VÝVOJ A EVALUÁCIA MULTILINGVÁLNEJ DATABÁZY PRE SYSTÉMY AUTOMATICKEJ TRANSKRIPCIE SPRÁV ELEKTRONICKÝCH MÉDIÍ

(ABOUT DEVELOPMENT AND EVALUATION OF MULTILINGUAL DATABASE FOR AUTOMATIC BROADCAST NEWS TRANSCRIPTION SYSTEMS)

Matúš PLEVA, Jozef JUHÁR, Anton ČIŽMÁR
Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky a informatiky Technickej univerzity v Košiciach, Letná 9, 042 00 Košice,
tel. +421-55-602 {2334,2333,2294}, E-mail: {Matus.Pleva,Jozef.Juhar,Anton.Cizmar}@tuke.sk

SUMMARY

This paper deals with the problem of broadcast news transcription, segmentation algorithms, training database, and related topics. The automatic transcription of broadcast news is an attractive application of speech technology. To support research in this domain LDC has created the Hub4 American Broadcast News corpus. As there are large differences between the American and European national broadcasts, institutions collaborating in the European action COST278 Spoken Language Interaction in Telecommunication joined together to compile an European Broadcast News Database.
At building the database on multimodal base (video, audio, text) a cooperation of institutions from seven European countries (Portugal, Belgium, Czech republic, Spain, Slovenia, Greece and Slovakia) has been established. This paper presents a description of this database and an introduction to the first steps of a segmentation software evaluation. Namely we are dealing with database building, specifying the focus conditions and speech detection problems.

Keywords: *broadcast news, focus conditions, transcription, multilingual, speech detection, segmentation.*

1. ÚVOD

Súčasná, na znalostiach založená, dynamicky sa rozvíjajúca spoločnosť kladie veľké nároky na včasnú informovanosť o všetkých oblastiach ľudskej činnosti[1]. Prostredníctvom elektronických médií sa denne vysiela množstvo informácií, ktoré z rôznych dôvodov nie je možné sledovať bez obmedzenia. Ďalšou dôležitou skutočnosťou je fakt, že nie každá informácia je rovnako dôležitá pre každého. Vo svete preto rastie potreba automatického distribuovaného archivačného systému spravodajských relácií elektronických médií a metód automatickej analýzy archivovaných nahrávok s cieľom rýchleho, tematicky orientovaného vyhľadávania a výberu špecifickej informácie.

Primárnou požiadavkou je prepis zvukovej nahrávky do textovej formy pomocou automatického rozpoznávania reči. V ďalšom kroku je potrebné takto získané informácie kategorizovať podľa témy (topic detection), podľa hovoriaceho (speaker detection).

Pre zvýšenie kvality procesu automatického rozpoznávania reči je dôležitá fáza optimálneho predspracovania rečového signálu. Práve problematikou predspracovania rečového signálu sa v rámci projektu COST 278 začala intenzívne zaoberať medzinárodná výskumná skupina, ktorej cieľom je vytvoriť univerzálny multilingválny systém spracovania elektronických spravodajských relácií. V súčasnosti už existuje systém s architektúrou Hub4 [1] vyvinutý pomocou americkej databázy elektronických spravodajských relácií (American Broadcast News corpus), vytvorenej konzorciom LDC[2] (Linguistic Data Consortium). Kvôli rozdielom medzi americkým a európskym spravodajským vysielaním sa v rámci akcie COST278 začalo pracovať na vybudovaní paneurópskej databázy spravodajských relácií, na ktorej momentálne participujú inštitúcie z Belgicka, Česka, Portugalska, Slovinska, Grécka, Španielska a Slovenska.

2. AUTOMATIZOVANÝ SYSTÉM TRANSKRIPCIE SPRAVODAJSKÝCH RELÁCIÍ

Problematiku vývoja automatizovaného systému transkripcie spravodajských relácií možno rozdeliť do nasledujúcich blokov:

1. *Predspracovanie.* Segmentácia reč/nereč (speech/non-speech), klasifikácia ruchu pozadia, segmentácia podľa hovoriaceho, zoskupovanie a označenie hovoriacich [2].

2. *Transkripcia.* Automatické rozpoznanie reči a jej prepis reči do textovej podoby, vhodnej na archiváciu a ďalšie spracovanie.

[1] *Príspevok bol vypracovaný v rámci projektu COST278 (http://cost278.org/- Spoken Language Interaction in Telecommunication) a za podpory projektu Grantovej agentúry Slovenskej Republiky VEGA č. 1/1057/04.*

[2] http://www.ldc.upenn.edu/

3. *Detekcia témy*. Nájdenie a označenie témy reportáže a zoskupovanie častí rozpoznaného textu do blokov podľa príslušnosti k téme a reportáži.

4. *Prehľadávanie*. Efektívne algoritmy triedenia a zoskupovania častí spravodajstva podľa oblastí, tém, geografickej polohy miesta odkiaľ správa pochádza, mena televízneho moderátora, resp. spravodajcu a jazyka, v akom bol príspevok vysielaný.

Kvalita kompletného transkripčného systému je závislá od kvality každého komponentu tohto systému. Bloky automatického rozpoznávania reči, detekcie témy aj vyhľadávacie algoritmy sú závislé od jazyka. Nezávislým od jazyka je blok predspracovania, preto sa javí ako najvhodnejší na evaluáciu multiligválnej databázy ako aj algoritmov predspracovania pomocou multiligválnej databázy (počet použitých jazykov, označovanie parametrov rečových segmentov, rečníkov, a iné).

3. BUDOVANIE DATABÁZY

V prvej fáze vývoja multilingválneho paneurópskeho transkripčného systému bolo potrebné vytvoriť databázu spravodajských relácií a ich ortografických transkripcií s manuálne klasifikovanými rečovými segmentmi. Každá zúčastnená strana v prvom kroku skompletizovala 3 hodiny televízneho spravodajstva verejnoprávnej alebo súkromnej televízie, vysielaného v jazyku danej krajiny. Následne sa vykonala ručná ortografická transkripcia zvukových dát pomocou nástroja Transcriber[3] tak, aby vyhovovala štandardu LDC[4] podľa prepisového protokolu NIST[5]. Vzhľadom na dosiahnutie maximálnej konzistencie databáz a kompatibilitu s konvenciu použitou v amerických databázach BN96 a BN97 sa táto fáza uskutočnila v INESC ID Lisabon na Workshope v júni 2003.

jazyk	Akro-nym	Počet Relácií	Celkové Množstvo (sek.)	Užitočné Množstvo (sek)	Počet slov	Počet rôznych slov
portugalský	PT	6	12661	11002	33949	5719
holandský	BE	6	9763	9201	26456	5018
český	CZ	10	10856	10132	27642	8834
galícia	GA	3	13497	11813	30730	6155
grécky	GR	3	10438	9523	22580	6174
slovinský	SI	3	10921	8627	22269	7239
slovenský	SK	9	11431	9789	25770	8887

Tab. 1 Dostupné dáta
Tab. 1 Available data

[3] http://www.etca.fr/CTA/gip/Projets/Transcriber/
[4] http://www.ldc.upenn.edu/Projects/Corpus_Cookbook/transcription/broadcast_speech/english/index.html
[5] www.nist.gov/speech/tests/bnr/hub4_96/h4spec.htm

V súčasnosti sú k dispozícii databázy jazyka portugalského, holandského (používaného v Belgicku), českého, slovinského, slovenského, gréckeho a jazyka Galície (časť Španielska na sever od Portugalska). Prehľad dostupných dát je v Tab. č. 1 (užitočné množstvo je dĺžka hovoreného slova).

4. KLASIFIKÁCIA AKUSTICKÝCH PODMIENOK

Vzhľadom k tomu, že počas spravodajských relácií sa neustále mení kvalita reči, šum v pozadí, a takisto sa menia samotní rečníci, je nutné adaptovať trénované modely na dané situácie. Súčasťou budovania databázy bola nielen manuálna ortografická transkripcia, ale aj popis jednotlivých rečových segmentov podľa akustických podmienok.

Podľa dohodnutej konvencie reč klasifikujeme do týchto skupín [3]:

- F0: *Základná vysielaná reč* (baseline broadcast speech) - táto podmienka popisuje reč, ktorá je smerovaná priamo do vysielacieho reťazca, a je zaznamenaná v tichom štúdiu, s odstupom signálu od šumu viac ako 20dB. Predpokladáme tiež, že táto reč vznikla čítaním pripraveného textu.
- F1: *Spontánna vysielaná reč* (spontaneous broadcast speech) - táto podmienka popisuje reč, ktorá je smerovaná jednému alebo viacerým konverzačným partnerom, teda odohráva sa spontánna konverzácia. Tento záznam je uskutočnený v tichom štúdiu, s odstupom signálu od šumu viac ako 20dB.
- F2: *Reč cez telefónnu linku* (speech over telephone channels) - táto podmienka popisuje reč získanú zo zdroja s úzkym prenosovým pásmom, napríklad z telefónu, mobilného telefónu, diktafónu, záznamníka alebo podobného média so šírkou pásma maximálne 4kHz.
- F3: *Reč s hudbou v pozadí* (speech in the presence of background music) - táto podmienka určuje reč, ktorá zodpovedá podmienkam F0 alebo F1, len s tým rozdielom že je vysielaná s hudbou v pozadí. Pomer výkonu signálu a hudby je taký, aby reč bola zrozumiteľná bežnému poslucháčovi, teda predpokladáme rozpätie medzi 10 až 20 dB.
- F4: *Reč v degradovaných akustických podmienkach* (speech under degraded acoustical conditions) - táto podmienka popisuje reč, ktorá je degradovaná iným spôsobom ako hudbou v pozadí alebo použitím telefónnej linky. Zdroje degradácie môžu byť šum, šum prostredia, alebo nelineárne skreslenie. Odstup signálu od šumu (SNR) sa predpokladá v medziach 10 až 20 dB.

- F5: *Reč rečníka, hovoriaceho iným ako materinským jazykom* (speech from non-native speakers) - táto podmienka určuje reč, ktorá zodpovedá podmienkam F0, ale je hovorená rečníkom, pre ktorého nie je táto reč prirodzenou materinskou rečou. Táto reč je dostatočne zrozumiteľnou pre bežného poslucháča. Je hovorená plynulo rečníkom, ktorý má cudzozemský akcent. Napríklad britský rečník je cudzokrajným rečníkom pre americkú angličtinu. Ak rečník používa iný jazyk, označuje sa to v texte spolu s jazykom aký používa. Nepoužíva sa však klasifikácia F5 ak to je jeho materinský jazyk.
- FX: *rôzne* (miscellaneous) - Predstavuje reč, ktorá nespĺňa ani jednu predchádzajúcu podmienku, alebo reč, ktorá spĺňa viac ako jednu z podmienok F1 až F5. Napríklad cudzokrajný rečník s hudbou v pozadí.

V rámci projektu boli k dispozícii dáta siedmich jazykov. Pre porovnanie uvádzame v Tab. č. 2 klasifikáciu databázy podľa vyššie uvedených akustických podmienok, ktorá bola získaná aplikovaním testovacích segmentačných skriptov, vyvinutých na INESC ID Lisabon na ručné transkripcie dát.

			% viet daných akustických podm.							
Krajina	Počet slov	Rôznych slov	F0	F1	F2	F3	F4	F5	FX	Spolu
BE	26456	5018	38.6	11.5	0.5	2.9	45.1	0.0	1.4	1857
CZ	27642	8834	52.9	18.2	1.2	5.4	21.7	0.1	0.4	1547
GA	33029	6463	29.5	5.1	0.2	6.5	43.3	4.0	11.4	1673
GR	23748	6065	48.7	16.6	0.0	3.1	31.3	0.1	0.2	1624
PT	33949	5719	10.5	7.5	0.0	2.5	76.4	0.5	2.6	1987
SI	22269	7237	61.9	15.3	3.7	8.6	8.1	0.6	1.7	1292
SK	25770	8887	35.7	15.7	4.1	7.6	34.1	0.0	2.8	2023

Tab. 2 Dáta rozdelené podľa akustických podm.
Tab. 2 Data divided under focus conditions

5. POUŽITIE DATABÁZY A ĎALŠIE KROKY NUTNEJ ŠTANDARDIZÁCIE FORMÁTOV

Plánovaným základným použitím databázy v rámci akcie COST278, ale aj mimo nej, je trénovanie a testovanie rôznych segmentačných algoritmov a ich vzájomné porovnávanie. Práve porovnávanie presnosti a kvality jednotlivých algoritmov vyvolalo požiadavku na:

1. Štandardizovanie dátových súborov. Dohodnutými štandardnými dátovými súbormi, ktoré sa používajú, sú:
   - *.rm – obrazový aj zvukový záznam spravodajskej relácie skomprimovaný vo formáte RealMedia kvôli nízkemu dátovému toku a tým nižšej spotrebe miesta na záznamovom médiu. Primárne je určený iba na kontrolu nahrávky a hovoriaceho, môže sa však využiť aj pri automatickom spracovaní.
   - *.wav – monofónny zvukový záznam spravodajskej relácie so vzorkovacou frekvenciou 16 kHz v nekomprimovanej forme.
   - *.trs – štandardný výstupný formát programu Transcriber, ktorý bol využitý na manuálnu ortografickú transkripciu zvukových dát.
   - *.stm - formát vstupných dát (manuálne ortografické transkripcie) na trénovanie segmentačných algoritmov [4].
   - *.seg – výstupný formát segmentačných skriptov, ktorý bol definovaný na workshope v Lisabone. Pozostáva z položiek oddelených tabulátorom, ako je to vidno v Tab. č. 3 a štandardizuje výstupný formát skriptov, ktorý potom slúži na evaluáciu kvality a porovnanie s výsledkami iných algoritmov, či porovnanie s manuálnou segmentáciou zvukových dát.
2. Výber softvéru na evaluáciu kvality segmentačných algoritmov. Zvolený bol porovnávací skript Elis[6] (je dostupný na ftp serveri projektu), ktorý porovnáva výsledky automatickej (.seg súbor) a manuálnej (.stm súbor) segmentácie. Na tomto postupe je založené testovanie algoritmov segmentácie reči a porovnávanie ich výsledkov.
3. Vývoj skriptu na prevod z .trs do .stm formátu podľa rôznych požiadaviek (rozpoznávanie reči, akustických podmienok, detekcia reči, atď). Tento skript je potrebný vzhľadom na to, že je nutné urobiť prevod medzi výstupom Transcribera a štandardným .stm formátom, ktorý sa ďalej používa ako vstupný formát pre všetky skripty. Hoci Transcriber ponúka zápis aj v .stm formáte, nie je schopný korektne zapísať iné ako „Western“ kódové stránky. Tento skript je dostupný v databáze ako trs2stm.

Štart	{čas v sek.}
Koniec	{čas v sek.}
Reč / Nie-reč	{ Speech, Non-speech}
ID rečníka	{číslo}
Pohlavie	{Male, Female, Unknown, Child}
Klasifikácia ruchu pozadia	{Clean, Speech, Music, Ssh, Other}

Tab. 3 Dohodnutý formát .seg súborov
Tab. 3 The .seg file format convention

[6] http://chardonnay.elis.ugent.be/cost278/ELIS_evaluation_software.doc

6. ZÁVER

Získaná databáza trojhodinových blokov správ v siedmich jazykoch je užitočným nástrojom pre ďalší výskum v oblasti multilingválnych aplikácií. Možno ju využiť aj ako nezávislú testovaciu množinu transkripčných systémov audiovizuálnych dát.

Zároveň je táto databáza potenciálom vytvorenia multilingválneho nástroja pre výskum v oblasti transkripcie a archivácie spravodajských relácií v elektronických médiách.

Dosiahnuté výsledky ukazujú potrebu pokračovania vo výskume algoritmov detekcie reči, čoho dôkazom sú značné rozdiely v úspešnosti použitých algoritmov, ktoré testovali niektoré participujúce organizácie [5] (UGent, TULiberec).

Spoločné úsilie, ktoré bolo započaté tvorbou tejto databázy, prispeje k návrhu vylepšených a na európske spravodajské relácie optimalizovaných segmentačných algoritmov novej generácie na multilingválnej báze.

LITERATÚRA

[1] Eide, E. - Maison, B. - Kanevsky, D. - Olsen, P. - Chen, S. - Mangu, L. - Gales, M. - Novak, M. - Gopinath, R.: *Transcription Of Broadcast News With A Time Constraint: Ibm's 10xrt Hub4 System*. Proc. ICSLP'2000 - 6th International Conference on Spoken Language Processing, 00563.pdf, 2000.

[2] Meinedo, H. - Neto, J.: *Automatic Speech Annotation and Transcription in a Broadcast News task.*, automatic-speech-annotation-and.pdf, 2003. http://citeseer.nj.nec.com/571279.html

[3] Hain, T. - Woodland, P. C.: *Segmentation And Classification Of Broadcast News Audio*. Proc. ICSLP'1998 - 4th International Conference on Spoken Language Processing, SL980851.pdf, 1998. http://svr-www.eng.cam.ac.uk/reports/svr-ftp/hain_icslp98.ps.gz

[4] Gauvain, J. L. - Lamel, L. - Adda, G.: *Partitioning And Transcription Of Broadcast News Data.* Proc. ICSLP'1998 - 4th International Conference on Spoken Language Processing, SL980084.pdf, 1998.

[5] Vandecatseye, A. - Martens, J. P. - Neto, J. - Meinedo, H. - Mateo, C. G. - Dieguez, J. - Mihelic, F. - Zibert, J. - Nouza, J. - David, P, - Pleva, M. - Cizmar, A. - Papageorgiou, H. – Alexandris, Ch.: *The COST278 pan-European Broadcast News Database.* Proc. LREC' 2004 – 4th International Conference On Language Resources And Evaluation, 2004.

BIOGRAPHY

Matúš Pleva was born in Košice, Slovakia in 1977. In 1991 he graduated (MSc.) with honours at the department of Electronics and Multimedia Telecommunications of the Faculty of Electrical Engineering and Informatics at Technical University in Košice. He's working on his PhD. thesis in the field of Multilingual speech recognition and Dialogue systems.

Jozef Juhár was born in Poproč, Slovakia in 1956. He graduated from the Technical University of Košice in 1980. He received Ph.D. degree in Radioelectronics from Technical University of Košice in 1991, where he works as an Associate Professor at the Department of Electronics and Multimedia Communications. His research interests include digital speech and audio processing, speech/speaker recognition, speech synthesis and spoken dialogue systems in telecommunication networks.

Anton Čižmár was born in Michalovce, Slovakia in 1956. He is a graduate of the Slovak Technical University in Bratislava in 1980, at the Department of Telecommunications. He holds a Ph.D. degree in Radioelectronics from the Technical University of Košice in 1986, where he works as a Full Professor at the Department of Electronics and Multimedia Communications. His scientific orientation is telecommunication management, project management, broadband, information and telecommunication technologies, multimedia systems, telecommunications networks and services, speech processing, coding and recognition and finally coding and modulation techniques.