Prototypování rozpoznávačů řeči pro nové jazyky

Popis technologie a protokol o ověření

Martin Karafiát, František Grézl, Ekaterina Egorova, Miloš Janda a Jan Černocký, Michal Kašpar:

BUT Speech@FIT, Fakulta informačních technologií VUT v Brně
a Lingea s.r.o., Brno

Brno, srpen 2013

Projekt Ministerstva průmyslu a obchodu České republiky
„Multilingvální rozpoznávání a vyhledávání v řeči pro elektronické slovníky“
Program: TIP
Identifikační číslo: FR-TI1/034

Obsah

1 Úvod

Tato zpráva popisuje ověřenou technologii pro tvorbu multilingválních rozpoznávačů řeči, která je výstupem projektu MPO TIP FR-TI1/034 „Multilingvální rozpoznávání a vyhledávání v řeči pro elektronické slovníky“. Výstupem technologie je rozpoznávač řeči pro daný jazyk.

2 Data

Pro trénování rozpoznávačů jsou nutná řečová data přepsaná na úrovni jednotlivých slov. Technologie byla ověřena na databázi GlobalPhone [GP2010], je však použitelná pro jakoukoliv databázi. K produkci modelů pro základní funkcionalitu je vhodné mít k disposici minimálně 20 hodin řeči (po odstranění ticha) z daného jazyka, s přidáním dat se vlastnosti výsledného rozpoznávače zlepšují. K datům je rovněž nutné disponovat výslovnostním slovníkem.

3 Forced alignment

Pro trénování zvolených akustických modelů založených na neuronových sítích je nutné přesné časové zarovnání jednotlivých fonémů s řečovým signálem. To je vygenerováno pomocí tzv. vnuceného zarovnání (forced-alignment) pomocí GMM/HMM rozpoznávače. Ten je detailně popsán v [Karafiat2012], následující text obsahuje pouze základní body:

- Jedná se o systém založený na skrytých Markovových modelech s hustotami pravděpodobností modelovanými pomocí směsí Gaussových rozložení (GMM/HMM). Základními jednotkami pro akustické modelování jsou vázané stavy HMM (tied states), které překračují hranice slov (cross-word). Model obsahuje cca 3000 takových stavů a každý stav obsahuje 18 Gaussovek.
- Akustickými parametry je zde 13 základních perceptuálních lineárně prediktivních koeficientů vygenerovaných pomocí toolkitu HTK http://htk.eng.cam.ac.uk/ a doplněných rychlostními (delta) a akceleračními (delta delta) koeficienty. Je aplikována normalizace PLP koeficientů pomocí odečítání střední hodnoty a dělení směrodatnou odchylkou (mean and variance normalization), příslušné statistiky jsou odhadnuty na každé promluvě.
- Systém je trénován s jednoduchým kritériem maximální věrohodnosti (maximum likelihood) s postupným navyšováním počtu Gaussovek ve stavech (mix-up training). S natrénovaným systémem jsou pak pomocí Viterbiho dekódování s jedinou variantou získána zarovnání na jednotlivé fonémové stavy (počátek – střed – konec fonému).

Forced alignment je realizován toolkitem STK, dostupným na http://speech.fit.vutbr.cz/software/hmm-toolkit-stk.

4 Fonémové sady

Vzhledem k častému požadavku koncových uživatelů na vlastní fonémovou sadu, lišící se od sady použité v trénovaní databázi, je možné tyto sady mapovat. Ukázka takového mapování je na Obr. 1, levý sloupec obsahuje fonémy sady Lingea, pravý obsahuje fonémy rozpoznávače.

'		\pel	E	\poh	_2	l	l_d
'		w	v	e:	e:	E:	o:
%		E	_at	a	A	m	m
|		H	A	b	b	n	n_d
o:	o:	I	I	c	k	o	O
\pYb	y	\pex	E	d	d_d	p	p
ts	ts	\poh:	_2	e	E	r	R_backs
O:	o:	N	N	f	f	s	s
u:	u:	O	O	g	g	t	t_d
T	s	a:	a:	h	h	u	U
\pel:	_<	S	S_a	i	I	v	v
y:	y:	U	U	j	j	x	X
\pag	a:	V	A	\pou	_9	y	y
\paj	_3_backs	Z	S_a	k	k	z	z
\pch	C	\pog	o:	i:	i:		sil
\pxc	C						

Obrázek 1: Pravidla pro převod fonémových sad, příklad pro němčinu, vlevo je značení fonémů Lingea, vpravo GlobalPhone

5 Struktura a trénování rozpoznávače

Rozpoznávač je založen na kaskádové struktuře neuronových sítí. Detailní popis použitých technologií a diskuse využití jednotlivých funkčních bloků je obsažena v [Schwarz2009], níže jsou opět uvedeny pouze konkrétní hodnoty a detaily vygenerovaných modelů.

5.1 Parametrizace signálu

Parametrizace řečového signálu probíhá na vstupu se vzorkovací frekvencí 16 kHz v rámcích o 400 vzorcích (25 ms), jejichž posun (frame shift) je 160 vzorků (10 ms). FFT výkonové spektrum je zpracováno bankou 23 Melovských filtrů, z trajektorií v jednotlivých pásmech jsou následně odebrány úseky přes 31 rámců (celkem tedy 300 ms). Tyto trajektorie jsou rozděleny na levou a pravou polovinu (left kontext – right kontext, proto je tento systém označován jako LC-RC) a každá je dekorelována pomocí diskrétní cosinové transformace (DCT), z níž je zachováno 11 koeficientů. Parametrizace je realizována pomocí knihovny BS-CORE http://phonexia.com/docs/bsapi/.

5.2 Výpočet posteriorních pravděpodobností fonémových stavů

Výpočet pravděpodobností jednotlivých stavů fonémů je realizován pomocí trojice neuronových sítí:

- Dvě kontextové sítě, každá zpracovávající jednu polovinu kontextu. Počet neuronů ve vstupní vrstvě je 253 (23 pásem krát 11 DCT koeficientů), skrytá vrstva má 1500 neuronů a výstupní vrstva má rozměr odpovídající počtu fonémových stavů dle tabulky 2. Kontextové sítě produkují posteriorní pravděpodobnosti fonémových stavů.

- Merger, který výstupy dvou kontextových sítí spojuje do jediné sady posteriorní pravděpodobnosti fonémových stavů. Velikost jeho vstupní vrstvy je dána dvojnásobkem počtu fonémových stavů podle jazyka, skrytá vrstva má opět velikost 1500 neuronů a počet neuronů výstupní vrstvy je dán opět počtem fonémových stavů daného jazyka.

5.3 Trénování neuronových sítí

Všechny tři neuronové sítě jsou trénovány pomocí standardního kritéria maximální entropie na akustických datech a fonémových zarovnáních získaných pomocí GMM/HMM systému. Zarovnání v tomto případě bylo fixní, nebylo pře-generováno pomocí neuronových sítí. K trénování je použit toolkit TNet http://speech.fit.vutbr.cz/software/neural-network-trainer-tnet.

5.4 Rozpoznávání

Výsledné posteriorní pravděpodobnosti jsou využity v rozpoznávacím systému ve standardním Viterbiho algoritmu. Pro ověřovanou aplikaci realizuje Viterbiho algoritmus nalezení optimálního zarovnání referenční a uživatelovy promluvy na fonémy (forced alignment), podobně jako při tvorbě cílů pro trénování neuronové sítě. Viterbiho algoritmus pracuje pomocí triviální rozpoznávací sítě, kterou tvoří jediná výslovnostní varianta daného slova. Viterbiho algoritmus je implementován pomocí BS-CORE.

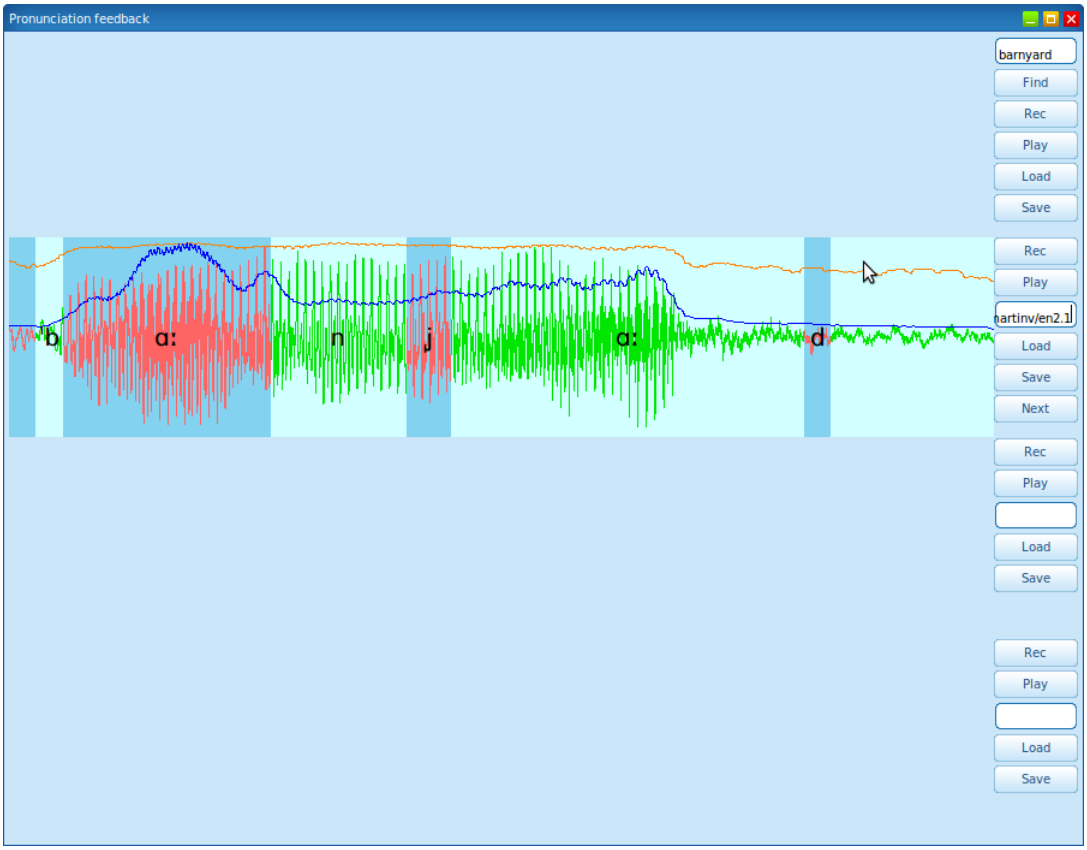
6 Ověření technologie

Testování probíhalo na čtyřech jazycích v demonstrátoru aplikace pro ověřování výslovnosti Lingea (obrázek 2). Testované jazyky byly *angličtina, němčina, turečtina a španělština*.

6.1 Testovací protokol

Pro každý testovaný jazyk

1) Vybere se množina slov pokrývající všechny fonémy vyskytující se ve výslovnostních slovnících Lingea
2) Vybranou množinu slov namluví jeden nebo několik různých mluvčích.
3) Záznamy se rozdělí na jednotlivá slova do souborů, jejichž názvy jsou slova v nich namluvená. Z každého původního záznamu vznikne jedna sada testovacích slov.
4) Pro slova z jednotlivých sad se postupně vytvářejí a zobrazují grafy výslovnosti, slova se opakovaně přehrávají s běžícím ukazatelem aktuální polohy v grafu, a postupně se kontrolují hranice jednotlivých fonémů, viz ilustrace na obrázku 2.



Obrázek 2: Testovací aplikace.

Po zkontrolování celé nahrávky daného slova se zapíše jeho celkové hodnocení ve formátu:

```
mluvci/sada.pokus hodnoceni jednotlivych slov oddelena mezerami
```

kde hodnocení je dáno následovně:

1 – všechny intervaly odpovídají skutečným polohám fonémů
2 – většina intervalů odpovídá skutečným polohám fonémů
3 – většina intervalů neodpovídá skutečným polohám fonémů
4 – žádný interval neodpovídá skutečné poloze fonému

6.2 Osoby účastnící se testování

Testovací data byla namluvena několika nerodilými mluvčími. Poslechovou kontrolu na demonstrátoru provedl, výsledky zaznamenal a použitelnost pro aplikace vyhodnotil Michal Kašpar. Zprávu zpracoval Jan Černocký.

6.3 Výsledky

Následující tabulky uvádějí detailní výsledky pro jednotlivé jazyky:

Jazyk	**Španělština**
Testovaná slova	pullover, vajilla, ping-pong, facultad, gañir, embriaguez, enfurruñado, zigzaguear, individualización, Chad, follaje, enjuiciamiento, 95,

	maldispuesto, caprichosidad, virginidad, biocombustible, ganchillo, Qatar, engañaba, bibliografía, refunfuñar, chachi, Devuelves, chinchilla, enjuague, fluctuaciones, comprometimiento, Iraq, efectividad, fragilidad, guija, cigüeñal, cachipolla, gruñir, superproducción, cigüeña, ambigüedad, jeringuilla, borgoña, vellosidad, champiñón, llegue, subjetivismo, allá, sed
Mluvčí/soubor a detailní výsledky	A/sp9.1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Souhrné výsledky známka/počet	1: 44 2: 2 celkem 46 slov
Mluvčí/soubor a detailní výsledky	B/sp9.1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Souhrné výsledky známka/počet	1: 43 2: 3 celkem 46 slov
Hodnocení	Španělské modely poskytují velice kvalitní fonémové zarovnání.

Tabulka 1: Hodnocení zobrazování výslovnosti - španělština.

Jazyk	**Němčina**
Testovaná slova	aber, Ärztin, Jacke, Terrain, oben, sperrig, Appartement, Papaya, jünger, öffnet, ungefähr, Thriller, Regie, Ehe, nachfüllen, Cousin, übel, gegenüber, verbunden, Parasailing, woanders, Abend, müssen, tagsüber, Zwischenstation, Abfallzerkleinerer, Onkel, Sweatshirt, geeignet, Erfolg, Vollpension, über, Sternanis, erinnern, Juan, SMS, können, Gingerale, Essen, Lohnerhöhung, unser, Bananensplit, Croutons, Rührei, Ihrer, U-Bahn, Ehemann, Akku, Schuljahr, Euro, immer, Forellensaison, PS, Deck, Ufer, Ober, Feuerzeug, Shortcut, öffnen, Behandlungsmöglichkeiten, Check-in-Schalter, Ihre, Deodorant, Sun, irgendwie, Seniorenermäßigung, Löchern, ähnlich, gehe, sechzig, Crêpe
Mluvčí/soubor a detailní výsledky	C/ge2.1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 3 1 1 3 1 2 1 2 1 1 1 1
Souhrné výsledky známka/počet	1: 65, 2: 5, 3: 2 celkem 72 slov

Hodnocení	Problémy nastávají u slov s větším množstvím šumu a u dlouhých slov s opakujícími se fonémy. Zvláště problematický případ nastane, je-li foném nalezen v nahrávce dříve místo jiného fonému: jedna chyba tak „rozhází“ podstatně delší časový úsek. Natahují se i plozívy (například g/k), problém je i s k/t/D na konci. Typickými příklady slov s problémy jsou např. „Behandlungsmoglichkeiten“, „Deodorant“, „irgendwie“. Celkově je však dojem ze zarovnání v němčině pozitivní.

Tabulka 2: Hodnocení zobrazování výslovnosti - němčina.

Jazyk	**Turečtina**
Testovaná slova	hoşgörüsüz, KDV, patinajcı, yığmak, vahvahlamak, gaga, TV, sığlık, köpoğlu, boğmaca, reportajçı, boğmak, gönderebildiniz, burjuvazi, hatchback, URL, WC, çiçekböceği, jip, görünüyorsunuz, çığlık, sığmak, çöpçü, doğdunuz, fotoğrafya, patinajçı, burjuva, pedagojik, bagaj, hoşgörüsüzlük, özgeçmiş, sığdırmak, yığdırmak, başçavuş, yığma, coğrafya, öhö, Moğolca, hafifmeşrep, doğmuş, dezenformasyon, fotoğrafçı, saflaştırılmış, püskürtücü, hijyen, lezzetsiz, lâğvetmek
Mluvčí/soubor a detailní výsledky	D/tr9.1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
Souhrné výsledky známka/počet	1: 36, 2: 11 celkem 47 slov
Hodnocení	V turečtině docházelo k problémů nejčastěji na konci slova u hlásek „k“. Celkově je však dojem opět příznivý.

Tabulka 3: Hodnocení zobrazování výslovnosti - turečtina.

Jazyk	**angličtina**
Testovaná slova	barnyard, earphone, overestimation, abjuration, jalousie, umlaut, layup, negligee, indoors, orangeade, earthshine, Usenet, outcome, turbogenerator, theurgy, miniaturization, epithelioma, consubstantiation, oilbird, coccidioidomycosis, Uruguayan, holdall, update, therefor, oozy, upthrow, supervision, etiology, governmental, churchyard, antiabortion, outhouse, uncoordinated, exempli gratia, disappointing, oilfield, peracetic, indecision, archangel, gee-gee, armchair, demiurge, wheatgerm, offshore, thitherto, offshoot, orang-utan, thereby,

	violoncello, outweigh, audio, microcircuitry, cardioid, yardarm, gigolo, trapezoid, aubergine, zapateado, earthman, thereto, oubliette, Taiwanese, anteater, earache, eyehole, paediatric, thruway, Bavarian, airway, eelpout, brouhaha, outvote, etherize, red-eye, aye-aye
Mluvčí/soubor a detailní výsledky	E/en2.1 1 1 1 1 1 2 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 2
Souhrné výsledky známka/počet	1: 68 2: 7 celkem 75 slov
Hodnocení	Anglické modely poskytují velice kvalitní fonémové zarovnání.

Tabulka 4: Hodnocení zobrazování výslovnosti - angličtina.

6.4 Závěr hodnocení

Při manuálním testování zarovnání hranic fonémů na testovacích nahrávkách se zjistilo, že často lze odhalit nesrovnalosti ve výslovnosti i pouhým spouštěním uživatelské výslovnosti s graficky vyznačenou polohou přehrávaného místa v grafu. Byl-li vysloven jiný foném, než je právě pod běžícím kurzorem, stává se při tomto způsobu přehrávání poměrně nápadným. Ještě nápadnější je pak vložení fonému, který v přepisu není. Zdá se, že podrobně popsaný graf výslovnosti zdůrazňuje detaily ve výslovnosti, které jsou na úrovni celých slov sotva postřehnutelné.

Testování nejen potvrdilo, že vyvinutá technologie je kvalitní a poskytuje modely s vysokou přesností, ale navíc ukázalo, jak užitečná může být pomůcka na této technologii založená i pro zkušeného nerodilého mluvčího.

7 Literatura

[GP2010] XLingual, GmbH & Co. KG: Documentation GLOBALPHONE: a Multilingual Text & Speech Database, Version 3.0, November 2010, Heidelberg, Germany.

[Karafiat2012] Karafiát, M., Janda, M., Černocký, J., Burget, L.: Region Dependent Linear Transforms in Multilingual Speech Recognition, In: *Proc. International Conference on Acoustics, Speech, and Signal Processing 2012*, Kyoto, JP, IEEESP, 2012, s. 4885-4888, ISBN 978-1-4673-0044-5

[Schwarz2009] Schwarz, P.: *Phoneme recognition based on long temporal context*, disertační práce, Brno, CZ, FIT VUT, 2009.