Zvaná přednáška

Efektivní přístup ke znalostem v audio-vizuálních záznamech

Igor SZÖKE[1,2], Michal FAPŠO[1], Josef ŽIŽKA[2], Vítězslav BERAN[1],
Jan "Honza" ČERNOCKÝ[1]

[1]*ÚPGM a IT4Innovations Centre of Excellence, Fakulta informačních technologií, VUT v Brně*
Božetěchova 2, 612 66 Brno
szoke@fit.vutbr.cz

[2]*ReplayWell, s. r. o.*
U vodárny 2a, 616 00 Brno

Abstrakt. Množství audiovizuálních dat se neustále zvětšuje. Část z nich obsahuje důležité informace - záznamy přednášek, konferencí, kurzů atd. Tato data jsou však pro současné vyhledávače (Google, Seznam) neviditelná. Pokud potřebujeme nalézt záznam, který se týká konkrétní problematiky nebo se jen potřebujeme rychle zorientovat v záznamu, současné vyhledávače nám mnoho nepomohou. V první části přednášky se podíváme, jak zviditelnit audiovizuální záznamy prezentací pro standardní textové vyhledávače. Z technického hlediska si ukážeme, jak vypadá celkové schéma systému, a pak se podrobněji podíváme na jednotlivé komponenty. Zejména na komponentu automatického rozpoznávače řeči (převod audia do textu s časováním), a komponentu automatické synchronizace a rozpoznání slajdů s videem (převod videa do textu s časováním). Dále si ukážeme, jak jsou zpracována audiovizuální data reprezentována pro vyhledávací engine Lucene. V druhé části přednášky se zaměříme na prezentaci uživatelům, a také na zpětnou vazbu od uživatelů. Systém je prakticky nasazen na Fakultě informačních technologií VUT v Brně, a také pro několik velkých konferencí.

Klíčová slova: audiovizuální záznam, převod řeči do textu, převod obrazu do textu, indexace a vyhledávání, web

1 Úvod

Během posledních let se stále častěji setkáváme s nahráváním různých konferencí, seminářů a přednášek.. Vhodně pořízené videozáznamy mohou mít velký přínos při vstřebávání informací. Přednášející může poukázat na zajímavé souvislosti, může použít elektronickou tužku nebo třeba vysvětlit nějaký numerický příklad. Jeho výklad může mít tedy zásadní vliv na pochopení dané problematiky. Častým problémem bývá ovšem orientace ve videozáznamech, které u přednášek mívají běžně délku i několika hodin. Jak tedy rychle najít ve videozáznamech požadovanou informaci? Mluvil o ní přednášející na začátku, uprostřed či na konci přednášky? Na kterém slajdu byla? Nezmínil se o ní i v jiné přednášce?
Druhým problémem spojeným s audiovizuálními záznamy je jejich dohledatelnost přes standardní Internetové vyhledávače jako jsou Google nebo Seznam. Roboty těchto vyhledávačů se často musí spokojit pouze s názvem přednášky a stručným abstraktem.

Přestože videozáznamy bývají běžně doplněny nějakým titulkem, krátkým popisem či klíčovými slovy, pro usnadnění orientace v archivu videí a dohledatelnosti jednotlivých záznamů je to stále málo. Pro tyto případy může být velkou pomocí systém pro převod řeči či obrazu do textu. A právě tyto technologie používá služba SuperLectures.com (Prednasky.com) prezentovaná v tomto článku.

Na projektu prohlížeče audiovizuálních záznamů jsme ve skupině zpracování řeči Speech@FIT na Fakultě informačních technologií VUT v Brně začali pracovat v roce 2005. V té době byla naše výzkumná skupina součástí evropského projektu Augmented Multi-party Interaction (AMI - EU-6FP-IST). Cílem AMI projektu bylo vytvořit systém pro podporu meetingů. Meetingy byly nahrány (audio a video), a dále zpracovány (automatický přepis řeči, tvorba abstraktů, analýza struktury meetingu, analýza interakce člověk-člověk atd.). Vzhledem k tomu, že se na Fakultě informačních technologií VUT od roku 2005 začaly nahrávat přednášky, rozhodli jsme se využít těchto záznamů pro prezentaci našich technologií rozpoznávaní řeči. Následně jsme se rozhodli vyvinutý systém odpoutat od "uzavřeného laboratorního experimentování". V současné době se pokoušíme vyvinutý systém s podporou Technologické Agentury ČR komerčně aplikovat pro zpracování záznamů z konferencí.

V druhé sekci tohoto článku nastíníme, jaký druh informací je přítomen v audiovizuálním záznamu přednášek či konferencí. Následující sekce bude věnována třem různým pohledům na efektivní přístup k informacím ze záznamu, a to: Technickém, Uživatelském a pohledu Provozovatele služby. Čtvrtá a pátá sekce se bude podrobněji zabývat technickou stránkou služby. Ve čtvrté si ukážeme, jak informaci vyextrahovat, a v páté jak ji indexovat. Prezentační forma služby je též důležitá součást. Budeme se jí věnovat v šesté sekci. Poslední sekce shrnuje zkušenosti uživatelů a provozovatele přednáškového portálu SuperLectures.com.

2 Co je to audio-vizuální záznam

Audio-vizuálním (AV) záznamem přednášky na vysoké škole či konferenci v kontextu tohoto článku rozumíme video záznam pořízený kamerou (nebo kamerami) doplněný o zvukovou stopu obsahující mluvený projev přednášejícího. Videozáznam může obsahovat záběr na přednášejícího (tak, jak jsme zvyklí z televize), ale ten nám obvykle mnoho zajímavých informací neposkytne. Mnohem užitečnější je vizuální záznam projekčního plátna. Při přednáškách se v dnešní době hojně používá právě projekční plátno, na které jsou promítány slajdy nebo je naživo demonstrováno řešení nějakého problému. Posluchač má tak k dispozici i názornou informaci, která se nemusí vyskytovat v mluveném projevu. Navíc lze doplnit AV záznam o další modality jako jsou skripta v elektronické podobě, abstrakty nebo odkazy na další zdroje či přednášky na podobné téma.

2.1 Informace v audiovizuálním záznamu

Informace v AV záznamu je skryta v různých modalitách: audio, obraz, text. Vzhledem k tomu, že veškerá komunikace s počítačem a vyhledávání je v současné době řešeno přes textovou modalitu, je naším cílem převést maximum informace a AV záznamu do textu. Pokud bychom měli formálně popsat druhy informace v AV záznamu, které by mohly být k dispozici pro zpracování a následné zpřístupnění, jednalo by se o:

- Mluvený projev - textová reprezentace zvukového záznamu.
- Důležitá klíčová slova obsažená v mluveném projevu.

- Slajdy - textová reprezentace jednotlivých slajdů promítaných na projekční plátno.
- Časování slajdů.
- Abstrakt přednášky.
- Přednášející, případně další zůčastněné osoby.
- Odkazy na doplňující materiály.
- Elektronická skripta.

2.2 Časová dimenze "dokumentu"

Při zpracování běžných textových dokumentů (webové stránky nebo články) neexistuje časová dimenze. Jeden dokument (stránka, článek) se považuje za atomickou jednotku. Pro případ vyhledávání v textových dokumentech se dokument může rozložit na jednotlivá slova, kdy každé má svůj index (patnácté slovo v tomto odstavci je "stránka") a dále s touto informací pracovat.
Na rozdíl od textových dokumentů obsahuje audiovizuální dokument právě onu časovou dimenzi. Každá informace (pronesené slovo, zobrazený slajd) má přesně definovaný interval, ve kterém byla zaznamenána. Pokud bychom pracovali se záznamem jako celkem a stačilo by nám například pouze vyhledávat celé dokumenty (přednášky), můžeme tuto časovou dimenzi zanedbat a pracovat s přepisem řeči či slajdů čistě jako s řetězcem textu. Pokud však chceme zachovat uživatelský komfort a nabídnout možnost nasměrovat posluchače přímo do konkrétního místa v záznamu, musíme časovou dimenzi zachovat a vhodným způsobem ji reprezentovat v indexu pro pozdější vyhledávání. Dále viz kapitola 5.

2.3 Pravděpodobnostní dimenze "dokumentu"

Podobně jako časová dimenze se u AV záznamů musíme vypořádat ještě s tzv. pravděpodobnostní dimenzí. U textových dokumentů je obvykle pravděpodobnost každého slova binární. Tedy slovo v textu na dané pozici buď je nebo není. Bohužel algoritmy pro převod řeči do textu pracují v pravděpodobností rovině. Ke vstupnímu řečovému záznamu vrací nejpravděpodobnější sekvenci slov, která odpovídá vstupnímu záznamu (obrázek 1). Tyto algoritmy jsou však schopny vyprodukovat i paralelní hypotézy (reprezentovány dopředným acyklickým grafem) a ke každému slovu doplnit i věrohodnost. Opět, pokud nám stačí pracovat pouze s nejpravděpodobnějším přepisem (zobrazit ho jako titulky), nemusíme se paralelními hypotézami a pravděpodobností zabývat. Pokud však chceme uživatelům zpřístupnit možnost vyhledávat v záznamech a výsledky by měly být co možná nejpřesnější, jsou pro nás grafy paralelních hypotéz nutností. Bližší vysvětlení jak zacházet s grafy hypotéz je v sekci 5.1.

```
0.00 0.00 <s> 0.00000
0.00 0.30 JO -2743.351074
0.30 0.68 PROTOŽE -3934.467529
0.90 0.90 TA -67.685501
0.90 1.10 FOURIEROVA -2338.387451
1.10 1.62 TRANSFORMACE -5231.450195
1.62 2.12 MÁ -4890.719238
2.53 2.64 NAPROSTO -1125.318604
3.37 3.52 PŘESNÉ -1547.275269
3.86 4.08 FREKVENČNÍ -2253.997070
```

```
4.08 4.27 ROZLIŠENÍ -1949.657471
4.27 4.73 NA -4534.286133
4.73 4.96 VŠECH -2360.818604
4.96 5.11 FREKVENCÍCH -1763.967285
5.11 5.11 </s> 0.000000
```

Obr. 1. Příklad nejpravděpodobnější sekvence rozpoznaných slov pro jeden segment. První a druhý sloupec reprezentují čas v sekundách, třetí sloupec slovo a čtvrtý sloupec věrohodnost (v logaritmu).

3 Prohlížeč audiovizuálních záznamů a tři pohledy na něj

Vzhledem k tomu, že se se službou SuperLectures.com nepohybujeme pouze v "teoretické" rovině laboratorního prostředí, ale také v "praktické" rovině komerčního nasazení, je vhodné se na celkovou problematiku získávání informací z AV záznamů dívat nejen očima technika vědce (inženýra a programátora), ale též uživatele (posluchače) a zákazníka (provozovatele systému).

3.1 Pohled 1. - uživatel

Uživatel (posluchač) je ten, kdo "definuje" nejen rozhraní, ale také jak má celá služba ve výsledku vypadat. Uživatel služby chce aby:

- Rychle našel to co hledá. Čas jsou peníze.
- Výsledný prohlížeč byl pěkný a měl intuitivní ovládání (mimo zaměření tohoto článku).
- Automatické titulky nebo přepis byly přesné.
- Vyhledávání v záznamech bylo rychlé (v nejhorším stovky milisekund na dotaz).
- Vyhledávání v záznamech bylo přesné.
- Celková orientace v záznamech byla rychlá a pohodlná.

3.2 Pohled 2. - provozovatel

Zákazník (provozovatel) je ten, kdo si nechává zpracovat data, a také za poskytnutí této služby platí. Může se jednat o organizátora konference, univerzitu či firma specializující se na školení. Zákazník služby chce aby:

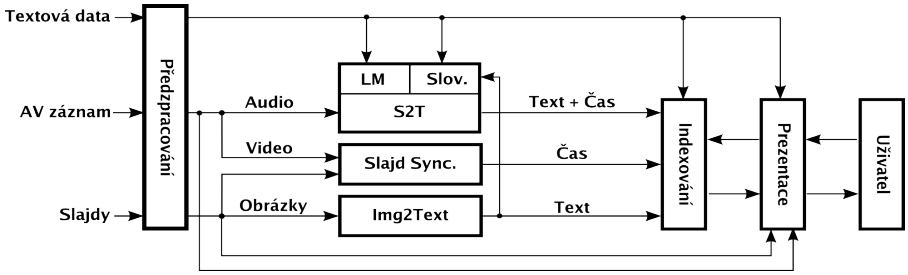
- Nasazení služby vydělávalo. Toto je mimo zaměření tohoto článku, ale implikuje to nutnost minimálních nákladů. Tedy, aby zpracování záznamů a provozování služby bylo co nejméně náročné na hardware a software. Pokud budeme uvažovat příjmovou stránku služby, tak za předpokladu, že zákazník záznamy volně zveřejní, chce, aby měly co největší dopad - lidé je snadno našli pomocí vyhledávačů jako je Google nebo Seznam.
- Z praktického hlediska by mělo být zpracování záznamů maximálně automatizované.

3.3 Pohled 3. - inženýr

Inženýr jakožto vývojář služby se musí přizpůsobit uživateli a provozovateli. Celou službu by měl navrhnout tak, aby byla modulární a dobře škálovatelná (což je opět mimo zaměření tohoto článku).

4 Extrakce informace z audiovizuálních záznamů

Tato sekce se bude věnovat technickému popisu důležitých částí systému. Nejprve se podívejme na diagram celkového zpracování záznamu (obrázek 2).



Obr. 2. Schéma celého systému pro zpracování audiovizuálních záznamu přednášek.

V části *předzpracování* se jedná o úpravu audia a videa (střih, intro, …), převod slajdů z PDF do obrázků (JPG) a vyčištění textových dat od formátovacích značek. Pokud jsou dispozici elektronická skripta či sborník konference, mohou se tato textová data dále analyzovat a vytvořit z nich jazykový model a výslovnostní slovník, kterými se zadaptuje rozpoznávač řeči. Tím je možné dosáhnout přesnějšího přepisu řeči. AV záznam se dále rozdělí na audio a video.
Po předzpracování následují jednotlivé bloky pro extrakci informace:

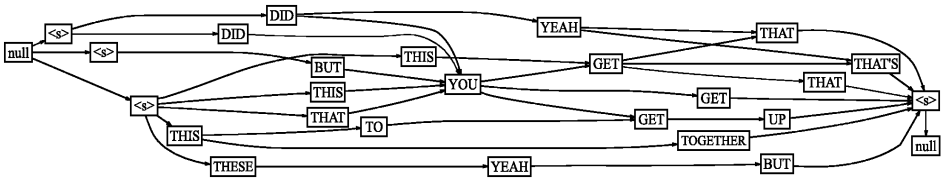
- *Převod řeči do textu* (S2T). Vstupní audio záznam převedeme do textové podoby s časovou informací (začátek každého slova), a také do grafu paralelních hypotéz (obrázek 3).
- *Automatická synchronizace slajdů* s videem (Slide Sync.). Synchronizace slajdů je postup, kdy se automaticky snažíme spárovat snímek z dodané prezentace (ve formátu PDF) s videozáznamem promítacího plátna.
- *Převod slajdů do textu* (Img2Text). Tento postup nahrazuje obecný a také velmi obtížný převod videa do textu. Pomocí spárování videa se slajdy a následného převodu slajdů do textu si celý proces zjednodušíme při zachování maximálního množství extrahované informace. Převod slajdů do textu je totiž relativně jednoduchý. Pokud máme k dispozici zdrojové PDF nebo PowerPoint, můžeme běžně dostupnými nástroji ze slajdů vyextrahovat text. Jako alternativní se nabízí možnost použití OCR (Optical Character Recognition) technik, kdy převedeme obraz do textu přímo rozpoznáním znaků v obraze.

Textová informace společně s časováním je zaindexována a uložena pro následné vyhledávání. Za indexací následuje prezentační vrstva, kde jsou veškerá zpracovaná data vhodným způsobem prezentována uživateli pomocí webového rozhraní v tzv. přehrávači. Uživatel interaguje s přehrávačem a v případě zadání vyhledávacího dotazu (query) je dotaz přeposlán na vyhledávací server (Lucene). Výsledky vyhledávání jsou pak prezentační vrstvou zobrazeny uživateli.

4.1 Převod řeči do textu

Zřejmě nejdůležitějším blokem v celém systému je převod řeči do textu (S2T - speech to text), který umožní "zviditelnit" většinu netextové informace. Vstupem do tohoto bloku je

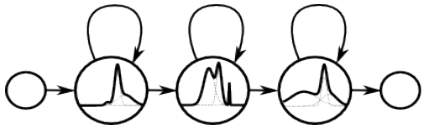
běžný audio záznam (WAV). Výstupem je pak sekvence slov a časových značek (obrázek 1) a také graf paralelních hypotéz (obrázek 3).



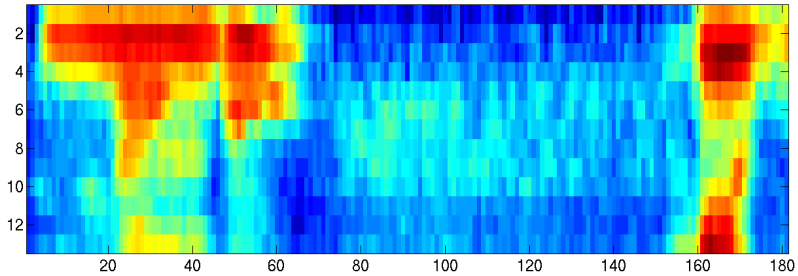
Obr. 3. Acyklický orientovaný graf paralelních hypotéz (též nazývaný "lattice"). Jednotlivé uzly reprezentují slova včetně časové informace začátku. Přechody pak reprezentují konkrétní výskyt slova (čas od/do) a jeho věrohodnost.

Detailní popis S2T [4] je nad rámec tohoto článku. Zde si vystačíme s faktem, že celý blok S2T obsahuje tři důležité komponenty: Akustický model, Jazykový model a Výslovnostní slovník (obrázek 6).

Akustický model je trénován na stovkách hodin pečlivě přepsaných zvukových záznamů (1000 hodin v případě angličtiny a 200 hodin v případě češtiny). Dále je adaptován na akustický kanál přednášek (akustika místnosti, typ mikrofonu a způsob promluvy řečníků). K první adaptaci bylo použito několik hodin pečlivě přepsaných záznamů z přednášek a konferencí. Vzhledem k tomu, že je akustická inter-variabilita "přednáškových" záznamů nízká, je dále pro jednotlivé přednášející použita adaptace bez učitele. Nejsou tedy třeba žádné ruční zásahy při zpracování dalších dat. Základní jednotkou v akustickém modelu je kontextově závislý foném (foném - základní zvuková jednotka řeči). Model fonému je reprezentován skrytým Markovovým modelem s pěti stavy, kde tři vnitřní stavy obsahuje směsici gausovek (obrázek 4). Tato směsice modeluje pravděpodobnostní rozložení parametrů. V rozpoznávání řeči obvykle pracujeme s časovým rámcem 10ms, který je reprezentován vektorem parametrů odvozených ze spektra (obrázek 5).

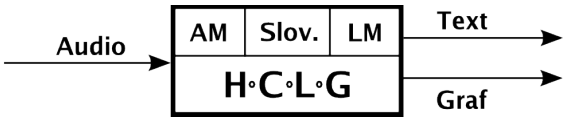


Obr. 4. Skrytý Markovův model akustické reprezentace kontextově závislého fonému. Druhý, třetí a čtvrtý stav obsahují směsici gausovek pro výpočet akustické věrohodnosti vstupního rámce řeči.



Obr. 5. Prvních třináct parametrů odvozených ze spektra pro větu "dobrý den". Osa x je čas, jeden sloupec odpovídá 10ms. Osa y jsou parametry odvozené ze spektra.

Jazykový model reprezentuje gramatiku jazyka. Vzhledem k tomu, že formální popis gramatiky přirozeného jazyka by byl extrémně složitý, spokojíme se se statistickým přístupem. Jazykový model obsahuje statistiky sekvencí slov (samostatná slova - unigramy, dvojice slov - bigramy a trojice slov - trigramy). Protože je obtížné pokrýt jazykovým modelem všechny specifické domény, obvykle dochází k adaptaci tohoto modelu na konkrétní doménu dat. Při adaptaci se statistiky slov získané z přepisu skutečné řeči (obecný jazykový model) smíchají se statistikami získanými například právě z elektronických skript či ze sborníku.
Výslovnostní slovník mapuje slova z jazykového modelu na fonémy, které jsou v akustickém modelu. Vytvořit výslovnostní slovník pro češtinu není velký problém, protože čeština je fonetický jazyk (píšeme skoro stejně jako čteme). Komplikací jsou pouze přejatá slova, u kterých musíme automaticky získanou výslovnost korigovat ručně. Pro angličtinu už je s tvorbou výslovností pomocí pravidel problém, takže se zde uchylujeme opět ke statistickému přístupu (například za pomocí váhovaných konečných stavových převodníků [1]).



Obr. 6. Reprezentace S2T pomocí kompozice váhovaných konečných stavových převodníků.

Technicky je převod řeči do textu řešen pomocí konečných váhovaných stavových převodníků - weighted finite state transducers (obrázek 6). Je vytvořena takzvaná rozpoznávací síť, která je kompozicí:

- *Gramatiky G* (akceptor) reprezentující jazykový model. Ta nám říká, jaká může být sekvence slov v promluvě a s jakou váhou (věrohodností) následují jednotlivá slova po sobě.
- *Převodníku L* reprezentující výslovnostní slovník. Ten převádí slova na řetězce fonémů.
- *Převodníku C* reprezentujícího kontextově závislé fonémy. Zde se jednotlivé fonémy doplní o kontextovou informaci, tedy jaký foném předchází a následuje.
- *Převodníku H* reprezentujícího skryté Markovovy modely kontextově závislých fonémů v akustickém modelu.

Po této kompozici vznikne rozsáhlý orientovaný graf, kde jednotlivé stavy mají přiřazeny směsice gausovek. Místy se v rozpoznávací síti vyskytují i speciální "slovní" uzly, které říkají, že se má na výstup vypsat slovo. Tato síť se načte do speciálního programu (tzv. dekodéru). Vstupní nahrávka (ta, která se má přepsat na text) se převede na matici příznaků, a také se vloží do dekodéru. Pak se spustí algoritmus rozpoznávání (tzv. token passing). Cílem je najít pro zadanou matici příznaků nejpravděpodobnější cestu grafem (rozpoznávací sítí). Po nalezení nejlepší cesty, se zjistí sekvence slovních uzlů, které leží na této cestě. Tato sekvence je výstup rozpoznávače - tedy text, který byl rozpoznán a měl by být řečen ve vstupní nahrávce. Dekodér je možné též přepnout do módu, kdy na výstup generuje graf hypotéz. Tento graf je podgrafem teoreticky nekonečné rozpoznávací sítě s tím, že jsme díky vstupní nahrávce (akustické informaci) značně omezili stavový prostor je na smysluplné sekvence slov.

Jednotlivé přechody v grafu obsahují akustické a jazykové věrohodnosti vygenerované dekodérem. Tyto věrohodnosti nejsou mezi jednotlivými záznamy normované, a tak je nelze přímo využít k odhadu jistoty, s jakou byla v globálním měřítku jednotlivá slova rozpoznána. Proto se tyto věrohodnosti přepočítávají na takzvané posteriorní pravděpodobnosti [6]. Tím dostaneme pro každé slovo hodnotu mezi 0 a 1, a jistota s jakou byla jednotlivá slova vyslovena v audiu je porovnatelná i mezi jednotlivými záznamy. To je důležité pro vyhledávání a řazení záznamů podle relevance k vyhledávacímu dotazu (query).

Automatická detekce klíčových slov

Pro lepší dohledatelnost zpracovaných záznamů z Internetu (SEO - Search Engine Optimization) jsme implementovali automatickou extrakci klíčových slov z textového přepisu řeči (například "skrytý Markovův model" nebo "převod řeči do textu"). Jedná se o metodu využívající branching entropy a vychází z článku [2]. Princip spočívá v předpokladu, že klíčové slovo (nebo frázi) přímo předchází a následuje ji velké množství různých slov (velká branching entropy). Zatímco uvnitř klíčové fráze je branching entropy nízká. Za slovem "Markovův" může následovat jen několik slov (například "model", "řetězec", ...) což implikuje nízkou entropii, ale za slovem "model" může být velké množství slov (například "je", "může", "má", ...) což značí velkou entropii. Pokud bychom tento postup popsali bodově, tak:

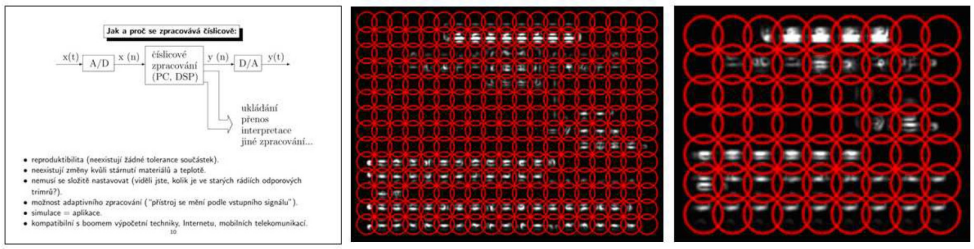
- Ze vstupního textu odstraníme "stop" slova (předložky, spojky, ...) a tvary převedeme do kořenové formy.
- Vytvoříme n-gramový slovník (například pro n rovno 1 až 4).
- Ke každému n-gramu vypočítáme počty všech různých přímých předchůdců a přímých následovníků.
- Z těchto počtů vypočítáme entropii předchůdce (levá branching entropie) a entropii následovníka (pravá branching entropie).
- Pokud je například pravá branching entropie 3-gramu ABC nízká (za n-gramem zprava následuje jen malý počet slov, například X nebo Y nebo Z), ale pravá entropie 4-gramu ABCX je vysoká (zprava následuje za ABCX velké množství různých slov), lze předpokládat, že X ohraničuje klíčovou frázi zprava.
- Stejně lze nalézt hranici klíčové fráze zleva.

Takto vygenerovaná klíčová slova jsou následně použita v kódu HTML stránky jako klíčové fráze pro zvýšení relevance záznamu přednášky.

4.2 Automatická synchronizace slajdů s videozáznamem

Jedná se o automatický proces, kde vstupem je videozáznam, který obsahuje v záběru celé projekční plátno a PDF soubor s promítanými slajdy. Z PDF se v kroku předzpracování vyextrahují jednotlivé snímky a ty se uloží jako knihovna obrázků.
V prvním kroku je ve video záznamu detekovaná plocha plátna (obecný čtyřúhelník) pomocí kombinace detektoru gradientu a detektoru obrázků. Detektor gradientu převede videozáznam do jasové složky a v něm lokalizuje stabilní čtyřúhelníky. Detektor obrázků načte jednotlivé slajdy z knihovny a hledá geometrickou transformaci mezi obrázkem slajdu a analyzovaným snímkem videa. Následně se pomocí RANSAC algoritmu vypočítá homografie slajdů do videa. Výsledkem obou detektorů je menší či větší množství čtyřúhelníků, ze kterých se vybere jeden reprezentant s největší věrohodností.

Ve druhém kroku se použijí souřadnice plátna v záznamu a provede se synchronizace obrázků v knihovně s video záznamem. Každý referenční slajd (obrázek z PDF) i hledaný slajd (obrázek z videa) je popsaný pomocí distribuce vizuálních slov (viz obrázek 7.). Nejprve se popíší lokální příznaky ve snímcích videa a ty se následně kvantizují pomocí slovníku vizuálních slov. Pomocí inverzního indexu jsou pak slajdy z videa vyhledány v referenční sadě. Nalezení kandidáti se transformují pomocí homografie a provede se validace přiřazení srovnáním jasových složek. Výsledkem je jeden nejlepší (nebo žádný) kandidát reprezentující daný slajd na video snímku (podle zadaného prahu minimální shody). Výstupem celého procesu jsou časové razítka ke každému slajdu v knihovně obrázků.



Obr. 7. Extrakce vizuálních slov ze snímku slajdu.

Jak už bylo zmíněno v úvodu, pro převod jednotlivých obrázků do textu používáme standardní nástroje. V případě anglického jazyka volíme nástroj PDF2Text, který vyextrahuje text přímo z PDF a v případě českého jazyka volíme z důvodů diakritických znamének OCR nástroj (například open source Tesseract[1]). Získaný text před indexování očistíme od slov obsahující nealfanumerické znaky.

5 Indexace informace

Pro indexaci a vyhledávání v extrahovaných informacích [3] jsme zvolili open source nástroj Lucene[2] implementovaný v Javě. Z Lucene využíváme takzvaná speciální pole pro:

- Přepis řeči - graf s pevným časováním.
- Slajdy - uložení informace o čísle slajdu včetně textového přepisu slajdu.

Dále používáme standardní textová pole pro:

- Název přednášky.
- Přednášející.
- Abstrakt.
- Klíčová slova.

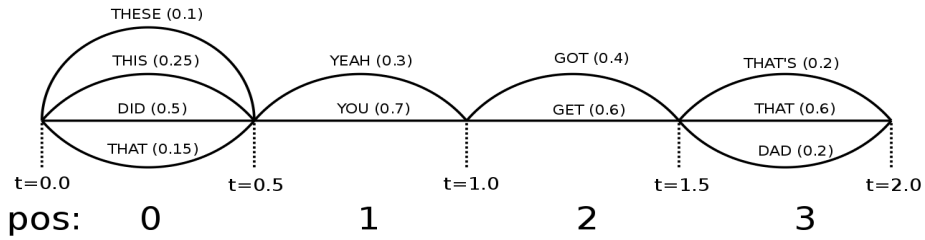
Vyhledáváme ve všech polích a kombinujeme výsledky pomocí TF-IDF váhování termů.

5.1 Indexace přepisu řeči

V sekci 4.1 jsme si popsali, jak vypadá výstup z rozpoznávače řeči - text nebo graf, kde každé slovo má přiřazen čas začátku a konce, a také v případě grafu posteriorní

1 http://code.google.com/p/tesseract-ocr/
2 http://lucene.apache.org/

pravděpodobnost. Nyní potřebujeme převést tuto reprezentaci do formy, kterou je schopen zpracovat Lucene. Taková forma se nazývá confusion network [5]. Jedná se o graf, kde paralelní hypotézy slov končí vždy v jednom uzlu a z něho se rozbíhají následující slova. V obecné confusion network je obvykle mezi slovy i "přeskakovací" přechod (null přechod) a jednotlivé uzly mají nerovnoměrné časování. Pro indexaci využíváme speciální typ confusion network, kde je časování uzlů pevně vzorkováno po půl vteřině (obrázek 8). Přeskakovací přechody neindexujeme.



Obr. 8. Příklad confusion network s pevným časováním (0,5s). řádek POS reprezentuje sloty (pozice) v indexu. Čísla v závorkách jsou posteriorní pravděpodobnosti slov.

V Lucene pak jednotlivé sloty ve speciálním poli obsahují paralelní hypotézy slov pro daný časový úsek (například *t=1,0..1,5*). Pro uchování informace o posteriorní pravděpodobnosti slouží proměnná *payload*, která je přiřazena ke každému slovu. Lucene pak při vyhledávání zohledňuje pravděpodobnost jednotlivých výskytů ve výsledcích dotazu. Při vyhledávání jsou čísla slotů nalezených výskytů vydělena dvěma. Tak získáme časovou informaci o výskytu hledaného slova v řeči ve vteřinách.
Při vyhledávání více slovních frází se sousední slova mohou nacházet uvnitř jednoho časového slotu nebo mohou být až tři sloty vzdálená. To je z důvodů, aby byla fráze nalezena i v případě, že se řečník uprostřed fráze na chvíli odmlčí.

5.2 Indexace slajdů

Pro indexování slajdů využíváme také přiřazení slov do slotů. Informaci o čísle slajdu (tedy o čase výskytu) uchováváme jako číslo slotu krát 1000. Slova z prvního slajdu jsou přiřazena do slotů 1, 2, 3, ..., slova z druhého slajdu jsou přiřazena do slotů 1001, 1002, 1003, atd. Při vyhledávání jsou čísla slotů nalezených výsledku vydělena 1000 a celočíselný výsledek udává číslo slajdu, ve kterém se nalézá výskyt hledaného slova. V prezentační vrstvě pak převedeme číslo slajdu na čas, který jsme získali automatickou synchronizací slajdů s videozáznamem.

6 Prezentace informace

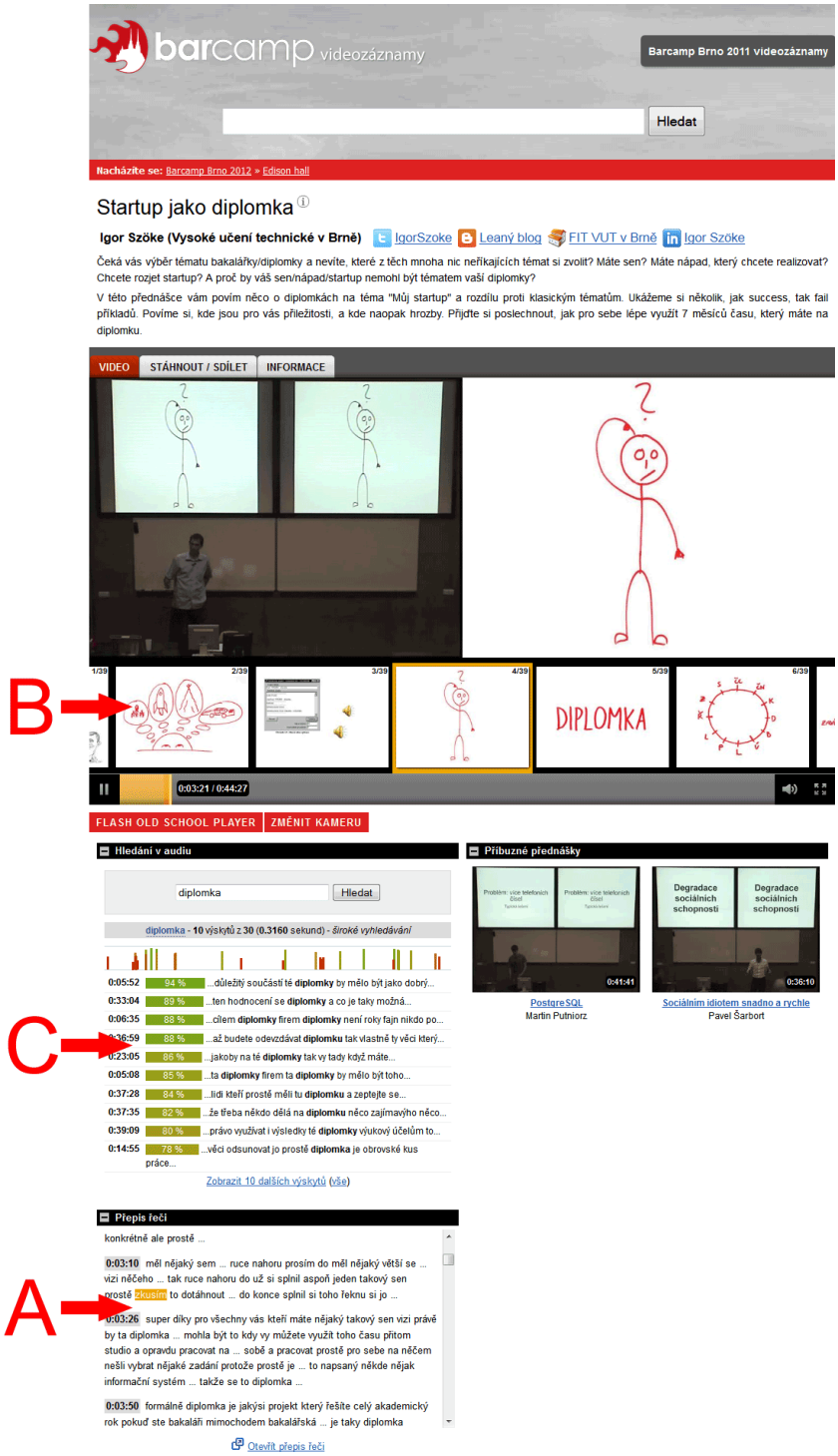
Prezentační vrstva je z pohledu uživatele tím nejdůležitějším prvkem celé služby. Podrobný popis implementace přesahuje rozsah tohoto článku, přehrávač audiovizuálních záznamů si tedy popíšeme jen stručně.
S postupným rozšiřováním HTML5 a nedostupností Flash pluginu na některých zařízeních (např. iPad) vznikla potřeba přehrávače, který by pracoval s HTML5 videem. Pro možnost přehrání videa na zařízeních nepodporujících HTML5 bylo nutné zachovat i nadále Flash

video. Přehrávač je napsán v JavaScriptu a je postaven nad API SuperLectures, skrze které lze získat v XML formátu kompletní informace pro jakoukoliv přednášku (kategorie, název, autoři, abstrakt, odkazy na řečníky, přepis řeči, slajdy, výsledky vyhledávání, atd.). API SuperLectures zároveň umožňuje vyhledávání v datech. Dotaz je předán v URL, odpověď přijde v XML. Nad tímto API vznikly i další přehrávače přednášek, které byly například řešeny v rámci diplomových prací.

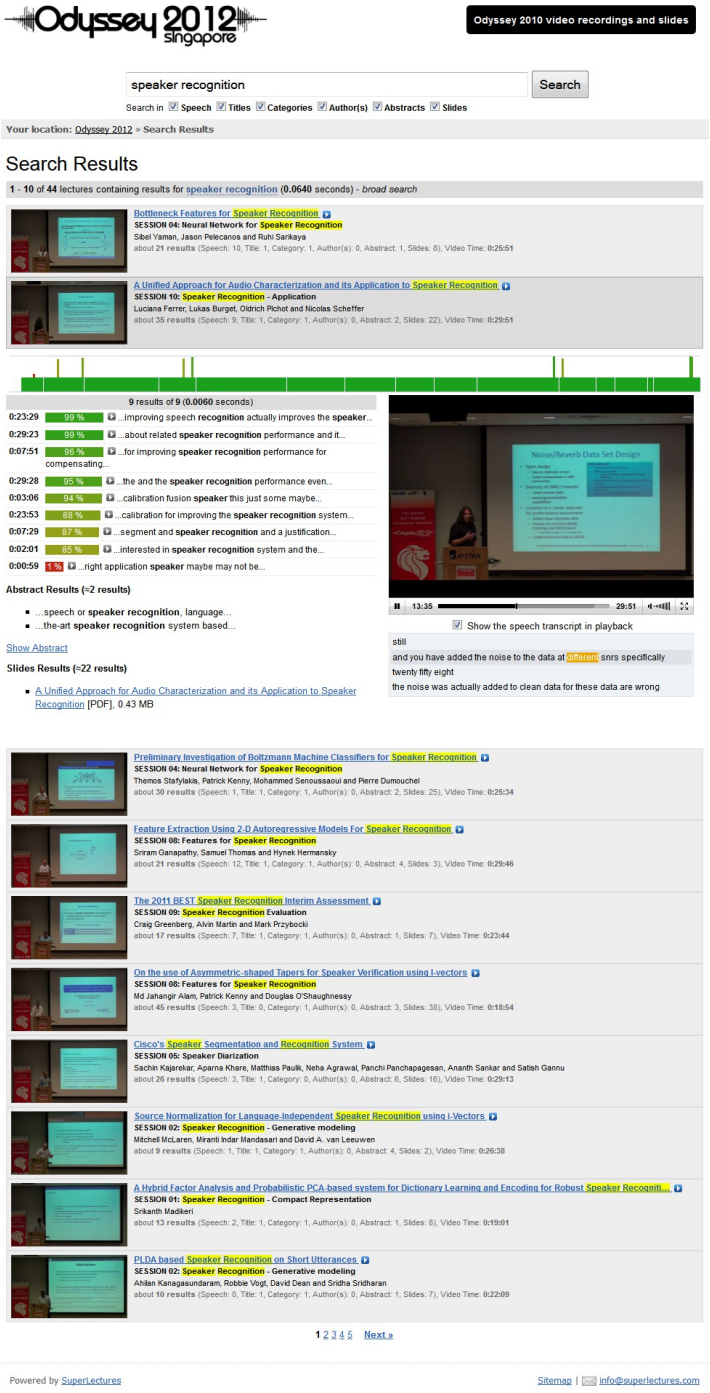
V primárním zobrazení se v přehrávači zobrazuje video záznam, který je případně doplněn o slajdy (obrázek 9a). Jednotlivé komponenty v přehrávači jsou navzájem synchronizovány. Z našeho pohledu je důležitá komponenta, zobrazující automaticky vygenerovaný textový přepis řeči (obrázek 9b). Jedná se o sekvenci nejpravděpodobnějších slov, která jsou získána z rozpoznávače. Pro přehlednost je přepis řeči rozdělen do segmentů. Ty jsou označeny časovými body a kliknutím na ně dojde k přesunu na odpovídající místo videozáznamu. Pro usnadnění orientace je v přepisu řeči zvýrazňováno aktuální slovo.

Komponenta synchronizovaných slajdů umožňuje připojit k přednášce obrázky slajdů a ty synchronizovat s videozáznamem. Mnohdy díky nízkému rozlišení videa či malým fontům je problém přečíst obsah slajdů z videozáznamu. Díky vygenerování náhledů přímo z PDF mohou být slajdy vždy čitelné. Důležitým prvkem této komponenty je časově synchronizovaný seznam prezentovaných slajdů, který usnadňuje navigaci ve videozáznamu. Tedy vybráním určitého slajdu dojde k přesunu na místo, kdy byl prezentován a opačně.

Propojení s vyhledávacím serverem je realizováno přes komponentu vyhledávání (obrázek 9c). Je možné omezit modality, ve kterých se vyhledává (takže uživatel může hledat například jen v audiu nebo jen ve slajdech). Výsledky jsou posléze zobrazeny v tabulce, ve které jsou setříděny podle skóre a sekundárně podle času. Kliknutím na výsledek dojde ke spuštění videozáznamu od patřičného místa. Jak je uživateli reprezentována funkce globálního vyhledávání je zobrazeno na obrázku 10.



Obr. 9. Stránka se záznamem přednášky. a) automatický přepis řeči do textu, b) synchronizované náhledy slajdů, c) výsledky vyhledávání v řeči.



Obr. 10. Stránka se výsledky globálního vyhledávání (vyhledávání ve všech záznamech z konference).

7 Nasazení a zpětná vazba

Vzhledem k tomu, že služba SuperLectures.com již překročila hranici "laboratorního experimentu" a snažíme se ji dostat do komerční praxe formou start-upu, zajímají nás nejen laboratorní data (úspěšnost, rychlost, ...), ale také chování uživatelů a odezva trhu.

7.1 Laboratorní výsledky

Z laboratorních výsledků prezentujeme úspěšnost rozpoznávání řeči. Pro český jazyk jsme provedli analýzu přesnosti přepisu řeči na záznamech přednášek z bakalářského studia na Fakultě informačních technologií VUT v Brně (tabulka 1). Použitou metrikou je míra chybně rozpoznaných slov (WER - word error rate). Testovací data jsou rozdělena na tři části. První část *Akustická a jazyková adaptace* ukazuje přesnost systému při supervised akustické adaptaci na řečníka (systém "slyšel" konkrétního přednášejícího během adaptace) a také jazykové adaptaci (byla použita skripta pro daný předmět). Druhá část ukazuje vliv *pouze jazykové adaptace* a třetí část ukazuje úspěšnost *rozpoznávání neviděných dat* - předměty z magisterského studia - (jak akusticky - řečník) tak jazykově (předmět).

	Data	WER %
Akustická a jazyková adaptace	8 x 5min	25,6
Jazyková adaptace	8 x 5min	28,3
Neviděná data (MGR)	8 x 5min	30,1
Celkem	**2h**	**28,0**

Tab. 1. Výsledky úspěšnosti převodu řeči do textu pro český jazyk. WER - word error rate - počet chybně rozpoznaných slov.

Pro anglické záznamy (tabulka 2) jsme na datech z konference Odyssey2010 dosáhli chybovosti 26,9%. Nutno dodat, že testovací záznamy jsou tvořeny z větší části prezentacemi nerodilých mluvčích se silným akcentem. Oproti tomu, chybovost záznamů rodilých mluvčích z konference InterSpeech2010 byla pod 20%.

	Data	WER %
Odyssey2010 - jazyková adaptace	48 x 2min	26,9
InterSpeech2010 - bez adaptace	93min	18,7

Tab. 2. Výsledky úspěšnosti převodu řeči do textu pro anglický jazyk. WER - word error rate - počet chybně rozpoznaných slov.

Podle zpětné vazby od uživatelů, je chybovost přepisu 30% a více již neakceptovatelná. Takto prezentované automatické titulky uživatele ruší při sledování záznamu. Chybovost pod 20% je na druhou stranu již přijímána bez výhrad. Zde se právě ukazuje výhoda indexování grafů s paralelními hypotézami kdy přesto, že se hledané slovo nedostane do textového přepisu (na první pohled není vidět), velmi často se vyskytuje jako paralelní hypotéza a lze ho tak dohledat.

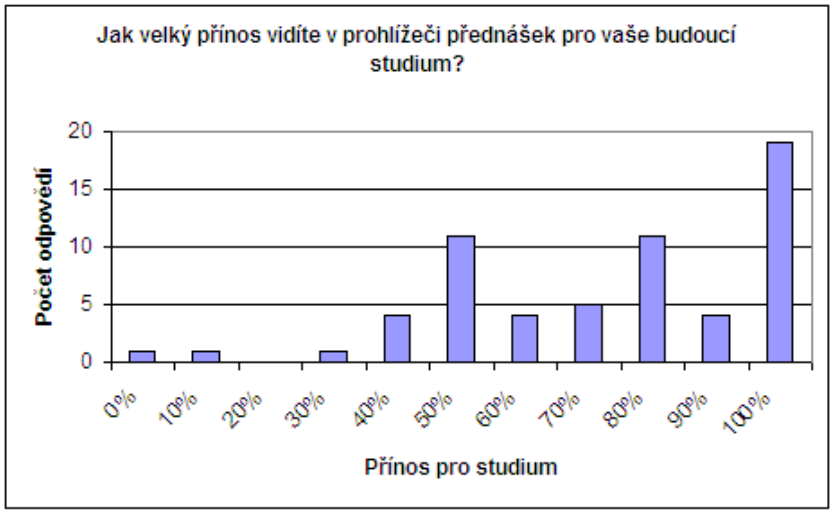
7.2 Praktické zhodnocení - přednášky bakalářského studia

Na počátku května 2010 byla spuštěna testovací verze prohlížeče přednášek určená pro studenty Fakulty informačních technologií Vysokého učení technického v Brně. Ta obsahuje videozáznamy přednášek z povinných kurzů bakalářského studijního programu tak, jak jimi typický student během 2007-2010 prošel. Tato verze mohla pomoci studentům při přípravě na zkoušky, zejména pak na státní bakalářskou zkoušku.
Pro zajímavost, ze statistik použití v období červen až srpen 2010 vyplývá, že systém vyzkoušelo více než 500 studentů (z celkových ~2600). Přibližně 150 studentů pak použilo systém více než 5krát. V žebříčku nejčastěji hledaných výrazů se na prvních místech jednoznačně umístily termíny `"státnic"/"státnice"`. Ty v mnoha případech navedly uživatele na místa ve videozáznamech, kdy se přednášející zmiňoval, co bude po studentech vyžadováno u státní závěrečné zkoušky. Mezi další často hledané výrazy pak patřily termíny jako `"Fourierova transformace"`, `"konvoluce"`, `"směrování"`, `"tranzistor"`, `"multiplexor"`, `"dioda"`, `"relace"` aj.
Studenti po svém přihlášeni do systému měli možnost vyplnit dotazník a napsat své podněty k prohlížeči přednášek. Celkově lze ze získaných odpovědí, kterých bylo více než 60, vyčíst pozitivní hodnocení systému. Na dotaz, zdali by student byl ochotný se podílet na zlepšování systému například opravami přepisů, doplňováním slovníků apod. přišlo 85% kladných odpovědí. V tabulce 3 je pak zobrazeno průměrné hodnocení některých vlastností prohlížeče. Obrázek 12 ukazuje, jak studenti hodnotili přínos vyhledávání z záznamech přednášek pro své budoucí studium. Pro zajímavost jsme v tabulce 4 vypsali patnáct nejčastěji vyhledávaných klíčových slov a obrázek 11 ukazuje návštěvnost služby během státnic.

	Průměrná známka
Přehlednost, orientace v prohlížeči:	1,34
Vyhledávání v řeči:	1,52
Úroveň automatického přepisu:	2,11
Kvalita videa a audia:	1,97
Grafická stránka prohlížeče:	1,39

Tab. 3 Výsledky dotazníku - hodnocení prohlížeče přednášek (1 nejlepší, 2, 3, 4, 5 nejhorší)

Obr. 11 Návštěvnost záznamů přednášek během bakalářských státnic v ak. roce 2009/2010.



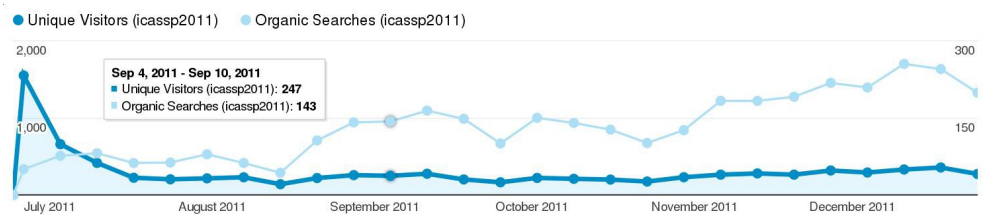
Obr. 12 Výsledky dotazníku – odhad míry přínosu prohlížeče přednášek pro budoucí studium

6. – 8. 2010	#	2010/2011	#
Státnic	261	Klasifikace	25
Státnice	170	Lemma	22
Konvoluce	69	Státnice	22
Směrování	58	Fourierova transformace	18
Tranzistor	52	Pumping	17
Multiplexor	50	Konvoluce	17
Dioda	45	Strom	16
Relace	44	Automat	15
Fourierova transformace	41	Pumping lemma	13
Minimalizace	37	Gramatika	12
Rasterizace	34	Gymnázium	11
Složitost	30	Rotace	11
Regulární jazyk	29	Tabulka	11
Unipolární	28	Z buffer	11
Státnicích	28	Normalizace	9

Tab. 4 Četnosti vyhledávaných klíčových slov v období státnic (první a druhý sloupec) a během následujícího akademického roku (třetí a čtvrtý sloupec).

7.3 Praktické zhodnocení - záznamy z mezinárodní konference IEEE ICASSP 2011

V červnu 2011 jsme zpracovali přes 250 orálních prezentací (celkem přes 80 hodin AV záznamů) z International Conference on Acoustics, Speech, and Signal Processing, kterou navštívilo přes 2000 účastníků. Pro zajímavost jsou na obrázku 13 vyobrazeny grafy návštěvnosti záznamů mezi červnem a prosincem roku 2011. Za pozornost stojí rostoucí trend návštěv z vyhledávače Google. Ten podporuje hypotézu, že jsou díky převodu řeči to textu záznamy snadno dohledatelné. Během stejného období bylo přes 6700 vyhledávací dotazů z 1450 sezení. Průměrný vyhledávací čas byl 0,016ms.



Obr. 13 Návštěvnost AV záznamů z konference ICASSP 2011 v období červen až prosinec 2011. Tmavě modrý graf jsou unikátní návštěvníci, světle modrý graf je počet návštěv z vyhledávače Google.

8 Závěr

Vývoj systému pro přístup ke znalostem v audiovizuálních přednáškách je typickým příkladem přenosu špičkových vědeckých poznatků do tvrdé reality. Teoretické, algoritmické a experimentální práce skupiny BUT Speech@FIT jsou ve světě uznávány, ale ověřili jsme si, že pro provoz a komerční úspěšnost je neméně důležité poznat požadavky uživatelů a poskytovatelů audiovizuálních dat, reagovat na ně, a přijít s uceleným technickým řešením, které přinese úspory času a finanční zhodnocení. Věříme, že služby nabízené ReplayWell si najdou širokou klientelu, a že poznatky a data z jejich masového nasazení nám naopak pomohou zlepšovat funkčnost jednotlivých subsystémů (přesnost rozpoznávání řeči i slajdů jsou závislé na množství trénovacích dat). A o tom by podle nás „transfer technologií“ mezi akademickou a komerční sférou měl být.

Poděkování:
Tato práce byla podporována výzkumným programem MŠMT 0021630528, Techologickou Agenturou České Republiky grant číslo TA01011328 a IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070. I. Szöke byl podporován Grantovou Agenturou České Republiky post-doktorský projekt č. GP202/12/P567.

Literatura

1. Bisani, M., Ney, H.,: Joint-Sequence Models for Grapheme-to-Phoneme Conversion. *Speech Communication* (2008), doi: 10.1016/j.specom.2008.01.002.
2. Chen, Y.N. , Huang, Y., Kong, S.Y., Lee L.S.: Automatic Key Term Extraction from Spoken Course Lectures Using Branching Entropy and Prosodic/Semantic Features. In

Proc. of IEEE Spoken Language Technology Workshop 2010, Berkeley, California, U.S.A., 2010, ISBN 978-1-4244-7902-3.

3. Fapšo, M., Smrž, P., Schwarz, P., Szőke, I., Schwarz, M., Černocký, J., Karafiát, M., Burget, L.: Information Retrieval from Spoken Documents, *In Proc. of the Seventh International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2006)*, Mexico City, MX, Springer, 2006, p. 410-416, ISBN 3-540-32205-1.
4. Hain, T., Burget, L., Dines, J., Garner, P., N., Grézl, F., El, H., A., Huijbregts, M., Karafiát, M., Lincoln, M., Wan, V.: Transcribing Meetings with the AMIDA System. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 2, 2012, US, p. 486-498, ISSN 1558-7916.
5. Mangu, L., Brill, E., Stolcke, A.,: Finding Consensus among words: lattice-based word error minimisation. Computer Speech and Language (2000) 373-400.
6. Szőke, I., Schwarz, P., Burget, L., Fapšo, M., Karafiát, M., Černocký, J., Matějka, P.: Comparison of Keyword Spotting Approaches for Informal Continuous Speech. *In Proc. of the Nineth European Conference on Speech Communication and Technology,* Lisabon, PT, 2005, p. 633-636, ISSN 1018-4074.

Annotation:
The amount of audiovisual data in growing. Part of the data as lecture or conference recordings contain important information. However this information is hidden and unreachable for standard web crawlers as Google. This paper deals with a system, which makes the information available for standard text based indexers and searchers. It is done by conversion of speech and video into text. Description of the audiovisual indexing and search system is provided in the first part of this paper. We briefly describe the speech-to-text and slide synchronization components. Next, the description of an indexing engine is given. The engine is capable to index not only text but also timing and probability of recognized speech. The second part is aimed at practical issues like user interface and customer feedback.