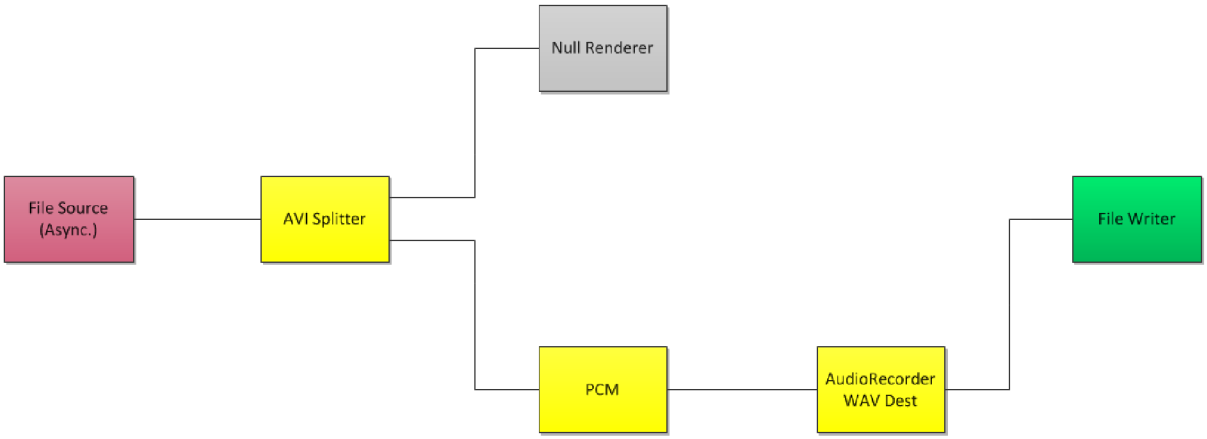
Extrakcia zvukovej stopy sa koná vytvorením a spustením filter grafu z vhodných filtrov. Pretože DirectShow predvolene podporuje aj formáty WAV a AVI, či formáty vstupného a výstupného súboru, takým spôsobom nie je potrebné vytvoriť a registrovať vlastné filtre pre načítanie a ukladanie dát.

Pre tento účel môžu byť nasledujúcie filtre vhodné:

- **File Source (Async.)** : Otvorí a číta lokálne súbory rôzneho typu a predáva data pre iný filter, ktoré parsuje data.
- **AVI Splitter** : Akceptuje iba vstupný formát AVI a rozdelí vstupný prúd na svoje komponenty (audio prúd, video prúd) pre budúce spracovanie alebo renderovanie. Má dve výstupy, prvý výstup slúži pre video kanál, druhý pre audio.
- **PCM** : Patri do skupiny audio kompresorov. Na vstup aplikuje algoritmus pulse code modulation.

- **AudioRecorder WAV Dest** : Vstupná audio data convertuje do WAV. Môže byť spojený napríklad s file writerom.
- **File Writer** : Uloží data na disk. Umožňuje vytvorenie nového súboru alebo prepisovanie existujúceho súboru.
- **Null Renderer** : Je to renderer filter, ktorý odmieta vstupné data, bez prezentovania používateľovi. V tomto prípade je spojený s prvým výstupom avi splittera, lebo video stopu netreba spracovať.

Po vzájomnom pripojení spomenutých filtrov vznikne nasledujúci filter graf pre extrakciu audio stopy z videa :



obr. 3 Filter graf pre extrakciu audio stopy z videa

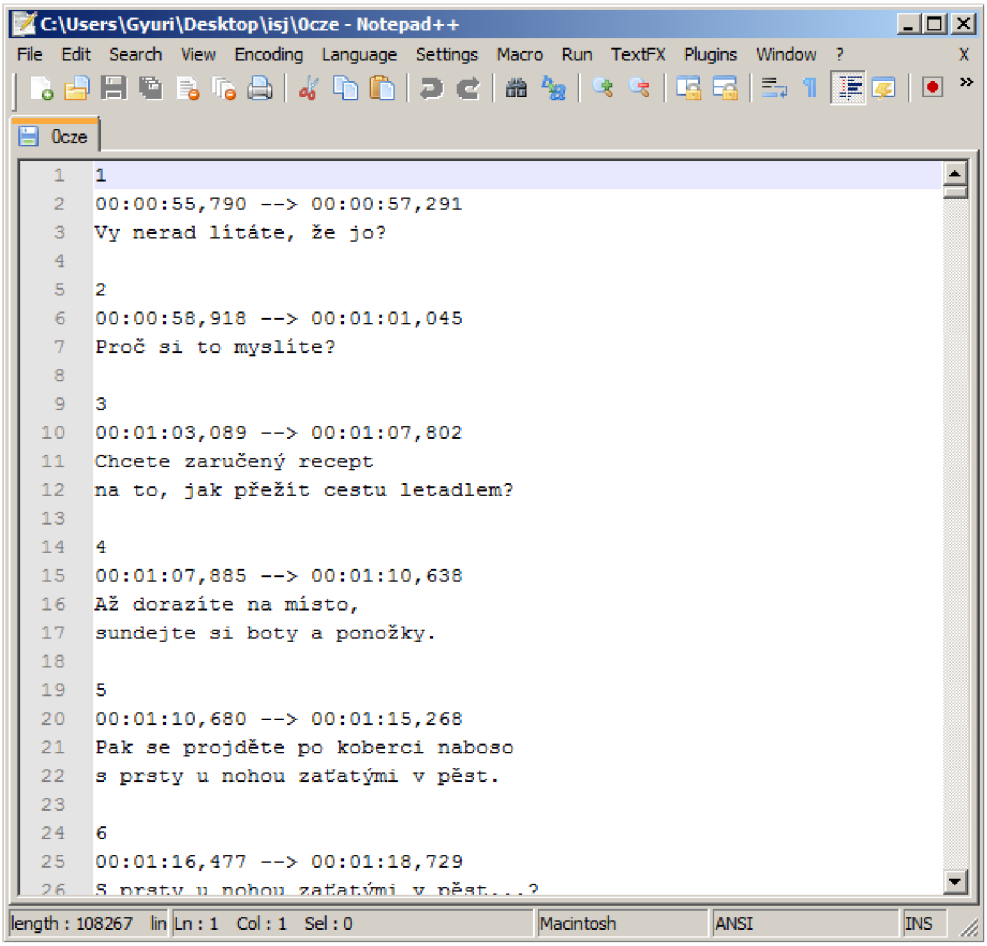
4 Vytvorenie titulkového súboru

V tejto kapitole je popísaná spôsob spracovania výstupu rozpoznávača, vytvorenie titulkového súbou. V prvej podkapitole je predstavený formát titulkového súboru.

4.1 SubRip text file format (SRT)

SubRip text file format [11] je jeden z najrozšírenejších formátov titulkových súborov. Príponou SubRip súborov je „.srt“, obsahujú sformátovaný text. Štruktúra SRT súboru sa skladá z blokov. Jednotlivé bloky obsahujú okrem textu titulkov ešte poradové číslo (počitané od 1) a časové

informácie o začiatku a o konci prezentácie textu na obrazovke. Používaný časový formát: **hours:minutes:seconds,milliseconds.** Používa sa desatinná čiarka ako oddeľovač pri desatinných čísel. Podporuje kódovanie UTF-8.



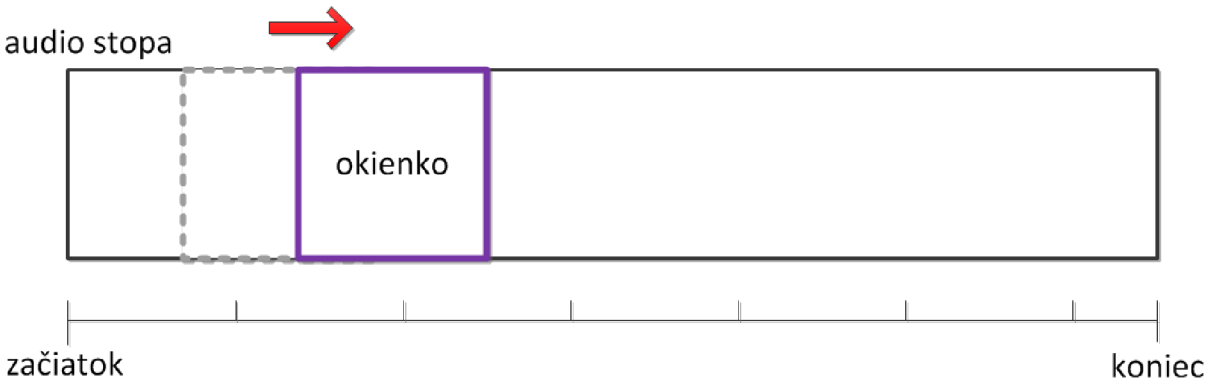
obr. 4 Ukážka titulkového súboru

4.2 Prepis reči na text

Pretože rozpoznávač pracuje v offline režime, pracuje s audio súbormi uložených na disku, je potrebné vytvoriť koncept, algoritmus na segmentáciu vstupného súboru, lebo si nemôžeme dovoliť prácu s veľkými dočasnými audio súbormi (myslíme veľkosť stovky MB až GB). Na druhej strane segmentácia nemôže zhoršiť statistiky správneho rozpoznania reči. To sa môže nastať vtedy, keď napríklad koniec segmentu je počas vyslovenia jedného slova. Pre riešenie problematiky je používaný algoritmus pracujúci s časovým okienkom.

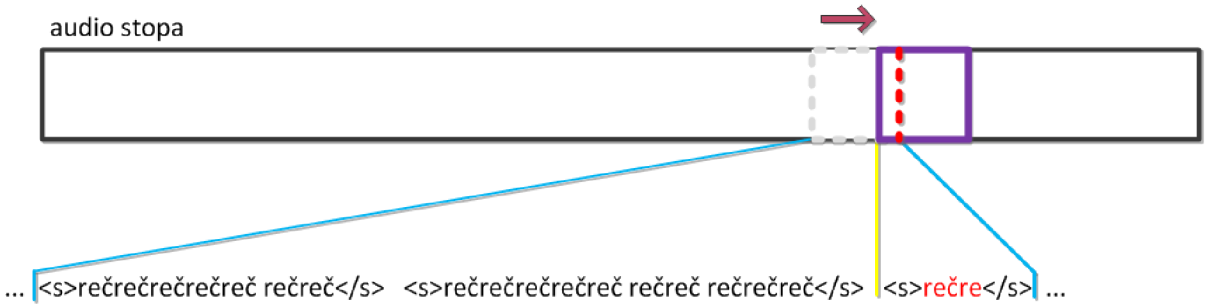
4.3 Segmentácia zvukového súboru

Pre zníženie potreby voľného miesta aplikácie na pevnom disku nie je celá audio stopa spracovaná naraz. Aplikácia pracuje s časovým okienkom dĺžky 20 sekúnd. Po spracovania zvukovej stopy v aktuálnom časovom intervalu však nie je vždy možné posunúť okienko na konci aktuálneho intervalu. Dôvodom toho je možná strata slova vďaka segmentácie. Je možné si to predstaviť tak, ako keby niekto skočil do reči rečníka (pôvodný rečník je prinútený na skončenie svojho hovoru). Dôvodom zvolenia dĺžky časového okienka na jednej strane je akceptovateľná veľkosť dočasného audio súboru (do 1 MB), na druhej strane v tomto prípade s veľkou pravdepodobnosťou bude obsahovať neovplyvnené skupiny slov segmentáciou zvukovej stopy a počet posunutí okienka bude menší ako pri menšom okne. Samozrejme „množstvo“ reči v rôznych typoch filmov aj v rôznych častí istého filmu môže byť odlišné, preto v mnohých situáciách často sa môže okienko posunúť po 20 sekundových intervaloch.



obr. 5 Posunutie okienka cez autio stopu

Spracovaním výstupu rozpoznávača po prepisovaní reči v aktuálnom okienke získame informácie o miesta posunutia.



obr. 6 Posunutie okienka na začiatok poslednej skupiny

Algoritmus spracovania zvukovej stopy s posunutím okienka je nasledujúci (pseudokód) :

```
init, je_koniec = false, nastav začiatok_okienka
while (!je_koniec) {
      if (začiatok_okienka + dĺžka_okienka > koniec_zvukovej_stopy)
      {
            koniec_okienka = koniec_zvukovej_stopy;
            je_koniec = true;
      }
      else  koniec_okienka = začiatok_okienka + dĺžka_okienka;
      posunutie okienka
      prepis reči na text v okienke
      spracovanie_vystupu();
}

spracovanie_vystupu() {
      trasformácia skupín slov okr. poslednej na titulky a zápis do
      titulkového súboru, inkr. flag nezapisovania posl. skupiny
      if (počet_skupín_v_okienke == 0) {
            flag = 0;
            začiatok_okienka = začiatok_okienka + dĺžka okienka;
      } else if (počet_skupín_v_okienke == 1 && flag == 2)  {
            //posl. skupina druhýkrát to isté
            transformácia na titulky a zápis do súboru;
            začiatok_okienka = začiatok_okienka + dĺžka okienka;
            flag = 0;
      } else  začiatok_okienka = koniec_poslednej_skupiny
}
```

O poslednej skupine slov nie je možné rozhodnúť, či bola skupina slov zakončená iba vďaka konci zvukového segmentu (nútený koniec), preto treba posunúť okienko na začiatok poslednej skupiny a znova odtiaľ spracovať audio stopu. V prípade keď ďalšie okienko okrem poslednej skupiny neobsahuje inú skupinu slov (nasleduje časť filmu bez reči), treba poslednú skupinu ihneď zapisovať do titulkového súboru a posunúť okno o celú dĺžku (viď flag nezapisovania vo pseudokódu).

4.4 Zápis titulkov to titulkového súboru

Tvorba a zápis titulkov do titulkového súboru sa koná postupne so spracovaním výstupov rozpoznávača. Každý rečový segment je zapísaný do titulkového súboru. Časové údaje začiatku a konca zobrazenia sú nastavené na základe začiatku a konca skupiny slov.

5 Implementácia

V tejto kapitole je popísaná implementácia aplikácie pre platformu Windows. Sú uvedené použité nástroje a techniky prípadne ukážky zdrojových súborov.

5.1 Vývojárske prostredie

Pre vývoj tejto aplikácie bol používaný Mirosoft Visual Studio 2010 a Microsoft Windows SDK 7.1. Aplikácia je implementovaná v jazyku C++ **(Visual C++)** [13] s využitím **Windows apllication programming interface (API)**.

Práca s DirectShow vyžaduje zahrnutie statických knižníc *Strmiids.lib*, *Quartz.lib* a *Strmbase.lib* do projektu. Strmbase.lib vznikne kompiláciou DirectShow Base Classes. Okrem statických knižníc každá DirectShow aplikácia používa hlavičkový súbor Dshow.h.

Ako to bolo naznačené v kapitole 2.1, balík BSAPI obsahuje batch file pre vytvorenie statickej knižnice bsapi.lib, čo je potrebné pre prácu s BSAPI vo Visual C++. Bsapi.lib treba taktiež zahrnúť do aplikácie.

5.2 Extrakcia zvuku

Ako už to bolo popísané v kapitole 3, pre získanie audio stopy je treba vytvoriť inštanciu filter graf manažéra, pridať príslušné filtre k vytvorenia filter grafu a vhodne ich spojiť.

5.2.1 Filter Graph Manager

Pretože filtre a filter graph manager sú taktiež COM objekty, môžu byť vytvorené pomocou COM funkciou **CoCreateInstance**. Pri vytvorení COM objektu treba špecifikovať mimo iných **CLSID** a **IID**. CLSID je 128 bitové hexadecimálne číslo medzi zložených zátvoriek, je globálnym jednoznačným identifikátorom COM class objektu. Hodnoty CLSID zaregistrovaných COM objektov sa nachádzajú v regisrty. IID je referenčným identifikátorom používaného rozhrania pre komunikáciu s COM objektom. V prípade úspešného vytvorenia COM objektu, získame prístup k objektu cez ukazovateľ na zadané rozhranie. Ukážka vytvorenia filter graf manažéra :

```
IGraphBuilder* pIGB;
HRESULT hr = CoCreateInstance(  CLSID_FilterGraph, NULL,
                                CLSCTX_INPROC_SERVER,
                                IID_IGraphBuilder, (void **)&pIGB );
```

DirectShow poskytuje dve varianty filter graf manažéra, v tomto prípade pri vytvorení je používaný makro **CSLID_FilterGraph** (vytvorí graph manager na „shared worker thread“). Prístup k manažéru bude možný cez rozhranie **IGraphBuilder** (jeho IID – **IID_IGraphBuilder**). IGraphBuilder ponúka okrem iných metódy na spojenie filtrov a pridania filtrov do grafu.

5.2.2 Vytvorenie filtrov – filter graf

Inštantcie filtrov sa väčšinou vytvárajú taktiež pomocou volania **CoCreateInstance**. Pre niektoré filtre však neexistuje preddefinované makro obsahujúce CLSID. V tomto prípade je vhodné používať System Device Enumerator (cez rozhranie **ICreateDevEnum**). System Device Enumerator vymenuje všetky registrované filtre na počítači podľa kategórií. Enumerátor určitého kategórie filtrov môžeme získať pomocou metódy **ICreateDevEnum::CreateClassEnumerator** na základe CLSID kategórie požadovaného filtra. V praxi táto metóda vrátí ukazovateľ na rozhranie **IEnumMoniker**. Volaním metódy **IEnumMoniker::Next** je možné vymenovať tzv. monikery, získame ukazovateľ na rozhranie **IMoniker**. Volaním **IMoniker::BindToObject** vytvoríme a inicializujeme požadovaný filter.

V prípade tejto aplikácie všetky potrebné filtre okrem **AudioRecorder WAV Dest** a **PCM** v znalosti príslušných hodnôt CLSID je možné vytvoriť pomocou CoCreateInstance. CoCreateInstance je volaný v rámci funkcie *AddFilterByCLSID*, ktorý vytvorí filter a pridá ho do filter grafu. Skoršie spomínané dve filtre sa musia vyhľadať a vytvoriť použitím System Device Enumerator až potom môžu byť pridané do filter grafu.

Po pridania File Source filtra nastavíme cestu vstupného videosúboru. Po pridania File Writer filtra treba nastaviť cestu výstupného dočasného audiosúboru súboru.

5.2.3 Spojenie filtrov

Spojiť dvoch filtrov je možné pomocou metódy **IGraphBuilder::Connect**. Táto metóda spojí dve vývody (pin) : výstup jedného filtra so vstupom druhého. Pracuje s ukazovateľmi na rozhranie **IPin**. IPin je základným rozhraním vývodov každého filtra. Ponúka metódy ako napr. metóda **QueryDirection** na urečenie smeru vývodu (vstup/výstup).

V prípade tejto aplikácie je vytvorená funkcia pre spojenie filtrov (**ConnectFilters**), ktorá využíva rôzne pomocné funkcie :

- **IsPinConnected** – Otestuje či vývod je zapojený alebo nie.
- **IsPinDirection** – Otestuje či smer vývodu zhoduje so zadaným smerom alebo nie.
- **MatchPin** – S využitím predchádzajúcich otestuje či stav daného vývodu zhoduje so zadaným stavom alebo nie. (voľnosť a smer vývodu)
- **FindUnconnectedPin** – Vyhľadáva prvý nezapojený spoj na danom filtre.

Po spojení filtrov PCM filtra nastavíme parametry jeho výstupu (8 kHz, MONO). Je to možné vyplnením WAVEFORMATEX štruktúry. Ukážka zdrojového kódu :

```
AM_MEDIA_TYPE* pMediaType;
CComPtr<IAMStreamConfig> pPCMStreamConfig;
pPCMStreamConfig = pPCMOutPin; //výstup PCM filtra, definované inde

pPCMStreamConfig->GetFormat(&pMediaType);
ASSERT(pMediaType->subtype == MEDIASUBTYPE_PCM);
ASSERT(pMediaType->formattype == FORMAT_WaveFormatEx);
WAVEFORMATEX* pFormat = (WAVEFORMATEX *) pMediaType->pbFormat;
pFormat->nChannels = 1;
pFormat->nSamplesPerSec = 8000;
pFormat->nBlockAlign = pFormat-> nChannels *
                                (pFormat>wBitsPerSample/8);
pFormat->nAvgBytesPerSec = pFormat->  nBlockAlign *
                                      pFormat->nSamplesPerSec;
pPCMStreamConfig->SetFormat(pMediaType);

/*poznámka : Ošetrenie neúspechov funkcií/metód kvôli prehľadnosti
vynechané.                                                        */
```

5.2.4 Rozhrania COM objektu

COM objekty môžu byť prístupné cez rôzne rozhrania, ktoré ponúkajú rozličné možnosti. Po vytvorenia objektu môžeme získať prístup k objektu cez iné rozhrania pomocou metódy

QueryInterface, ktorá je súčasťou základného rozhrania **IUnknown**. Všetky rozhrania v COM sú priamou alebo nepriamou cestou potomkami tohoto rozhrania.

Pretože IGraphBuilder ponúka metódy iba pre splnenie niektorých základných funkcií, je potrebné mať prístup k filter graf manažéru aj cez iné rozhrania napríklad pre kontrolu dátového toku v grafu. Príklad použitia QueryInterface :

```
IMediaControl *pMediaControl;
hr = pIGB->QueryInterface( IID_IMediaControl,
                           (void **)&pMediaControl );
```

Použité rozhrania filter graf manažéra v aplikácie:

- **IGraphBuilder** – Ponúka operácie potrebné pre vytvorenie grafu.
- **IMediaControl** – Kontrola dátového toku v grafu. Poskytuje okrem iných funkcie Run (spustí všetky filtre v grafu) a Stop (zastaví všetky filtre v grafu).
- **IMediaSeeking** – Ponúka metódy pre skoky na určité miesta v stope a na nastavenie rýchlosti prehrávania. Metódy ako **GetCurrentPosition**, **GetDuration**, **GetCapabilities, SetPositions** boli používané pri posunutiu okienka (pri implementovanie okienkového algoritmu). Metódy tohoto rozhrania pracujú predvolebne s časovou jednotkou dĺžky 100 ns (REFERENCE_TIME unit).
- **IMediaEventEx** – Tento rozhranie je odvodené od rozhrania **IMediaEvent**, ktoré obsahuje metódy pre náhradu predvolenej spravovania udalostí a pre získanie notifikácií udalostí. IMediaEventEx pridá metody, ktoré umožňujú získanie správ udalostí pre okno aplikácie, keď nastanú. Pomocou metódy **SetNotifyWindow** môžeme zaregistrovať okno na spracovanie notifikácií udalostí. V prípade tejto aplikácie je zaregistrovaná hlavné okno užívateľského rozhrania.

5.2.5 Proces extrakcie zvukovej stopy

Ako to už bolo spomenuté, extrakcia zvukovej stopy sa nekoná naraz. Je to spracované po menších segmentoch. Každý segment je spracovaný rozpoznávačom pred získaním nového segmentu zvuku. Extrakcia zvukového segmentu sa začína po nastavení začiatočného a koncového času filter grafu pomocou metódy IMediaSeeking::SetPositions, potom spustíme filter graf (IMediaControl::Start). Pri dosiahnutia koncového času sa pošle oznámenie o udalosti **EC_COMPLETE**. Spracovanie tejto udalosti zahrnuje zastavenie filter grafu (ImediaControl::Stop), spracovanie zvukového segmentu, nastavenie nový začiatočný a koncový čas filter grafu a spustenie filter grafu znova. Tento proces sa opakuje do konca spracovania celej zvukovej stopy.

Vytvorenie grafu sa koná iba raz pomocou funkcie **buildGraph**, spustiť graf je potrebné pre každý segment zvuku pomocou **runGraph** (za bližšie informácie viď zdrojové kódy a [9]).

5.3 Prepis reči na text

V tejto podkapitole bude popísaná prepis reči na text pomocou BSAPI a vytvorenie titulkového súboru z implementačného hľadiska.

5.3.1 Použitie BSAPI

Triedy BSCORE nie je možné vytvoriť priamo, ale sú prístupné cez rozhrania. Každá trieda má svoje vlastné rozhranie (trieda v BSAPI s „I“ na konci názvu) a vlastné identifikačné číslo v BSAPI. **SUnknownI** je základnou triedou pre všetky rozhrania. Obsahuje funkcie pre získanie identifikačného čísla, pre počítanie referencií a pre nastavenie spracovania udalostí.

Ako už to bolo spomenuté BSAPI poskytuje prístup k mnohým prostriedkom na rôzne účele. Pre prepis reči na text ponúka dve rozhrania **SOnlineSpeechRecognizerI** a **SOfflineSpeechRecognizerI**. Pretože v našom prípade zdroj reči je zvuková stopa filmu uložený na pevnom disku, SOfflineSpeechRecognizerI je použitý pre túto úlohu (SOnlineSpeechRecognizerI požaduje vstup z mikrofónu). Pre použitie rozpoznávača jednoduchým spôsobom treba vytvoriť inštanciu z triedy SofflineSpeechRecognizerI. Pred vytvorením rozpoznávača je treba overiť platnosť licencie. Pre tento účel BSAPI ponúka triedu **SLicenceManagerI**. Ukážka vytvorenia rozpoznávača s overením platnosti licencie :

```
SLicenseManagerI *plicman = BSAPIGetLicenseManager();
if (plicman)
{
     plicman->SetErrorHandler(&gErrorHandler);
     plicman->SetStateHandler(&gStateHandler);
     plicman->RegisterLicenseFile("license.dat");
}
```

Pri úspešnom vytvorenia manageru sa nastaví error hadler, state handler (výpis stavových informácií – verbose mode) a názov licencového súboru (license.dat).

```
SOfflineSpeechRecognizerI *psrec;

     psrec = static_cast <SOfflineSpeechRecognizerI>(
                          BSAPICreateInstance(SIID_OFFLINESREC2) );

     if (!psrec)
     {
          return 1;
     }
```

```
psrec->SetErrorHandler(&gErrorHandler);
plicman->SetStateHandler(&gStateHandler);

if (!psrec->Init(pconfig_file))
{
      psrec->Release();
      return 1;
}
```

V prípade úspešného vytvorenia rozpoznávača sa nastaví error handler a state handler a inicializuje sa rozpoznávač. Parameter metódy **Init** je názov konfiguračného súboru (viz. kapitola 2.1.3).

Po inicializácie rozpoznávača prepis reči na text sa koná volaním metódy SOfflineSpeechRecognizerI::ProcessFille.

Ukážka :

```
pSpeechRec->ProcessFile(pInputFile, pOutputFile);
```

Parametry pInputFile a pOutputFile sú ukazovatele na miesto pamäti, kde je uložené meno vstupného a meno výstupného súboru.

Informácie o chýb funkcií v BSAPI sú ľahko interpretovateľné :

- true znamená, že funkcia sa skončila úspešne
- false znamená koniec s chybou

V prípade, že je nastavený error handler pre rozhranie, detailné informácie o chybe sú posielané pre error handlera, inak sú vytlačené na štandardný chybový vystup. Error handler môže rovnako spravovať aj varovné hlásenia.

5.3.2 Proces prepisovania reči na text

Kontrola validity licencie a inicializácia rozpoznávača samozrejme sa koná iba raz na začiatku činnosti prepisovania. Pri získania každého zvukového segmentu (pri spracovania udalosti grafu EC_COMPLETE) sa spustí rozpoznávač a spracuje obsah dočasného audio súboru. Na konci činnosti treba uvoľniť rozpoznávač.

5.4 Užívateľské rozhranie

Kvôli jednoduchému ovládateľnosti aplikácie je vytvorené grafické užívateľské rozhranie s využitím Windows API [13]. Programovania užívateľských rozhraní je dosť rozsiahlou tematikou. Cieľom tejto kapitoly nie je návod programovania užívateľských rozhraní pomocou Windows API, ale zvýraznenie niektorých riešení pre vytvorenia užívateľského rozhrania pre túto aplikáciu.

5.4.1 Popis užívateľského rozhrania

Ovládanie tejto aplikácie nevyžaduje veľa užívateľských vstupov, vďaka čomu je užívateľské rozhranie je dosť jednoduchá.

Pre vykonanie činnosti aplikácia vyžaduje iba určenie správnej cesty vstupného video súboru vhodného typu a cesty výstupného súboru titulkového súboru používateľom. Pre tieto účele sú umiestnené na hlavnom okne užívateľského rozhrania dve textové polia na zadanie týchto ciest. Kvôli zjednodušeniu výberu vstupného a výstupného súboru je možné ich vybrať pomocou **File Save Dialog** a **File Open Dialog**. Tie sú dialógové okná (komponenty COM) pre vybranie vstupného a výstupného súboru pomocou Prieskumníka Windows (Windows Explorer).

Kvôli názornosti stavu činnosti aplikácie je implementovaný **Progress Bar**, ktorý znázorňuje mieru časti spracovanej audio stopy k dĺžky celej audio stope. Nastavenie rozsahu Progress Bara je možné pomocou posielaním správy **PBM_SETRANGE**, obsahom správy je dĺžka celej audio stopy v sekundách. Nastavenie aktuálnej pozície je možné posielaním správy **PBM_SETPOS** a obsahom správy je súčasná pozícia okienka v sekundách. Okrem Progress Bara je umiestnený ešte kontrol **Static Text** zobrazujúci mieru spracovanosti v percentách podľa vzorca:

$$PočetPercent = \frac{SúčasnaPozíciaOkienka}{DĺžkaAudioStopy} * 100$$

Applikácia umožní zobrazenie hlásenia state handlera a error handlera rozpoznávača počas behu aplikácie pomocou konzoly.

5.4.2 Thready

Pretože pri single-threaded aplikácií užívateľské rozhranie a výpočty zdielajú ten istý thread, tak aj menej náročné výpočty môžu zblokovať užívateľské rozhranie počas priebehu výpočtov. Prepis reči na text je navyše dosť náročnou operácou. Z tohoto dôvodu operácie sa vykonajú v odlišných threadoch od threadu užívateľského rozhrania. V našom prípade pre prepis každého segmentu sú vytvorené nové thready (thread predchádzajúceho prepisu sa vždy končí pred vytvorením nového).

6 Výsledky a hodnotenie

V tejto kapitole je popísané testovanie aplikácie na rôznych častiach rôznych filmov, sú uvedené výsledky vykonaných testov a ich hodnotenie. Testy sú zamerené na rôzne situácie vzhľadom na pomer reči a ostatných zvukov v pozadí (hudba, reč iných rečníkov, šum atď.).

6.1 Autorské práva

Aplikácia bola testovaná na vybraných častiach dvoch animačných filmov, dvoch hraných filmov a jednej univerzitnej prednášky. Používané filmy sú chránené autorskými právami a bez súhlasu autora nemôžu byť šírené, z toho dôvodu bohužiaľ k tejto práci nemôžu byť priložené testovacie materiály. Avšak súhlas autora nie je potrebný v prípade skopírovania jedného exemplára pre vlastné účeli. (ak originál nemá ochranu proti kopírovaniu; pre viac informácií viď [14]).

6.2 Testy

Pri popisu testov je uvedený názov filmu, začiatok vybranej časti vo formáte „hodiny:minúty:sekundy,stotiny sekúnd“ a dĺžka vybranej časti. Výstupné SRT súbory sú prístupné v prílohe. Ku každému výstupnému súboru je prikladaný textový súbor obsahujúci skutočnú reč v textovej forme.

Výsledky a získané znalosti sú ťažko vyjadriteľné iba samými štatistickými údajmi, preto pre každý test je uvedený slovný popis rozdelený na rôzne intervaly celkovej doby filmu na základe zmien podmienok (napríklad zmena scény, objaví hudba v pozadí atď.). Pri každom teste je uvedený charakter zvuku postáv a štatistické údaje rozpoznania v tabuľke rozpísané pre každý charakter zvlášť. Tieto tabulky obsahujú celkový počet slov v reči, počet správne spoznaných slov rozpoznávačom, počet nespoznaných slov rozpoznávačom a počet slov, ktorých rozpoznávač spoznal vo foneticky podobnej forme (výslovnosť spoznaného slova, či slovného spojenia je podobné výslovnosti pôvodného slova).

6.1.1 Test 1

Názov filmu : Simpsonovi – Taková nenormální rodinka [15]
Pôvodný názov filmu : The Simpsons – There's No Disgrace Like Home
Začiatok vybraného časti filmu : 0:2:19,75
Dĺžka vybranej časti : 0:0:59,89

Analýza časti :

a) od 0 s do cca. 6 s

Hudba v pozadí sa postupne silňuje, na konci intervalu sa skoro úplne stíši. Reč je ľahko zrozumiteľná, ale rozpoznávač nespozná slová pri silnejšej hudbe.

Charakter reči postáv :

- *postava1* : muž; afektovaná reč, líši sa od každodennej reči

Štatistické výsledky sú uvedené v **tabuľke A. 1** umiestnenej v prílohe.

b) od 6 s do caa. 13 s

Žiadna hudba ani reverb, reč je zrozumiteľný. Reč postáv je afektovaný. Rozpoznávač často spozná odlišné, ale podobne znejúce slová, slovné spojenia (foneticky podobné).

Charakter reči postáv :

- *postava1* : muž; afektovaná reč, líši sa od každodennej reči
- *postava2* : muž, imitovaný tón reči chlapca; afektovaná reč, líši sa od každodennej reči
- *postava3* : žena, imitovaný tón reči dievčaťa; extrémny prípad

Štatistické výsledky sú uvedené v **tabuľke A. 2** umiestnenej v prílohe.

c) od 6 s do konca

Veľmi tichá, skoro zanedbateľná jemná hudba v pozadí (sláčikový orchester), slabý reverb. Reč každej postavy je zrozumiteľná, ale afektovaná, výrazne dvoch postáv. Tie sú extrémne prípady, jeden z nich má naviac má veľmi vysoký tón zvuku. Reč obidvoch je zrozumiteľná pre človeka, ale rozpoznávač nerozpozná ich reč. Viackrát sa nastane zaujímavá situácia, keď do reči rečníka skočí druhý rečník a rozprávajú naraz cez nejakú časovú dobu (vo väčšine prípadov iba polovice slov sú hovorené naraz). V takých situáciach rozpoznávač buď vôbec nespozná slová, alebo spozná celkom iné slová.

Charakter reči postáv :

- *postava1* : muž; afektovaná reč, líši sa od každodennej reči
- *postava2* : muž, afektovaná reč, líši sa od každodennej reči
- *postava3* : muž; imitovaný tón reči chlapca; afektovaná reč, líši sa od každodennej reči
- *postava4* : muž; afektovaná reč, líši sa od každodennej reči
- *postava5* : muž, imitovaný tón reči chlapca; afektovaná reč, líši sa od každodennej reči
- *postava6* : muž; afektovaná reč, líši sa od každodennej reči

Štatistické výsledky sú uvedené v **tabuľke A. 3** umiestnenej v prílohe.

6.1.2 Test 2

Názov filmu : Simpsonovi – Taková nenormální rodinka
Pôvodný názov filmu : The Simpsons – There's No Disgrace Like Home
Začiatok vybranej časti filmu : 0:13:28,54
Dĺžka vybranej časti : 0:0:59,89

Analýza časti :

od 0 s do konca

Žiadna hudba ani reverb. Reč každej postavy je ľahko zrozumiteľná, ale afektovaná, líši sa od každodennej reči. Podobne ako pri predchádzajúcom testu reč extrémne afektujúcej postavy s výsokým tónom zvuku rozpoznávač vo väčšine prípadov vôbec nespozná, alebo spozná iné slová.

Charakter reči postáv :

- *postava1* : muž; afektovaná reč, líši sa od každodennej reči
- *postava2* : muž, imitovaný tón ženskej reči; afektovaná reč, líši sa od každodennej reči
- *postava3* : muž, imitovaný tón reči chlapca; afektovaná reč, líši sa od každodennej reči
- *postava4* : žena, imitovaný tón reči dievčaťa; extrémny prípad

Štatistické výsledky sú uvedené v **tabuľke A. 4** umiestnenej v prílohe.

6.1.3 Test 3

Názov filmu : Simpsonovi – Taková nenormální rodinka
Pôvodný názov filmu : The Simpsons – There's No Disgrace Like Home
Začiatok vybraného časti filmu : 17:58,05
Dĺžka vybranej časti : 0:0:59,89

Analýza úryvka :

a) od 0 s do cca. 21 s

Žiadna hudba slabý reverb. Reč každej postavy je ľahko zrozumiteľná, ale afektovaná, líši sa od každodennej reči. Podobne ako pri predchádzajúcom testu reč extrémne afektujúcej postavy s výsokým tónom zvuku rozpoznávač vo väčšine prípadov vôbec nespozná, alebo spozná iné slová.

Charakter reči postáv :

- *postava1* : muž; afektovaná reč, líši sa od každodennej reči
- *postava2* : muž; afektovaná reč, líši sa od každodennej reči
- *postava3* : muž, imitovaný tón ženskej reči; afektovaná reč, líši sa od každodennej reči
- *postava4* : muž, imitovaný tón reči chlapca; afektovaná reč, líši sa od každodennej reči
- *postava5* : žena , imitovaný tón reči dievčaťa; extrémny prípad

Štatistické výsledky sú uvedené v **tabuľke A. 5** umiestnenej v prílohe.

b) od 21 s do cca. 41 s

Hlučné zvýsknutie, nárek a iné zvuky v pozadí. Reč rečníka ešte zrozumiteľný, v niektorých situácií trošku ťažšie. Reč je afektovaná, líši sa od každodennej reči. Rozpoznávač nespozná žiadné slovo.

Charakter reči postáv :

- *postava1* : muž; afektovaná reč, líši sa od každodennej reči

Štatistické výsledky sú uvedené v **tabuľke A. 6** umiestnenej v prílohe.

c) od 41 s do cca. 51 s

Žiadna hudba slabý reverb. Reč rečníkov je ľahko zrozumiteľný. Reč je afektovaná, líši sa od každodennej reči. Podobne ako v predchádzajúcich prípadoch často sa to stane, že rozpoznávač spozná odlišné ale podobne znejúce slová (foneticky podobné výrazy).

Charakter reči postáv :

- *postava1* : muž; afektovaná reč, líši sa od každodennej reči
- *postava2* : muž; afektovaná reč, líši sa od každodennej reči

Štatistické výsledky sú uvedené v **tabuľke A. 7** umiestnenej v prílohe.

d) od 51 s do konca

Hlučné zvýsknutie, nárek a iné zvuky v pozadí. Reč rečníka ešte bez problémov zrozumiteľný, v niektorých situácií trošku ťažšie. Reč je afektovaný, líši sa od každodennej reči. Rozpoznávač nespozná žiadné slovo.

Charakter reči postáv :

- *postava1* : muž; afektovaná reč, líši sa od každodennej reči
- *postava2* : žena; afektovaná reč, líši sa od každodennej reči

Štatistické výsledky sú uvedené v **tabuľke A. 8** umiestnenej v prílohe.

6.1.4 Test 4

Názov filmu : Hele kámo, kdo tu vaří? [16]
Pôvodný názov filmu : Waiting...
Začiatok vybranej časti filmu : 12:59,90
Dĺžka vybranej časti : 0:0:59,99

Analýza časti :

a) od 0 s do cca. 31 s

Žiadna hudba slabý reverb, veľmi slabý zvuk v pozadí (hostinec). Reč každej postavy je ľahko zrozumiteľná, ale afektovaná, trošku teatrálna, líši sa od každodennej reči. V mnohých prípadoch rozpoznávač spozná namiesto správnych podobne znejúce slová. Zaujímavé je, že keď rečník na konci znesie prízvuk viac ako v bežnej reči, rozpoznávač buď nespozná celé slovo alebo spozná iba časť tohoto to slova.

Charakter reči postáv :

- *postava1* : muž; afektovaná reč, líši sa od každodennej reči
- *postava2* : muž; afektovaná reč, líši sa od každodennej reči

Štatistické výsledky sú uvedené v **tabuľke A. 9** umiestnenej v prílohe.

b) od 31 s do konca

Žiadna hudba slabý reverb, mierny šum v pozadí (kuchyňa). Reč každej postavy je ľahko zrozumiteľná, ale afektovaná, trošku teatrálna, líši sa od každodennej reči. Podobne ako v predchádzajúcom prípade rozpoznávač v mnohých situáciach spozná namiesto správnych podobne znejúce slová a v prípade, že rečník na konci slova znesie prízvuk viac ako v bežnej reči, rozpoznávač buď nespozná celé slovo alebo spozná iba časť tohoto to slova.

Charakter reči postáv :

- *postava1* : muž; afektovaná reč, líši sa od každodennej reči
- *postava2* : muž; afektovaná reč, líši sa od každodennej reči

Štatistické výsledky sú uvedené v **tabuľke A. 10** umiestnenej v prílohe.

6.1.5 Test 5

Názov filmu : Hele kámo, kdo tu vaří?
Pôvodný názov filmu : Waiting...
Začiatok vybranej časti filmu : 34:59,74
Dĺžka vybranej časti : 0:0:59,99

Analýza časti :

a) od 0 s do cca. 41 s

Hlučná muzika v pozadí, do konca časti sa zoslabuje. Reč každej postavy je zrozumiteľná, ale afektovaná, trošku teatrálna, líši sa od každodennej reči. Rozpoznávač skoro žiadne slovo nespozná do dosiahnutia do určitej úrovni hlučnosti hudby. Po dosiahnutí tejto úrovni rozpoznávač postupne začne spoznať slová. Spočiatku však spozná ešte nesprávne slová.

Charakter reči postáv :

- *postava1* : muž; afektovaná reč, líši sa od každodennej reči
- *postava2* : muž; afektovaná reč, líši sa od každodennej reči
- *postava3* : muž; afektovaná reč, líši sa od každodennej reči

Štatistické výsledky sú uvedené v **tabuľke A. 11** umiestnenej v prílohe.

b) od 42 s do konca

Žiadne zvuky v pozadí, slabý reverb. Reč každej postavy je ľahko zrozumiteľná, ale afektovaná, trošku teatrálna, líši sa od každodennej reči. Niekoľkokrát sa nastane že dvaja rečníci hovoria naraz. V takých prípadoch rozpoznávač nespozná slová. V ostatných prípadoch sú spoznané pôvodné alebo podobne znejúce slová. V prípade, že rečník na konci slova znesie prízvuk viac ako v bežnej reči, rozpoznávač buď nespozná celé slovo alebo spozná iba časť tohoto to slova.

Charakter reči postáv :

- *postava1* : žena; afektovaná reč, líši sa od každodennej reči
- *postava2* : muž; afektovaná reč, líši sa od každodennej reči
- *postava3* : muž; afektovaná reč, líši sa od každodennej reči
- *postava4* : muž; afektovaná reč, líši sa od každodennej reči
- *postava5* : muž; afektovaná reč, líši sa od každodennej reči

Štatistické výsledky sú uvedené v **tabuľke A. 12** umiestnenej v prílohe.

6.1.6 Test 6

Názov filmu : Hele kámo, kdo tu vaří?
Pôvodný názov filmu : Waiting...
Začiatok vybranej časti filmu : 52:59,60
Dĺžka vybranej časti : 0:0:59,99

Analýza časti :

a) od 0 s do cca. 25 s

Mierny šum v pozadí (hostinec). Reč každej postavy je dobre zrozumiteľná, ale silne afektovaná, teatrálna, líši sa od každodennej reči. Väčšina spoznaných slov je odlišné od pôvodných, len sa podobajú na niektoré časti pôvodnych slov. Relatívne veľké množstvo slov rozpoznávač vôbec nespozná. Na začiatku tejto časti filmu je možné počuť časť vety z pozadia, je oviac tichší ako bežná reč vo filmu, jej intenzita sa zbližuje k šumu v pozadia, ale je zrozumiteľná. Tieto slová nie sú spoznané vôbec. V prípade, že rečník na konci slova znesie prízvuk viac ako v bežnej reči, rozpoznávač buď nespozná celé slovo alebo spozná iba časť tohoto to slova.

Charakter reči postáv :

- *postava1* : žena; silne afektovaná reč, líši sa od každodennej reči
- *postava2* : žena; silne afektovaná reč, líši sa od každodennej reči

Štatistické výsledky sú uvedené v **tabuľke A. 13** umiestnenej v prílohe.

b) od 54 s do konca

Mierny šum v pozadí (kuchyňa), podobná situácia ako v predchádzajúcom prípade. Väčšina spoznaných slov je odlišné od pôvodných, podobajú sa niektoré časti pôvodnych slov. Relatívne veľké množstvo slov rozpoznávač vôbec nespozná.
Charakter reči postáv :

- *postava1* : žena; afektovaná reč, líši sa od každodennej reči
- *postava2* : žena; afektovaná reč, líši sa od každodennej reči

Štatistické výsledky sú uvedené v **tabuľke A. 14** umiestnenej v prílohe.

c) od 54 s do konca

Mierny šum v pozadí (kuchyňa), scéna tá istá ako v predchádzajúcom prípade. Popudlivá, zúrivá reč, emocionálna, extrémny prípad. Žiadne slovo nie je spoznané.

Charakter reči postáv :

- postava1 : žena; afektovaná reč, líši sa od každodennej reči

Štatistické výsledky sú uvedené v **tabuľke A. 15** umiestnenej v prílohe.

6.1.7 Test 7

Názov filmu : prednáška VUT FIT, predmet ISJ_2011-04-14 [17]
Začiatok vybranej časti filmu : 10:09,68
Dĺžka vybranej časti : 0:0:59,97

Analýza časti :

a) od 0 s do konca

Veľmi tichý šum v pozadí. Reč nie je afektovaná. Väčšina slov je spoznané správne. Ostatné spoznané slová sa foneticky podobajú na pôvodné.

Charakter reči postáv :

- postava1 : muž, neafektovaná reč

Štatistické výsledky sú uvedené v **tabuľke A. 16** umiestnenej v prílohe.

b) od 0 s do konca

Veľmi tichý šum v pozadí. Reč nie je afektovaný. Neartikulované slová spoznané zle, tie sú ťažko zrozumiteľné aj pre človeka. Ostatné spoznané slová sa foneticky podobajú na pôvodné.

Charakter reči postáv :

- *postava1* : muž, neafektovaná reč

Štatistické výsledky sú uvedené v **tabuľke A. 17** umiestnenej prílohe.

Poznámka : Okrem týchto testov aplikácia bola ešte testovaná na ďalšie dve filmy (kreslený film *Futurama* [19] a hraný film *Mesto – The Town* [18]). Výsledky týchto testov ukazujú podobné výsledky ako hore uvedené výsledky, preto ich dokumentovanie sa zdá zbytočnému v tejto správe (výstupné tiltulkové súbory a pôvodná reč v textovej forme sa nachádzajú na priloženom CD).

6.3 Získané znalosti

Výsledky testov ukazujú na nasledujúce nedostatky :

- afektovaná reč sa spozná s nižšou efektivitou : namiesto pôvodných slov sú spoznané podobne znejúce slová/slovné spojenia (napríklad : „tloustá“ - „tu stát“ atď.); v prípade, že rečník na konci slova znesie prízvuk viac ako v bežnej reči, nie je slovo spoznané poriadne
- prítomnosť muziky kazí efektivitu rozpoznania, prípadne to znemožňuje (už aj hudba s miernou intenzitou)
- v prípade, že viac rečníkov hovorí naraz, ich reč sa nerozpozná
- reč s extrémnym tónom hlasu nebude nerozpoznaná

7 Techniky zlepšenia rozpoznania reči

Výsledky testov aplikácie ukazujú, že v situáciách, keď zvuková stopa okrem reči obsahuje aj iné zvuky napríklad hudbu, reč iných rečníkov či šum v pozadí, efektivita rozpoznania môže dramaticky znížiť. Zlepšenie kvality rozpoznania v takých situácií, či takzvaná **robustné rozpoznanie reči (robust speech recognition)**, je zložitou záležitosťou. Táto kapitola sa zaoberá s vplyvom muziky, šumu a iných zvukov v pozadí a s technikami zlepšenia výsledkov. Dôkladný popis a experimentovanie s týmito metódami presahuje rámec tejto práce. Cieľom tejto kapitoly je naznačenie možností na základe rôznych technických správ. Iba Wienerov filter a experimentovanie s ním budú popísané podrobnejšie v samotnej podkapitole.

7.1 Efekty šumu na rozpoznanie reči

Efekty šumu na rozpoznávanie reči znázorňujú nasledujúce tri experimenty s izolovaným rozpoznávačom slov [20]:

(1) Trénovanie rozpoznávača na „čistej“ reči a testovanie na rušnej reči s rôznymi hodnotami SNR : Výsledky ukážu, že presnosť rozpoznania sa zníži so znižovaním SNR.

(2) Trénovanie a testovnie na reči so zhodnými vlastnosťami cez odlišné hodnoty SNR : Klesanie presnosti je o mnoho uhladenejšie ako pri teste 1.

(3) Testovanie na reči s SNR=18dB a trénovanie cez rôzne SNR hodnoty : Klesanie presnosti rozpoznania je úmerné k diferencií úrovní SNR medzi testovacom a trénovacom reči.

Na základe výsledkov týchto testov môžeme zhrnúť dva dopady šumu na rozpoznanie reči:

- *Nezhoda štatistiky trénovacej a testovacej reči* - Táto vlastnosť spôsobí zníženie efektivity rozpoznávača trénovaného na nerušenej reči. Väčšie diferencie hodnôt SNR trénovacej a testovacej reči spôsobujú väčšie zhoršenie výkonu rozpoznávača.
- *Redukovanie vzdialenosti medzi jednotkami reči (speech units)* - Zhoršuje schopnosť rozlišovania akustického modelu. Mnohé dôležité rozlišovacie informácie reči, hlavne s nízkou energiou, môžu byť skreslené alebo môžeme ich stratiť.

Napriek tomu, že po prispôsobení trénovacích a testovacích dát sa ukážu dobré výsledky, implementovanie takého systému je veľmi náročné, lebo vo skutočnosti nepoznáme presne tie trénovacie dáta, ktoré sa odpovedajú testovacím dátam reálneho sveta. Okrem toho, že testovacie podmienky a vlastnosti testovacích dát nie sú predvídateľné, najčastejšie je iba dostupná databáza rečových dát bez šumu. Cieľom robustného rozpoznávania reči je zdokonalenie, zvýšenie efektivity systémov trénovaných pomocou takej databázy.

7.2 Techniky zlepšenia

Súčasne najmodernejšie techniky rozpoznávania reči môžu dosiahnuť dobré výsledky pri testovaní v „čistých“ akustických prostredí. V situáciach, keď je reč rušený šumom, vďaka napríklad nepriaznivého akustického prostredia, výsledky sa výrazne zhoršujú. Dôvodom toho je najčastejšie odlišnosť medzi testovacích a trénovacích podmienok. Akustické interferencie, ako rôzné typy šumov pozadia, zmenia štatistiku reči. Problém môže byť približovaný z dvoch smerov :

Kompenzácia šumu (compensation)

Prispôsobenie aktustického modelu k šumu (model adaptation)

Pre kompenzáciu šumu existuje množstvo metód napríklad : špektrálna substrakcia, Wienerov filter, feature normalization a ďalšie. Tieto techniky sa snažia redukovať nepriaznivé efekty odlišného akustického prostredia odhadnutím vlastností reči bez šumu.

Modely adaptation techniky zmenia distribúciu čistej reči vzhľadom na efektu aditívneho šumu. Patria tu metódy : parallel model combination, maximal likehood linear regression, maximum aposteriori atď. Pretože tieto techniky sa iba pokúsia prispôsobiť trénovacie a testovacie štatistiky, ich výkon nikdy nebude presahovať výkon dosiahnuteľný v prípade zhodných trénovacích a dát.

7.2.1 Missing feature theory (MFT)

Medzi najmodernejšie techniky patria techniky založené na **missing feature theory (MFT)** [20], ktoré vychádzajú z charakteristík činnosti sluchového systému človeka pri rozpoznaní reči na základe spoľahlivých vlastností reči. Techniky založené na MTF kompenzujú rušené spektrálne vektory v dvoch krokoch : identifikujú, že v tejto reprezentácie ktoré vlastnosti reči chýbajú, v druhom kroku sa prebieha rekonštrukcia chýbajúcich vlastností. Pretože tieto techniky žiadne predpoklady vzhľadom na šum, sú schopné zvládnuť aj šum s nestacionárnymi vlastnosťami.

7.2.3 Vector autoaggressive model (VAR)

Možnosti VAR ešte nie sú využité. Okrem iných oblastí použitia sa dodnes experimentuje s použitím na získanie medzi-rámcových (inter-frame) štatistík na rekonštrukciu vlastností reči s dobrou efektivitou.

7.2.4 Databázy AURORA-2, -3, -4, -5

Tieto databázy obsahujú rečové dáta so šumom pre vývoj robustných rozpoznávacích systémov. Databázy AURORA :

- AURORA-2 obsahuje rečové dáta s pridaným šumom umelým spôsobom
- AURORA-3 obsahuje rečové dáta zaznamenané vo vnútri automobilu
- AURORA-4 podobne ako AURORA-2 obsahuje dáta pridaným šumom, ale aj s rôznymi vzorkovacími frekvenciami
- AURORA-5 obsahuje dáta, ktoré simulujú rušený rečový vstup pre hands-free v izbe.

Podľa technickej správe [26] použitie AURORA-2 databázy slabo znížia efektivitu rozpoznania reči bez šumu, ale zlepšili výsledky v situáciach pri prítomnosti šumu. Celkové výsledky sa zlepšili o 53,01%.

7.3 Wienerov filter

Cieľom Wienerovho filta je redukovanie množstva šumu v signálu pomocou porovnania k odhadnutého požadovaného bezšumového signálu. Wienerov filter nie je adaptivný, predpokladá stacionárne vstupy.

Wienerov filter je navrhnutý filtrovania šumu z rušeného signálu. Je založený na štatistické princípy. Typické filtre sú navrhnuté pre požadovanú frekvenčnú odozvu, naproti tomu Wienerov filter funguje inak. Predpokladá znalosť spektrálnych vlastností pôvodného signálu a šumu, hľadá filter LTI (lineárna časová invarianta), ktorého výstup je čo najbližšie k pôvodného signálu. Wienerov filter je charakterizovaný nasledujúcimi vlastnosti :

- predpoklady : signál a aditívny šum sú stacionárne lineárne stochastické procesy so známymi spektrálnymi charakteristikami alebo so známou auto-koreláciou a krížovou koreláciou.
- Požiadavky : filter musí byť fyzicky rozšíriteľný
- kritéria výkonu : minimálna stredná štvorcová chyba (minimum mean square error - MMSE)

Šumený signál si môžeme predstaviť nasledujúcim spôsobom :

$$(1) \qquad y[n] = x[n] + n[n]$$

Kde y je šumený signál, x je signál bez šumu a n je šum. Pomocou filtra h chceme získať signál, ktorý je veľmi podobné signálu x , čiže $\tilde{x}$, to znamená, že filter h minimalizuje diferenciu medzi x a $\tilde{x}$. Prvým krokom je minimalizácia strednej štvorcovej odchýlky :

$$(2) \qquad \| x - \tilde{x} \|_2$$

To vieme, že $\tilde{x}$ sa rovná $h * y$ (h konvoluovaný y) , takže :

$$(3) \qquad \| x - h * y \|_2$$

Rozširujeme výraz pomocou známymi algebraickými pravidlami. Pomocou Fourierovej transformácie môžeme nájsť spektrálnu hustotu výkonov.

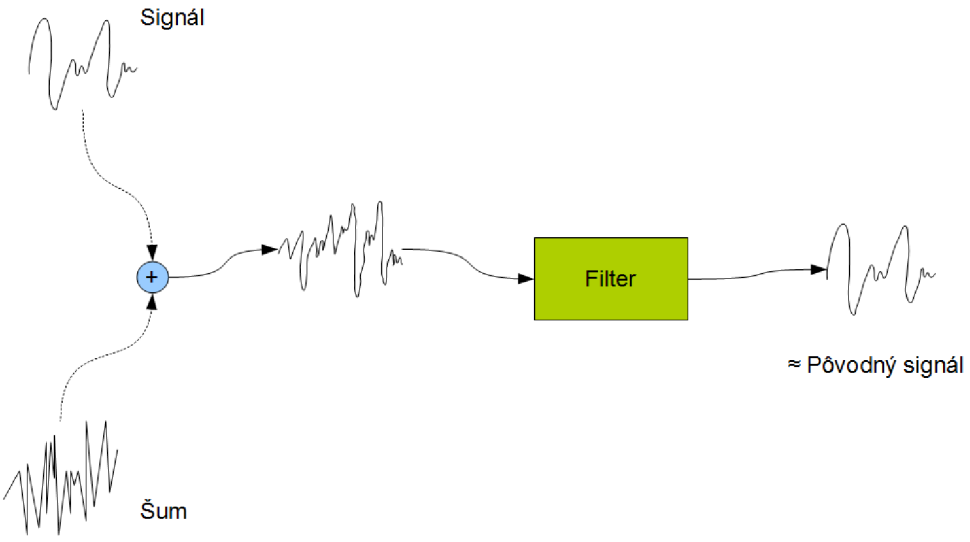
$$(4) \qquad \sum_j (X_j - H_j Y_j)^2$$

$$(5) \qquad \sum_j (X_j - H_j(X_j + N_j))^2$$

Zjednodušením výrazu dostaneme formulu pre *H*, požadovaného filtra optimalizovaného na minimalizáciu diferencie medzi *x* a $\tilde{x}$:

$$(6) \qquad H_f = \frac{(|X(f)|)^2}{(|X(f)|)^2 + (|N(f)|)^2}$$

Kde *X(f)* je výkon signálu a *N(f)* je výkon šumu.
Ako to už bolo naznačené pre použitie tohoto filtra musíme vedieť výkon spektra aktuálneho signálu a šumu. Šum je často aproximovaný Gaussovskému.



Obr. 7.1 Filtrovanie s Wienerovom filtrom

7.3.1 Supresia šumu pomocou TSNR a HRNR algoritmov

V tejto podkapitole je stručne popísaná zaujímavá implementácia Wienerovho filtra s metódami TSNR a HRNR [21]. Filter založený na sledovania a priori SNR pomocou metódou riadenou rozhodnutím (decision-directed method).

Dvojkroková metóda redukovania šumu (TSNR) odstráni „reverbation“ efekt, pričom sa uchovajú výhody metódy riadenej rozhodnutím. Okrem toho klasické techniky krátkodobej redukcie šumu, vrátane TSNR, predstavujú harmonické skreslenie v zlepšenej reči. Pre riešenie tohoto problému je implementovaná metóda **harmonickej regenerácie (HRNR).** Pomocou HRNR je možné

zjemňovať *a priori* SNR, ktorý je používaný na výpočet spektrálneho zisku za účelom uchovania harmonických frekvencií reči.

V prílohe je umiestnená implementácia filtra v jazyku MATLAB s licenčným súborom. Filtrovací algoritmus bol otestovaný na audio stope z filmu. Pôvodná reč v textovej forme a výstupy rozpoznávača pred a po aplikovaní filtra sú umiestnené v prílohe.
Výsledky vykonaných expreimentov ukázajú odlišné tendencie. Pri prítomnosti muziky v pozadí sa ukazovalo zlepšenie (v krátkom testu percent rozpoznaných slov sa zvýšila z 17% na 39%), v iných situáciach zlepšenie nebolo však značné, alebo opačne, niektoré slová sa stratili.

7.4 Hudba v pozadí

Hudba v pozadí je častým problémom pri rozpoznávania reči. Hlavnou vlastnosťou hudby ako škodlivého signálu je nestacionárnosť. Metódy aplikovateľné na izolovanie kvazistacionárnych signálov a zlepšenie reči vo väčšine prípadov nedosahujú očakávané výsledky. Na základe experimentov [24], spôsob zlepšenia rozpoznania reči rušeného nestacionárnym signálom ako je napríklad hudba je prispôsobenie stratégie systémov rozpoznania. Regióny zvukovej stopy s nízkym SNR vyžadujú zvláštne spracovanie. Problém je spôsob lokalizácie týchto regiónov. Na druhej strane kompenzačné metódy by mali byť schopné vysvetliť dočasné vzory rušeného signálu. Pre taký účel dvojdimenzionálny Hidden Markov Model (HMM) môže byť vhodný. V tomto prípade HMM musí poznať hudbu dopredu, alebo musí byť schopný sa naučiť parametry hudby na základe testovacích dát.

7.5 Aplikácie pre redukovanie šumu

7.5.1 Qualcomm-ICSI-OGI front end

Software je dostupný na adrese [23]. Okrem iných obsahuje samotnú aplikáciu **nr** na redukovanie šumu pomocou Wienerovho filtra. Vstupy aplikácie sú: vstupný audio súbor a súbor obsahujúci výstupy detekcie zvukovej aktivity (VAD - voice activity detection), takzvaných **silence flags** (príznaky ticha). Pre každý rámec je určený jeden príznak, ktorý udáva, ktorý označí rámec ako reč alebo nie (speech/non speech). Pre vykonanie VAD je určená aplikácia **silence_flags**, výstupom tejto aplikácie je ASCII súbor s príznakmi ticha. Aplikácia nr číta súbor s príznakmi a redukuje šum odhadnutím na základe výpočtu všetkých „nerečových“ rámcov. Pri testovaní tejto aplikácie bola používaná databáza AURORA-2. Aplikácia pracuje s bezhlavičkovými audio súbormi formátu NIST.

7.5.2 Redukovanie šumu pomocou SoX

Applikácia SoX (Sound eXchange) [22] je známa aj ako švajčiarsky nôž pre spracovanie zvuku. Okrem iných obsahuje aj funkciu redukovania šumu. Túto možnosť aplikácie je možné využiť pomocou prepínača **noisered**.

8 Záver

8.1 Súhrn

Výsledná aplikácia vytvorý titulky k AVI filmom (obsahujúce maximálne dvojkanálové audio) pomocou BSAPI. Extrakcia zvukovej stopy sa koná po segmentoch postupne s prepisovaním reči na text kvôli nízkej spotreby miesta na pevnom disku. Výsledný súbor je typu SRT.

Hlavným cieľom tejto práce bolo vytvorenie aplikácie na vytvorenie filmových titulov k AVI filmom a experimentovanie s rozpoznávačom v reálnych situáciách (aké poskytujú filmy), a prípadné odhalenie nedostatkov. Boli vymenované rôzne techniky pre zvýšenie efektivity rozpoznania (napr. metódy redukovania šumu) a bol uvedený model dokonalej aplikácie pre automatické vytvorenie titulkov k filmom. Wienerov filter pomocou algoritmov TASR a HRNR patrí medzi pokročilé techniky redukovania šumu. Aplikovanie filtra mal určitý prínos pri prítomnosti hudby v pozadí ale v ostatných prípadoch jeho efekt bol zanedbateľný, či negatívny.

Vytvorená aplikácia nie je dokonalý produkt, má byť považovaný za experimentálny nástroj, krorý slúži k mapovaniu schopností BSAPI pri aplikovania na filmoch. Dosiahnuté výsledky môžu byť uvažené za neúspšné.

8.2 Pokročilý systém pre aut. tvorbu titulkov

Vytvorenie pokročilého systému pre automatickú tvorbu tiulkov je zložitá záležitosť. Je to vôbec pochybné, či sa dá taký systém realizovať s dnešnými najmodernejšími technikami. Pri vytvorení a testovaní tejto aplikácie sa sformolovali požiadavky takého systému. Splňenie týchto požiadaviek by umožnilo minimalizovanie rozdielov medzi automaticky vytvorenými titulkami a tilulkami vytvorenými človekom.

Hlavnými požiadavkami sú robustné rozpoznávacie systémy :

- Rozpoznávač reči – Robustný rozpoznávač reči pre prepis reči na text aj pri prítomnosti rušivých činiteľov.
- Rozpoznávač viet a ich modality – Dôležitou súčasťou pokročilého systému pre automatickú tvorbu titulkov je rozpoznanie viet a ich modality. Rozpoznanie modality viet je dôležité hlavne pri vyznačovaní interpunkcie. Pre poskytnutie dokonalého textového výstupu je nepostrádateľný. Bez interpunkčných znamienok sa stráca pôvodný význam viet. V súčasnoti existujú rôzne expreimentačné metódy pre také účeli. V technických správ [27] a [25] sú popísané také systémy pre rozpoznanie modality reči. Sú zaujímavé audio-vizuálne metódy pre rozpoznanie modality viet, ale vo filmoch sú ťažko použiteľné.
- Identifikácia rôznych rečníkov – Je dôležitá najmä v takých situáciach, keď do reči rečníka „skočí“ iný rečník, alebo všeobecne keď koniec reči rečníka je z časového hľadiska veľmi blízko k začiatku reči druhého rečníka. Bez takého systému nie je možné v každom prípade oddeliť reč rôznych rečníkov v textovom výstupu. Novšia verzia BSAPI je už schopný identifikovať reč rôznych rečníkov pomocou diarizácie (rozhrania SSpeakerID2I a

SDiarizationI ktorými). Búdúca práca na projekte by mala zhrnúť experimentovanie s novým systémom na filmoch.

8.3 Práca v budúcnosti

Získané znalosti experimentovaním, odhalili rôzne nedostatky aplikáce a rozpoznávača. Špatné rozpoznanie niektorých slov môže zapríčiniť aj nepresnonsť operácie seek pri segmentácie zvukového súboru. Napriek tomu, že rozpoznávač obsahuje band-pass filter (prepustí pásmo od 60Hz do 3400Hz), efektivitu rozpoznania reči môže zvýrazne znížiť šum alebo muzika v pozadí, reč súčasne hovoriaceho alebo afektovaná reč. Reč vo filmoch sa líši od bežnej reči (zvýrazňejšie akcenty, afektácie). Pri testoch sa objavila tá zaujímavosť, že pre dané slovo bol spoznaný foneticky podobný, ale inak odlišný výraz, čo môže znamenať potrebu pretrénovania rozpoznávača. Je potrebné zlepšiť efektivitu rozpoznávania pri prítomnosti šumu alebo iných negatívnych činiteľov na rozpoznanie reči, prípadne s nestacionárnou charakteristikou aká je hudba. To znamená veľkú výzvu. V kapitole 8.2 boli uvedené požiadavky na pokročilý systém pre automatickú tvorbu titulkov. Cieľom budúcej práce môže byť vytvorenie takého systému. Veľmi zaujímavá je problematika robustného rozpoznania reči a rozpoznania modality viet. Ďalším cieľom po disiahnutí vytvorenia kvalitných titulkov môže byť vytvorenie aplikácie pre automatickú tvorbu viacjazyčných titulkov českého filmu pomocou prekladateľským API.

Literatúra

[1] *Phonexia* [online]. 2011-03-02 [cit. 2011-05-15].
URL http://www.phonexia.com/

[2] *Brno Speech Application Interface Documentation* [online]. 2011-02-15 [cit. 2011-05-15].
URL http://www.phonexia.com/docs/bsapi/

[3] *WAVE Audio File Format* [online]. 2009-09-08 [cit. 2011-05-15].
URL http://www.digitalpreservation.gov/formats/fdd/fdd000001.shtml

[4] *Linear Pulse Code Modulated Audio (LPCM)* [online]. 2009-09-08 [cit. 2011-05-15].
URL http://www.digitalpreservation.gov/formats/fdd/fdd000011.shtml

[5] *AVI (Audio Video Interleaved)* [online]. 2009-05-13 [cit. 2011-05-15].
URL http://www.digitalpreservation.gov/formats/fdd/fdd000059.shtml

[6] *The Component Object Model* [online]. 2011-01-27 [cit. 2011-05-15].
URL *http://msdn.microsoft.com/en-us/library/ms694363%28v=vs.85%29.aspx*

[7] *Microsoft Windows Software Development Kit* [online]. 2011 [cit. 2011-05-15].
URL http://msdn.microsoft.com/en-us/windows/bb980924

[8] *DirectShow* [online]. 2011-05-03 [cit. 2011-05-15].
URL http://msdn.microsoft.com/en-us/library/dd375454%28v=vs.85%29.aspx

[9] *DirectShow Reference* [online]. 2011-05-03 [cit. 2011-05-15].
URL http://msdn.microsoft.com/en-us/library/dd375467%28v=VS.85%29.aspx

[10] *Supported Formats in DirectShow* [online]. 2011-05-03 [cit. 2011-05-15].
URL http://msdn.microsoft.com/en-us/library/dd407173%28v=vs.85%29.aspx

[11] *SubRip text file format* [online]. 2011-04-07 [cit. 2011-05-15].
URL http://en.wikipedia.org/wiki/SubRip#SubRip_text_file_format

[12] *Visual Studio Home | Microsoft Visual Studio* [online]. 2011 [cit. 2011-05-15].
URL http://www.microsoft.com/visualstudio/en-us/home

[13] *Learn to Program for Windows in C++* [online]. 2010-10-05 [cit. 2011-05-15].
URL http://msdn.microsoft.com/en-us/library/ff381399%28v=vs.85%29.aspx

[14] *Co je autorské právo | Filmy nejsou zadarmo* [online]. 2011 [cit. 2011-05-15]. URL http://www.filmynejsouzadarmo.cz/cs/co-je-autorske-pravo/

[15] *There's No Disgrace Like Home - Wikipedia, the free encyclopedia* [online]. 2011-05-06 [cit. 2011-05-15]. URL http://en.wikipedia.org/wiki/There%27s_No_Disgrace_Like_Home

[16] *Hele Kámo, kdo tu vaří? / Waiting... (2005) | ČSFD.cz* [online]. 2011 [cit. 2011-05-15]. URL http://www.csfd.cz/film/156707-hele-kamo-kdo-tu-vari/

[17] *Záznamy - Video server 1* [online]. 2011-04-15 [cit. 2011-05-15]. URL https://video1.fit.vutbr.cz/av/records.php?id=19172&categ_id=768

[18] *Město / Town, The (2010) | ČSFD.cz* [online]. 2011 [cit. 2011-05-15]. URL http://www.csfd.cz/film/269227-mesto/

[19] *The Mutants Are Revolting - Wikipedia, the free encyclopedia* [online]. 2011-05-06 [cit. 2011-05-15]. URL http://en.wikipedia.org/wiki/The_Mutants_Are_Revolting

[20] Xiao, X. *Speech enhancement with applications in speech recognition*. , 2005. 81 s. First year report. School of Computer Engineering, Nanyang Technological University, Singapore.

[21] Plapous, C.; Marro, C.; Scalart, P.: Improved Signal-to-Noise Ratio Estimation for Speech Enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, ročník 14, č. 6, nov. 2006: s. 2098-2108, ISSN 1558-7916, doi:10.1109/TASL.2006.872621.

[22] *SoX - Sound eXchange | HomePage* [online]. 2011-04-20 [cit. 2011-05-15]. URL http://sox.sourceforge.net/

[23] *Qualcomm-ICSI-OGI front end archive* [online]. 2008 [cit. 2011-05-15]. URL http://www.icsi.berkeley.edu/Speech/papers/qio/

[24] Raj, B.; Parikh, V.; Stern, R.: The effects of background music on speech recognition accuracy. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, ročník 2, apr 1997, s. 851-854 vol.2, doi:10.1109/ICASSP.1997.596069.

[25] Král, P.; Klečková, J.; Ceirsara, Ch. : Scentece Modality Recognition in French based on Prosody, *World Academy of Science, Engineering and Technology 8 2005*, 2005, s. 185-188

[26] Zhu, Q.; Iseli, M.; Cui, X.; aj.: Noise Robust Feature Extraction for ASR Using the Aurora 2 Database. In *Proc. of EuroSpeech*, s. 185-188.

[27] Sztahó, D.; Nagy, K.; Vicsi, K.: *Automatic Sentence Modality Recognition in Children's Speech, and Its Usage Potential in the Speech Therapy*. Berlin, Heidelberg: Springer-Verlag, 2009, ISBN 978-3-642-03319-3, s. 266-275, doi:10.1007/978-3-642-03320-9_25.
URL http://portal.acm.org/citation.cfm?id=1601853.1601881

Príloha A

Tabuľky výsledkov testov

	postava1
N	5
N_{ok}	0
N_{fp}	2
N_{miss}	3
P_{ok} [%]	0

tabuľka A.1 Výsledky časti a) testu č. 1

	postava1	postava2	postava3
N	12	4	2
N_{ok}	5	2	0
N_{fp}	6	1	0
N_{miss}	1	1	2
P_{ok} [%]	39		

tabuľka A.2 Výsledky časti b) testu č. 1

	postava1	postava2	postava3	postava4	postava5	postava6
N	64	8	8	4	1	16
N_{ok}	22	4	0	0	0	8
N_{fp}	24	2	1	2	0	4
N_{miss}	18	2	7	0	1	4
P_{ok} [%]	34					

tabuľka A.3 Výsledky časti c) testu č. 1

	postava1	**postava2**	**postava3**	**postava4**
N	69	32	12	25
N_{ok}	40	14	1	0
N_{fp}	13	7	3	0
N_{miss}	16	11	8	25
P_{ok} [%]	47			

tabuľka A.4 Výsledku testu č. 2

	postava1	**postava2**	**postava3**	**postava4**	**postava5**
N	5	3	9	3	2
N_{ok}	3	1	0	0	0
N_{fp}	1	2	1	0	0
N_{miss}	1	0	8	3	2
P_{ok} [%]	18				

tabuľka A.5 Výsledky časti a) testu č. 3

	postava1
N	27
N_{ok}	0
N_{fp}	0
N_{miss}	27
P_{ok} [%]	0

tabuľka A.6 Výsledky časti b) testu č. 3

	postava1	**postava2**
N	6	12
N_{ok}	1	8
N_{fp}	1	4
N_{miss}	4	0
P_{ok} [%]	50	

tabuľka A.7 Výsledky časti c) testu č.3

	postava1	postava2
N	9	6
N_{ok}	0	0
N_{fp}	0	0
N_{miss}	9	6
P_{ok} [%]	0	

tabuľka A.8 Výsledky časti d) testu č.3

	postava1	postava2
N	51	27
N_{ok}	30	18
N_{fp}	9	2
N_{miss}	12	7
P_{ok} [%]	61	

tabuľka A.9 Výsledky časti a) testu č.4

	postava1	postava2
N	48	5
N_{ok}	15	1
N_{fp}	12	1
N_{miss}	21	3
P_{ok} [%]	30	

tabuľka A.10 Výsledky časti b) testu č. 4

	postava1	postava2	postava3
N	11	23	14
N_{ok}	0	2	0
N_{fp}	0	2	0
N_{miss}	11	21	14
P_{ok} [%]	4		

tabuľka A.11 Výsledky časti a) testu č. 5

	postava1	postava2	postava3	postava4	postava5
N	9	6	7	25	18
N_{ok}	3	4	0	10	2
N_{fp}	1	2	1	3	1
N_{miss}	5	0	6	12	15
P_{ok} [%]	29				

tabuľka A.12 Výsledky časti b) testu č. 5

	postava1	postava2
N	14	7
N_{ok}	1	1
N_{fp}	1	0
N_{miss}	12	6
P_{ok} [%]	10	

tabuľka A.13 Výsledky časti a) testu č. 6

	postava1	postava2
N	14	17
N_{ok}	0	4
N_{fp}	4	3
N_{miss}	8	10
P_{ok} [%]	13	

tabuľka A.14 Výsledky časti b) testu č. 6

	postava1
N	16
N_{ok}	0
N_{fp}	0
N_{miss}	16
P_{ok} [%]	0

tabuľka A.15 Výsledky časti c) testu č. 6

	postava1
N	143
N_{ok}	108
N_{fp}	17
N_{miss}	18
P_{ok} [%]	75

tabuľka A.16 Výsledky časti a) testu č. 7

	postava1
N	71
N_{ok}	46
N_{fp}	7
N_{miss}	18
P_{ok} [%]	65

tabuľka A.17 Výsledky časti b) testu č. 7

Legenda:
N – celkový počet slov
N_{ok} – správne spoznané slová
N_{fp} – počet foneticky podobných tvarov
N_{miss} – počet nespoznaných slov
P_{ok} – počet správne spoznaných slov v percentách

Príloha B

Obsah CD

- Visual Studio 2010 project obsahujúci zdrojové kódy aplikácie
- súčasti balíku BSAPI
- dokumentácia k aplikácie
- výstupy testov a textové súbory obsahujúce pôvodný reč
- zdrojový kód Wienerovho filtra a testovacie dáta
- technická správa v elektronickej podobe