Agentúra na podporu výskumu a vývoja
Mýtna 23, P.O.BOX 839 04, 839 04 Bratislava 32

Formulár ZK

Záverečná karta projektu

Názov projektu

Automatické titulkovanie audiovizuálneho obsahu pre osoby so sluchovým postihnutím

Evidenčné číslo projektu **APVV-15-0517**

Zodpovedný riešiteľ **prof., Ing. Jozef Juhár, CSc.**

Príjemca **Technická univerzita v Košiciach - Fakulta elektrotechniky a informatiky**

Názov pracoviska, na ktorom bol projekt riešený

1. Technická univerzita v Košiciach - Fakulta elektrotechniky a informatiky, Katedra elektroniky a multimediálnych telekomunikácií
2. Ústav informatiky Slovenskej akadémie vied v Bratislave, Oddelenie analýzy a syntézy reči

Názov a štát zahraničného pracoviska, ktoré spolupracovalo pri riešení

-

Udelené patenty/podané patentové prihlášky, vynálezy alebo úžitkové vzory, ktoré sú výsledkami projektu

-

Najvýznamnejšie publikácie (knihy, články, prednášky, správy a pod.) zhrňujúce výsledky projektu – uveďte aj publikácie prijaté do tlače

[1] Vavrek, J., Viszlay, P., Lojka, M., Juhár, J., Pleva, M. (2018) Weighted fast sequential DTW for multilingual audio Query-by-Example retrieval. Journal of Intelligent Information Systems, Volume 51, Issue 2, ISSN 0925-9920, pp. 439-455, https://doi.org/10.1007/s10844-018-0499-2 (Current Content, ISI Thomson IF=1.107)
[2] Hládek, D., Staš, J., Ondáš, S., Juhár, J., Kovács, Z. (2017) Learning string distance with smoothing for OCR spelling correction. Multimedia Tools and Applications, Volume 76, Issue 22, ISSN 1380-7501, pp. 24549-24567 https://doi.org/10.1007/s11042-016-4185-5 (Current Content, ISI Thomson IF=1.530)
[3] Pleva, M., Bours, P., Ondáš, S., Juhár, J. (2017) Improving static audio keystroke analysis by score fusion of acoustic and timing data. Multimedia Tools and Applications, Volume 76, Issue 24, ISSN 1380-7501, pp. 25749-25766 https://doi.org/10.1007/s11042-017-4571-7 (Current Content, ISI Thomson IF=1.530)
[4] Vavrek, J., Feciľák, P., Juhár, J., Čižmár, A. (2017) Classification of broadcast news audio data employing binary decision architecture. Computing and Informatics, Volume 36, Issue 4, ISSN 1335-9150, pp. 857-886 (Current Content, ISI Thomson IF=0.504), Dostupné online: http://www.cai.sk/ojs/index.php/cai/article/view/2017_4_857/845
[5] Guoth, I., Rusko, M., Ritomský, M., Trnka, M., Darjaa, S. (2017) Exploitation of phase-based features for emotional arousal evaluation from speech (Abstract). The Journal of the

Acoustic Society of America, Vol. 141, No. 5, ISSN 0001-4966, page 3468 https://doi.org/10.1121/1.4987206 (ISI Thomson IF=1.547)
[6] Rusko, M., Trnka, M., Darjaa, S., Ritomský, M., Guoth, I. (2017) Influence of noise on the speaker verification in the air traffic control voice communication (Abstract). The Journal of the Acoustic Society of America, Vol. 141, No. 5, ISSN 0001-4966, page 3469 https://doi.org/10.1121/1.4987211 (ISI Thomson IF=1.547)
[7] Staš, J., Viszlay, P., Lojka, M., Koctúr, T., Hládek, D., Juhár, J. (2018) Automatic transcription and subtitling of Slovak multi-genre audiovisual recordings. In: Human Language Technology, Challenges for Computer Science and Linguistics, Vetulani, Z., Mariani, J., Kubis, M. (Eds), LNCS, Volume 10930, Springer, Cham, ISBN 978-3-319-93781-6, pp. 42-56, https://doi.org/10.1007/978-3-319-93782-3_4
[8] Lojka, M., Viszlay, P., Staš, J., Hládek, D., Juhár, J. (2018) Slovak broadcast news speech recognition and transcription system. In: Advances in Network-Based Information Systems, Barolli, L. et al. (Eds), LNDECT, Volume 22, Springer, Cham, ISBN 978-3-319-98529-9, pp. 385-394, https://doi.org/10.1007/978-3-319-98530-5_32
[9] Staš, J., Hládek, D., Juhár, J. (2018) Modeling of filled pauses and prolongations to improve Slovak spontaneous speech recognition. In: Cognitive Infocommunications, Theory and Applications, Klempous, R., Nikodem, J., Baranyi, P. (Eds), TIEI, Volume 13, Springer, Cham, ISBN 978-3-319-95995-5, pp. 153-176, https://doi.org/10.1007/978-3-319-95996-2_8.
[10] Koctúr, T., Ondáš, S., Juhár, J. (2017) Speech corpus generation based on n-gram confidence measure classification. In: Proc. of the 59th International Symposium ELMAR 2017, Zadar, Croatia, ISBN 978-953-184-230-3, ISSN 1334-2630, pp. 149-152, DOI: 10.23919/ELMAR.2017.8124456

Uplatnenie výsledkov projektu

Hlavnou oblasťou uplatnenia výsledkov projektu je oblasť automatickej resp. poloautomatickej tvorby titulkov k audiovizuálnemu obsahu, vhodných najmä (ale nielen) pre osoby so sluchovým postihnutím. S týmto cieľom bol v rámci riešenia projektu vyvinutý systém na automatický prepis reči do textu, ktorého architektúra vychádza z klient-server koncepcie. Na tejto architektúre sú založené tri pilotné verzie systémov na automatické titulkovanie audiovizuálneho obsahu (podrobnejšie charakterizované nižšie v časti „Súhrn výsledkov riešenia a naplnenia cieľov projektu“), ktoré sú vhodné na každodenné používanie moderátormi a redaktormi v televíznom vysielaní, redaktormi, ktorí publikujú na spravodajských webových portáloch, či rôznymi inými používateľmi, ktorí vytvárajú vlastné video blogy.
Ďalšou z ďalších zaujímavých a už teraz sa otvárajúcich možností uplatnenie systému automatický prepis a indexácia obsahu audiovizuálnych archívov a tvorba a spravovania inteligentných databáz s neštruktúrovaným obsahom (napr. v zdravotníctve, vo verejnej správe apod.).
Systém na automatický prepis reči do textu novej generácie, založený na najmodrenejších metódach, algoritmoch a svetových poznatkoch, bude tiež základom pokračujúcich výskumných aktivít, plánovanie nových projektov základného a aplikovaného výskumu, ako aj predpokladom na prehĺbenie medzinárodnej spolupráce v tejto oblasti.
V neposlednej miere systém tvorí základnú platformu pre výchovu a vzdelávanie budúcich špecialistov v oblasti informačných a komunikačných systémov a technológií napr. formou riešenia záverečných prác vo všetkých troch stupňoch vysokoškolského štúdia s cieľom analyzovať nové poznatky, vznikajúce v tejto oblasti a implementovať ich do existujúcich modulov alebo navrhovať moduly nové.

Súhrn výsledkov riešenia projektu a naplnenia cieľov projektu v slovenskom jazyku (max. 20 riadkov)

Hlavným cieľom projektu bol aplikovaný výskum v oblasti spracovania prirodzenej reči a jazyka a vývoj špeciálne prispôsobeného systému na automatické titulkovanie audiovizuálneho obsahu v slovenčine na báze automatického rozpoznávania plynulej reči s veľkým slovníkom, ktorý je určený najmä pre osoby so sluchovým postihnutím. Tento cieľ sa naplnil vývojom pilotnej verzie systému na automatický prepis reči do textu novej generácie, založenej na použití hlbokých neurónových sietí, architektúra ktorého vychádza z klient-server koncepcie. Funkčnosť a aplikovateľnosť systému je demonštrovaná pomocou troch

samostatne použiteľných komponentov (podsystémov):
1. Podsystém na off-line titulkovanie audiovizuálneho obsahu – je určený na titulkovanie televíznych príspevkov, ktoré sú vytvorené s dostatočným predstihom pred samotným vysielaním redaktormi pracujúcimi v teréne. Tieto sú prepísané systémom na automatické rozpoznávanie spontánnej reči (umiestnenom na vzdialenom serveri) v čo možno najlepšej kvalite, pričom redaktor ich má možnosť dodatočne upraviť a prispôsobiť vysielaniu pomocou nami vytvorenej webovej služby, ktorá obsahuje vstavaný editor titulkov. Systém obsahuje tiež vstavané algoritmy na optimálne prerozdelenie titulkov tak, aby boli zobrazené vo vysielaní čo najdlhšie.
2. Podsystém na online titulkovanie audiovizuálneho obsahu – uvažuje so živými vstupmi redaktorov v televíznom vysielaní spravodajstva. Bol vyvinutý rýchly online systém na automatické rozpoznávanie spontánnej reči, ktorý sleduje prichádzajúci rečový vstup z mikrofónu redaktora s logikou rýchleho spracovania nahrávky a zobrazovania titulkov s čo najmenším oneskorením.
3. Podsystém na online zarovnávanie textu s audiovizuálnym obsahom – pracuje s audiovizuálnym obsahom, ku ktorému už existuje časovo nezarovnaný textový prepis. Vstupom môžu byť napr. texty zo štúdiovej čítačky – teleprompterа, scenáre k divadelným hrám a pod. Systém po spustení takto pripravený text vo veľmi krátkom čase synchronizuje (zarovná) so zvukovou stopou audiovizuálneho obsahu. Titulky sú následne zobrazované v reálnom čase už s vyslovením prvých slov.
Z kvantitatívneho hľadiska boli ciele projektu prekročené prakticky vo všetkých naplánovaných ukazovateľoch, najmä však:
- Bol výrazne prekročený počet plánovaných publikácií (4 karentované, 2 impaktované a 42 publikácií v recenzovaných časopisoch a zborníkoch z domácich a medzinárodných vedeckých konferencií);
- Bol prekročený objem zozbieraných a spracovaných akustických dát (370 hodín nových manuálne anotovaných rečových nahrávok televíznych a športových novín a diskusných relácií); s celkovým objemom cca 1000 hodín manuálne anotovaných rečových nahrávok a 1500 hodín automaticky spracovaných rečových dát sa oba tímy zaradili medzi pracoviská európskej úrovne;
- Vo významnej miere boli do vedeckého výskumu zapojení študentov všetkých stupňov VŠ štúdia;
- Dosiahnuté výsledky zaujali nielen odbornú verejnosť, ale aj poskytovateľov televízneho vysielania či spravodajské médiá publikujúce na Internete.
Stanovené ciele projektu aplikovaného výskumu boli splnené v celom svojom rozsahu. Dosiahnuté výsledky sú pôvodné a v ukazovateľoch používanej metodiky a dosahovanej presnosti automatického prepisu reči do textu dosahujú úroveň, porovnateľnú so svetovým výskumom. Vytvorili sa tak podmienky na praktické uplatnenie rečových technológií v slovenskom jazyku v oblasti sprístupnenia informačným zdrojov sluchovo hendikepovaným občanom, ale súčasne aj podmienky na nadväzujúci výskum a vývoj, aplikovateľný v ďalších oblastiach spoločenského života (médiá, zdravotníctvo, verejná správa atď).

Súhrn výsledkov riešenia projektu a naplnenia cieľov projektu v anglickom jazyku (max. 20 riadkov)
The main objective of the project was applied research in the field of speech and language processing and development a specially adapted automatic speech recognition system with a large vocabulary for automatic subtitling of audiovisual content in Slovak, especially for people with hearing disabilities. This goal was accomplished by developing a pilot version of the automatic speech transcription system based on the use of deep neural networks, the architecture of which is based on the client-server concept. The functionality and applicability of the system is demonstrated by three separate components (subsystems):
1. Sybsystem for offline subtitling of audiovisual content – designed for creation of subtitles by journalists to be ready before broadcasting. The subtitles are created by a system for automatic speech recognition for spontaneous speech (located on remote server) in the best quality available, while they can be further edited and customized for broadcasting by using our web-based service with built-in subtitle editor. The system also contains well designed algorithms for the subtitles distribution along the broadcast, so they are displayed as long as possible for convenient reading.
2. Subsystem for online subtitling of audiovisual content – covering the live feed during the

broadcast news. The fast spontanneous speech recognition system was designed, monitoring the journalist's microphone input, so that the speech recognition can be performed fast and the subtitles generated with the smallest delay possible.
3. Subsystem for online text and audiovisual content alignment – works with audiovisual content with text transcription available, but without the time alignment between them. The system can be used by anchors reading a text from teleprompter or in theaters during performance. The principle is to align the text with audiovisual content as fast as possible, while the subtitles are displayed in advance with first spoken words.
From a quantitative point of view, the objectives of the project were virtually exceeded in all planned indicators, but in particular:
- The number of planned publications (4 copies, 2 impacted and 42 publications in reviewed journals and proceedings from domestic and international scientific conferences) was significantly exceeded;
- The volume of collected and processed acoustic data (370 hours of new manual annotated speech of television and sports news and discussion sessions) was exceeded; with a total of about 1,000 hours of manual annotated speech and 1,500 hours of automated speech data, both teams ranked among European-level workplaces;
- Significantly, the involvement of students of all levels of university studies has been involved in scientific research;
- The results achieved were not only for the professional public, but also for TV broadcasters and news media on the Internet.
The objectives of the applied research project have been met in its entirety. The results obtained are original and, in the indicators of the methodology used and the achieved accuracy of automatic speech transcription, reach a level comparable to that of the world research. The conditions for the practical application of speech technologies in the Slovak language in the area of access to information resources for hearing-impaired citizens, as well as conditions for follow-up research and development applicable to other areas of social life (media, health, public administration, etc.) have been created.