

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Jan Milota

Vývoj hlasově ovládaných webových her pomocí CloudASR

Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: Mgr. Ing. Filip Jurčiček, Ph.D.

Studijní program: Informatika

Studijní obor: Obecná Informatika

Praha 2015

Děkuji panu doktorovi Jurčíčkovi za profesionální vedení této práce a za jeho neutuchající víru v mou schopnost samostatného vývoje.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V dne

Podpis autora

Název práce: Vývoj hlasově ovládaných webových her pomocí CloudASR

Autor: Jan Milota

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: Mgr. Ing. Filip Jurčíček Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: Cílem práce je navrhnout a vyvinout software pro výuku jazyků hrou za použití webových technologií a čerstvě vznikající CloudASR knihovny. Běžný uživatel provozuje interakci se svým prohlížečem skoro výhradně prostřednictvím myši a klávesnice. Díky softwaru, který tato práce reprezentuje, má nyní uživatel příležitost zabřednout do někdy ne úplně populární výuky jazyka i za pomoci svého hlasu. Což nabízí zmíněné výuce netušené možnosti, obzvláště stran uživatelské interaktivity. Důraz byl kladen na uživatelskou přívětivost, grafickou fidelitu a na kompetitivní aspekt výuky, využívajíc Facebookovou integraci a bodové hodnotící žebříčky.

Klíčová slova: automatické rozpoznávání řeči, ASR, hry, web, HTML5, Javascript
Title: Development of speech enabled web games using CloudASR

Author: Jan Milota

Department: Institute of Formal and Applied Linguistics

Supervisor: Mgr. Ing. Filip Jurčíček Ph.D., Institute of Formal and Applied Linguistics

Abstract: The main goal of this thesis is to design and implement a piece of software for playful language learning, using web technologies and the fresh CloudASR library. A common user interacts with their web browser almost exclusively using a mouse and keyboard. Thanks to the software this thesis represents the user has an opportunity to delve into sometimes unpopular language learning process using his natural voice. This fact presents new and exciting possibilities, mainly regarding user interactivity. A lot of stress has been put to user friendliness, graphical fidelity and to the competitive aspect of language education, exploiting Facebook integration and point-scoring leader boards.

Keywords: automatic speech recognition, ASR, games, web, HTML5, Javascript

Obsah

Úvod	2
1 Automatic Speech Recognition	4
1.1 Co je to ASR	4
1.2 Postup rozpoznávání	4
1.3 Reprezentace a zpracování signálu	5
1.3.1 Fourierovská analýza	5
1.3.2 Cepstrální analýza	6
1.4 Zdroje omezení mluvené komunikace	6
1.5 Modelování systému	7
1.5.1 Maximum a Posteriori formulace	7
1.5.2 Jazykové modely	8
1.5.3 Akustické modely	9
1.5.4 Lexikální modely	10
2 e-Learning jako zdroj informací	12
2.1 Kapitola	12
2.2 Kapitola	12
3 Použité technologie	13
3.1 Kapitola	13
3.2 Kapitola	13
4 Implementace	14
4.1 Kapitola	14
4.2 Kapitola	14
Závěr	15
Seznam použité literatury	16
Seznam tabulek	19
Seznam použitých zkratk	20
Přílohy	21

Úvod

Základním komunikačním kanálem mezilidské interakce je mluvená řeč. Ačkoli trend v poslední době spíše k separaci této komunikace virtuálními médii, je mluvené slovo i nadále nenahraditelnou součástí socializace a výměny informací. Proč tedy nepřenést schopnost porozumět lidské mluvě i na stroje, jakožto právě na ona separující virtuální média? Současná technologie je k tomu jistě dostačující, výpočetní výkon lidstva roste po zběsilé křivce a technologické know-how po ještě zběsilejší.

Aby tento cíl mohl být dosažen, je kritické mít dobrý ASR (Automatic Speech Recognition) systém. Těchto se pohybuje v éteru mnoho, volně dosažitelných, rozličné kvality a spolehlivosti. Pro specifické využití v této práci byl zvolen systém **CloudASR**¹, jehož autorem je kolega Ondřej Klejch.

TrogASR (Translating Online Game using Automatic Speech Recognition) je, jak možná název napovídá, online hra, zaměřující se na překlad slov a frází z jazyka do jazyka za využití automatizovaného rozpoznávání řeči.

Cílem této práce je návrh a implementace graficky přitažlivé hry pro webové prohlížeče, která poskytne uživatelům další, neotřelou, možnost, jak si vyzkoušet a zdokonalit své jazykové schopnosti. Tato hra pak bude sloužit jako kompetitivitou hnaný motor pro sběr verifikovaných dat pro použitou **CloudASR** knihovnu.

Velký důraz je kladen na grafickou stránku věci; uživatelsky neatraktivní hra nemá valnou naději na šíření mezi uživateli samotnými samovolně. Navíc se na ošklivou aplikaci nikdo nebude dívat po delší časový úsek, než je jedno sezení, a už vůbec nikdo se k takovéto aplikaci nebude vracet. Proto bylo veškeré stylování a malování a navrhování layoutů a přechodů prováděno s výraznou pílí.

Dále je důraz kladen i na zmíněnou kompetitivitu. Každý má touhu vidět své jméno někde vysoko na žebříčku, zanecháváje své přátele na tomto žebříčku daleko pod sebou. Proto byla zapojena i integrace s Facebookovým API, jež přináší kýženého výsledku poměrně nenásilnou formou. Na „Like“ a „Login“ tlačítka jsou uživatelé zvyklí a nebojí se jej extenzivně používat.

Následující text popisuje a dokumentuje vývoj této aplikace a myšlenkové pochody při něm stojící. První kapitola analyzuje využití ASR v e-learningu. Druhá kapitola seznámí čtenáře s využitými technologiemi a frameworky pro vývoj aplikace **trogASR**. Kapitola třetí nabídne čtenáři vhled do řešení a implementace aplikace **trogASR**, do použitých návrhových vzorů a do problematiky s nimi

¹<https://github.com/UFAL-DSG/cloud-asr/>

spojené. A konečně, kapitola čtvrtá shrne celou práci.

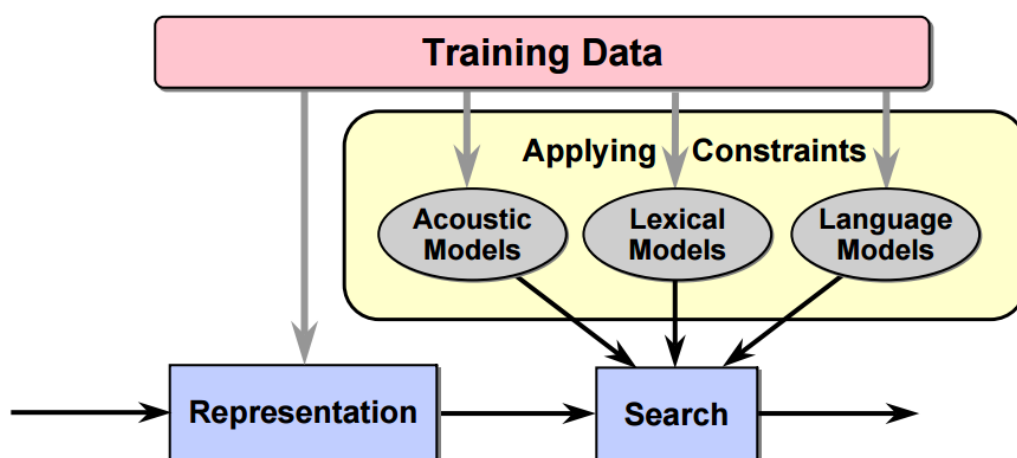
1. Automatic Speech Recognition

1.1 Co je to ASR

ASR (Automatic Speech Recognition) je teoretický model, založený na pravděpodobnostních vlastnostech akustických pozorování, který formalizuje převod mluveného slova do textu. Přeneseně se takto označují i konkrétní systémy a implementace, které tohoto modelu využívají.

Ultimátním cílem je získat systém, který bude v reálném čase převádět přirozenou lidskou řeč na text se stoprocentní úspěšností pro všechna slova, nezávisle na zkreslení vstupních zařízení, okolním ruchu, nebo přízvuku mluvčích. Již v minulosti ASR systémy dosahovaly i kolem devadesátiprocentní úspěšnosti [3]. Nyní lze nalézt systémy s úspěšností i větší, hlavně však systémy optimalizovanější a výkonnější.

1.2 Postup rozpoznávání



Obrázek 1.1: Hlavní součásti ASR systému [4]

Obrázek 1.1 ukazuje, že samotný návrh systému je poměrně přímočarý.

Zásadními milníky systému jsou:

- Jak zpracovat signál?
- Jak se popasovat s omezeními?
- Jak nalézt optimální výstup?

1.3 Reprezentace a zpracování signálu

Pro reprezentaci a zpracování signálu se běžně využívají dva druhy analýzy – analýza pomocí *discrete-time* Fourierovy transformace (DTFT) a Cepstrální analýza (CA).

1.3.1 Fourierovská analýza

Discrete-time Fourierova transformace je vztah:

$$X(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} x[n]e^{-j\omega n} \quad (1.1)$$

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n} d\omega \quad (1.2)$$

jemuž pro konvergenci postačuje podmínka:

$$\sum_{n=-\infty}^{+\infty} |x[n]| < +\infty \quad (1.3)$$

Ačkoli je $x[n]$ diskrétní, $X(e^{j\omega})$ je spojitá, 2π -periodická a pro její dualitu platí:

$$y[n] = x[n] * h[n] \quad (1.4)$$

$$Y(e^{j\omega}) = X(e^{j\omega})H(e^{j\omega}) \quad (1.5)$$

a

$$y[n] = x[n]w[n] \quad (1.6)$$

$$Y(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(e^{j\theta})X(e^{j(\omega-\theta)})d\theta \quad (1.7)$$

V důsledku můžeme zavést *short-time* fourierovskou analýzu (STFA – Short-Time Fourier Analysis) jako:

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{+\infty} w[n-m]x[m]e^{-j\omega m} \quad (1.8)$$

a v případě, že $X(e^{j\omega})$ reprezentuje DTFT signálu, který pokračuje i mimo okno brané v potaz (nebo je tam konstantní nula), můžeme psát, že:

$$X_n(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(e^{j\theta})e^{j\theta n} X(e^{j(\omega+\theta)})d\theta \quad (1.9)$$

1.3.2 Cepstrální analýza

Původně CA vznikla jako nástroj analýzy časového rozvoje seismických signálů vzniklých zemětřeseními, či explozemi. Měla pomoci nalézt především hloubku generátoru seismických signálů analýzou odezvy těchto signálů [6].

CA využívá předpokladu, že signál je výstup lineárního časově invariantního (LTI – Linear Time-Invariant) systému; tedy že je to konvoluce vstupu a impulsové odezvy. Pokud se chceme zabývat charakterizací tohoto modelu, musíme projít procesem dekonvoluce. A CA je právě postup, využívaný pro takovouto LTI dekonvoluci [5].

Vezmeme-li v potaz pozorování, že:

$$x[n] = x_1[n] * x_2[n] \iff X(z) = X_1(z)X_2(z) \quad (1.10)$$

a vyjádříme-li komplexní logaritmus $X(z)$ jako:

$$\hat{X}(z) = \log X(z) = \log X_1(z) + \log X_2(z) \quad (1.11)$$

pak pokud je tento logaritmus jednoznačný a $\hat{X}(z)$ je validní z-transformace, jsou nové konvolované signály aditivní v této nové cepstrální doméně:

$$\hat{x}(n) = \hat{x}_1(n) + \hat{x}_2(n) \quad (1.12)$$

Pokud se navíc omezíme na jednotkovou kružnici $z = e^{j\omega}$, pak cepstrální transformací označujeme:

$$\hat{X}(e^{j\omega}) = \log |X(e^{j\omega})| + j \arg X(e^{j\omega}) \quad (1.13)$$

1.4 Zdroje omezení mluvené komunikace

Při předávání informace mluvenou řečí může běžně docházet ke zkreslení řečené informace díky několika nepříjemným faktorům, které mohou (a budou) ASR systému působit nepříjemnosti. Těmto omezením se nelze vyhnout, proto jsou ASR systémy nuceny brát v potaz rozličné nepřesnosti a snažit se tyto eliminovat.

Práci už tak nepříjemnou nezlehčuje ani fakt, že omezení nejsou konstantně daná; lze je pouze taxonomizovat a obcházet.

Patří sem například:

- Akustická omezení – lidský hlasový aparát trpí, jako každý jiný komplexní systém, také svými neduhami

- Fonetická omezení – „kolemjdoucí paní“ - „kolem jdou cíp a ní“
- Fonologická omezení – například splynutí „s“ a „š“ do jednoho dlouhého fonému - „lezeš shora“
- Fonotaktická omezení – kupříkladu zjednodušení konsonantických shluků u menších dětí („uličnice“ - „ulitite“), nebo nesmyslná, rádoby přejatá slova („sympathetic“ - „sympatetický“)
- Syntaktická omezení – „jedu na výlet do Českých Budějovic“ - „Českých výlet na Budějovic jedu do“
- Sémantická omezení – „zapal sirku“ - „zapal si ruku“
- Lexikální omezení – „šupláneček“ ani „niplavý“ nejsou česká slova

a mnohá další.

1.5 Modelování systému

V ASR procesu se stroj pokouší nalézt co možná nejlepší shodu získaného signálu (reprezentace přirozené řeči) a posloupnosti písmen (slov, vět). Abychom něčeho podobného mohli dosáhnout, musíme navrhnout dostatečně komplexní a efektivní model, který nám (a našim strojům) dá do ruky nástroj pro formulaci problému, jeho porozumění a metodologii potřebnou k jeho rozlousknutí.

1.5.1 Maximum a Posteriori formulace

Tato formulace se „pokouší nalézt nejpravděpodobnější sekvenci slov W^* při daném akustickém vstupu A “ [2]. MAP přístup [1] lze typicky vyjádřit jako:

$$W^* = \arg \max_W P(W|A) \quad (1.14)$$

Věta 1. *Bayesova věta*

Pro náhodné nezávislé jevy X a Y platí:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \quad (1.15)$$

Za použití 1.15 můžeme upravit pravou stranu rovnosti 1.14, čímž dostaneme vztah:

$$W^* = \arg \max_W \frac{P(A|W)P(W)}{P(A)} \quad (1.16)$$

kde $P(A|W)$ je nazývána charakteristikou akustického modelu (AM – Acoustic Model) a $P(W)$ charakteristikou jazykového modelu (LM – Language Model). Jelikož $P(A)$ je konstantní vůči této maximalizaci, můžeme se jí hladce zbavit a psát:

$$W^* = \arg \max_W P(A|W)P(W) \quad (1.17)$$

1.5.2 Jazykové modely

Modelování jazyka poskytuje nástroj, jak odlišit slova a fráze, které zní podobně, či stejně (homografy, homofony, homonyma). Kupříkladu v americké angličtině jsou věty „recognize speech“ a „wreck a nice beach“ sémanticky diametrálně odlišné, foneticky jsou to však věty téměř homofonní.

Jelikož vyhledávání je v drtivé většině případů pouze jednosměrné, můžeme pravděpodobnost posloupnosti slov $W = w_1, \dots, w_n = w_i^n$ vyjádřit pomocí řetězového pravidla a napsat jako:

$$P(w_i^n) = \prod_{i=1}^n P(w_i|w_1^{i-1}) \quad (1.18)$$

Protože sekvence W může mít jednoduše příliš mnoho prvků a navíc pravděpodobnost $P(w_i)$ nemusí nutně záviset na úplně celé historii $h_i = w_1^{i-1}$, slučujeme h_i do tříd ekvivalence $\Phi(h_i)$, čímž dospíváme k:

$$P(w_i^n) \approx \prod_{i=1}^n P(w_i|\Phi(h_i)) \quad (1.19)$$

Jak možno nahlédnout, pro kvalitní odhad je kritické určit dobré zobrazení Φ . Čím lepší návrh tohoto zobrazení, tím lepší informaci získáme o w_i , využívaje historie $\Phi(h_i)$.

Zde nastupují na scénu n -gram modely [7][8], které v tomto případě zavádějí $n-1$ předchozích slov pro reprezentaci historie, tj. $\Phi(h_i) = w_{i-(n-1)}^{i-1}$, čímž získáme:

$$P(w_i^n) \approx \prod_{i=1}^n P(w_i|w_{i-(n-1)}^{i-1}) \quad (1.20)$$

Například uvážíme-li *bigram* (n -gram, $n = 2$), pravděpodobnost věty „České Budějovice jsou krásné město“, tj. posloupnosti slov:

$$W = Ceske, Budejovice, jsou, krasne, mesto$$

lze vyjádřit jako:

$$\begin{aligned}
P(W) &= P(Ceske | <start>) P(Budejovice | Ceske) \\
&\quad P(jsou | Budejovice) P(krasne | jsou) \\
&\quad P(mesto | krasne) P(<konec> | mesto)
\end{aligned}$$

Pravděpodobnosti $P(w_i | w_{i-(n-1)}^{i-1})$ můžeme dále odečíst z natrénovaných dat vztahem:

$$P(w_i | w_{i-(n-1)}^{i-1}) = \frac{c(w_{i-(n-1)}^i)}{c(w_{i-(n-1)}^{i-1})} \quad (1.21)$$

kde $c(\xi_p^q)$ představuje četnost výskytů posloupnosti ξ_p^q .

Kvůli možné nepřiměřené míře řídkosti dat je nasnadě uvážit možnost, že dělitel $c(w_{i-(n-1)}^{i-1})$ bude nulový, tedy že posloupnost ještě nebyla natrénována.

Pro eliminaci těchto faktorů lze využít například některý z vyhlazovacích algoritmů (Jelinek-Mercer [9], Katz [10], Good-Turing [11]), nebo drobně pozměnit paradigma a přizvat na pomoc neuronové sítě [12].

Buď jak buď, *bigramy* stále zachovávají svou pozici v modelování jazyka — lze je jednoduše zakomponovat do Viterbiho vyhledávání [13] — a *trigramy* (*n-gram*, $n = 3$) pokračují v bytí dominantním modelem.

1.5.3 Akustické modely

Modelováním akustiky můžeme naopak získat nástroj pro zachycení a vypořádání se například s akustickým šumem, se zkreslením vstupního zařízení, či s netradičním přízvukem mluvčího.

Akustický model statisticky reprezentuje zvuky, jenž poskládány dohromady tvoří vyřčené slovo. Každé takovéto statistické reprezentaci může být přiřazena nálepka ji reprezentující (běžně foném). Aby model mohl poskytovat rozumné výsledky, je potřeba jej nejprve natrénovat na nějakém jazykovém korpusu, například pro tento účel lze využít Baum-Welchova trénovacího algoritmu [15].

Skryté Markovovy modely (HMMs – Hidden Markov Models) [14] jsou jednou z možných interpretací akustického modelu (jiné interpretace mohou zahrnovat například neuronové sítě [16], nebo *dynamic time warping* [17]).

Při použití HMMs jeden HMM reprezentuje každý foném, slova vznikají konkatencí menších HMMs, věty na oplátku konkatencí těchto HMMs a tak dále.

Proč zrovna „skrytý“ Markovův model? Tento název vychází z definice „skrytého“ Markovova procesu (HMP – Hidden Markov process), který popíšeme za pomoci kolegů Alibaby a Rychlonožky a jejich uren.

Vezměme myšlenkový experiment, kde v místnosti za zavřenými dveřmi žije Alibaba se třemi svými urnami u_1, u_2, u_3 . Každá urna obsahuje známou množinu kuliček $\{k_{u_i}\}_1^n$, kde platí $(\forall u \in \{u_1, u_2, u_3\})(\forall j \in \{1, \dots, n\})(k_{uj} \in \{c_1, \dots, c_n\})$, tedy každá kulička z každé urny je jedné z barev c . Alibaba bude nyní ďábelským způsobem tahat kuličky z urn, a sice tak, že l -tou kuličku K_l vytáhne náhodně a pouze s ohledem na K_{l-1} . Tomuto procesu se říká Markovův proces. Jelikož Alibaba na běhání nikdy nebyl, dá taženou K_l Rychlonožkovi, který vyběhne ven dveřmi a složí K_l k nohám nezávislého pozorovatele, hned vedle kuličky K_{l-1} do řady.

Pozorovatel sice zná složení kuliček v urnách, ale nemá ani tušení, co se za dveřmi děje — Markovův proces je skrytý — takže může pouze smutně sledovat sekvenci kuliček, rodící se mu pod nohama. Po $1 < m < n$ tazích a Rychlonožkovu sprintech bude mít pozorovatel v zorném poli posloupnost kuliček K_1^m . Problém ale nastává, chce-li pozorovatel určit z jaké urny byla tažena K_m . Ani v případě, že $(\exists i \in \{1, 2, 3\})(K_1^m \subseteq \{k_{u_i}\}_1^m)$ si totiž nemůže být jist.

Nicméně, pozorovatel se může alespoň pokusit odhadnout pravděpodobnost, že K_m byla z urny i .

1.5.4 Lexikální modely

Lexikální modely popisují vztah mezi jednotkami akustickými (fonémy – části slova mluveného, popsané akustickými modely) a lexikálními (lexémy – jazykový odraz reality na vědomí).

Pro ilustraci lexému vezměme slova „ředkvička“ a „ředkvičky“; mají rozdílné gramatické a sémantické významy (singulár a plurál téhož), lexikální význam je ale stejný (malý, načervenalý, jedlý kořen).

Jednou z výzev navrhování ASR systému je popsat tyto vztahy pro jazyk, kde lexikální příznaky nejsou zcela zřejmé. Jak ale [18] uvádí, pro standardní stochastické HMM ASR systémy existuje lexikální model deterministický, tj. existuje bijektivní zobrazení mezi trénovanými HMMs a lexémy.

Pro jazyky obtížné, s ne zcela formálními lexikálními příznaky, lze lexikální model odvodit z některého lexikálního modelu již známého – aplikací známého modelu na nový jazyk a jeho postupnou adaptací.

Zavedeme-li $\Theta = \{\Theta_A, \Theta_L\}$ množinu parametrů akustických a jazykových modelů, kde $\Theta_A = \{\theta_a, \theta_p, \theta_l\}$ jsou parametry akustického modelu, respektive lexikonu, respektive lexikálního modelu, můžeme s využitím 1.14 a 1.15 rozvést:

$$W^* = \arg \max_W P(W|A, \Theta) \quad (1.22)$$

$$= \arg \max_W \frac{P(A|W, \Theta_A)P(W|\Theta_L)}{P(A|\Theta)} \quad (1.23)$$

$$= \arg \max_W P(A|W, \Theta_A)P(W|\Theta_L) \quad (1.24)$$

Hledání nejpravděpodobnější sekvence W^* pro akustický vstup A pak můžeme díky lexikálnímu modelování transformovat na problém hledání nejpravděpodobnější posloupnosti stavů Q^* :

$$Q^* = \arg \max_Q P(Q, A|\Theta) \quad (1.25)$$

$$= \arg \max_Q \prod_{t=1}^T p(a_t|q_t = l^i, \Theta_A)P(q_t = l^i|q_{t-1} = l^i, \Theta) \quad (1.26)$$

$$= \arg \max_Q \sum_{t=1}^T (\log p(a_t|q_t = l^i, \Theta_A) + \log P(q_t = l^i|q_{t-1} = l^i, \Theta)) \quad (1.27)$$

kde T je celkový počet uvažovaných oken, $Q = \{q_1, \dots, q_T\}$ jde přes všechny posloupnosti možných HMM stavů, $q_t \in \{l^1, \dots, l^i, \dots, l^I\}$ a I značí počet lexémů. Obvykle je výraz $\log p(a_t|q_t = l^i, \Theta_A)$ označován jako lokální emisní skóre (LES – Local Emission Score) a výraz $\log P(q_t = l^i|q_{t-1} = l^i, \Theta)$ jako přechodové skóre (TS – Transition Score).

2. e-Learning jako zdroj informaci

2.1 Kapitola

2.2 Kapitola

3. Použité technologie

3.1 Kapitola

3.2 Kapitola

4. Implementace

4.1 Kapitola

4.2 Kapitola

Závěr

Seznam použité literatury

- [1] BAHL, L. R., JELINEK, F., MERCER, R. L.
A maximum likelihood approach to continuous speech recognition.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983
- [2] DEORAS, A., FILIMONOV, D., HARPER, M., JELINEK, F.
Model combination for Speech Recognition using Empirical Bayes Risk minimization.
IEEE Spoken Language Technology Workshop (SLT), 2010
- [3] BAHL, L., BAKER, J., COHEN, P., DIXON, N.R., JELINEK, F., MERCER, R.L., SILVERMAN, H.F.
Preliminary results on the performance of a system for the automatic recognition of continuous speech.
Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '76, 1976
- [4] GLASS, James, ZUE, Victor
6.345 Automatic Speech Recognition.
Massachusetts Institute of Technology: MIT OpenCourseWare, Spring 2003.
License: Creative Commons BY-NC-SA
- [5] TOHKURA, Y.
A Weighted Cepstral Distance Measure for Speech Recognition.
IEEE Trans. ASSP, Vol. ASSP-35, No. 10, 1414-1422, 1987
- [6] BOGERT, B.P., HEALY, M.J.R., TUKEY, J.W.
The frequency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking.
Time Series Analysis, M. Rosenblatt, Ed. New York: Wiley, ch.15, 1963
- [7] BYEONGKYU, Ko, DONGJIN, Choi, CHANG, Choi, JUNHO, Choi, PANKOO, Kim
Document Classification through Building Specified N-Gram.
Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), Sixth International Conference 2012
- [8] MASATAKI, H., SAGISAKA, Y., HISAKI, K., KAWAHARA, T.
Task adaptation using MAP estimation in N-gram language modeling.

Acoustics, Speech, and Signal Processing (vol.2), IEEE International Conference 1997

- [9] JELINEK, F., MERCER, R.
Interpolated estimation of Markov source parameters from sparse data.
Proceedings of Workshop on Pattern Recognition in Practice, pg.381-397, 1980
- [10] KATZ, S.
Estimation of probabilities from sparse data for the language model component of a speech recognizer.
Acoustics, Speech and Signal Processing, IEEE Transactions, 1987
- [11] CHURCH, K., GALE, W.
A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams.
Computer Speech & Language, 1991
- [12] CHIEN-LIN, Huang, HORI, C., KASHIOKA, H.
Semantic inference based on neural probabilistic language modeling for speech indexing.
Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference 2013
- [13] GOBLIRSCH, D.M.
Viterbi beam search with layered bigrams.
Spoken Language, Fourth International Conference, 1996
- [14] PORITZ, A.
Hidden Markov models: a guided tour.
Acoustics, Speech, and Signal Processing, International Conference, 1988
- [15] HAN, Shu, HETHERINGTON, I.L., GLASS, J.
Baum-Welch training for segment-based speech recognition.
Automatic Speech Recognition and Understanding, IEEE Workshop 2003
- [16] KINGSBURY, B.
Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling.
Acoustics, Speech and Signal Processing, IEEE International Conference 2009

- [17] TARAR, S.
Speech analysis: Desktop items activation using Dynamic time warping algorithm.
Computer Science and Information Technology (ICCSIT), 3rd IEEE International Conference 2010
- [18] RASIPURAM, Ramya, MAGIMAI.-DOSS, Mathew,
Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model.
Idiap Research Report, 02-2014

Seznam tabulek

Seznam použitých zkratek

AM	Acoustic Model
ASR	Automatic Speech Recognition
CA	Cepstral Analysis
DTFT	Discrete-Time Fourier Transform
HMM	Hidden Markov Model
HMP	Hidden Markov process
LES	Local Emission Score
LM	Language Model
LTI	Linear Time-Invariant
MAP	Maximum a Posteriori
STFA	Short-Time Fourier Analysis
TrogASR	Translating Online Game using Automatic Speech Recognition
TS	Transition Score

Přílohy