INSPO – Internet a informační systémy pro osoby se specifickými potřebami
16. března 2013, Kongresové centrum Praha

AUTOMATICKÉ VYTVÁŘENÍ DOPROVODNÉ ZVUKOVÉ STOPY TELEVIZNÍHO VYSÍLÁNÍ PRO SLUCHOVĚ POSTIŽENÉ

Jindřich MATOUŠEK, Zdeněk HANZLÍČEK, Daniel TIHELKA
Katedra kybernetiky, Fakulta aplikovaných věd, Západočeská univerzita v Plzni
jmatouse@kky.zcu.cz, zhanzlic@kky.zcu.cz, dtihelka@kky.zcu.cz

Anotace: *Příspěvek popisuje poslední vývoj v oblasti automatického vytváření doprovodné zvukové stopy televizního vysílání, určené sluchově postiženým divákům České televize. Doprovodná zvuková stopa neobsahuje podkresovou hudbu ani žádné ruchy a zvuky prostředí, ale pouze řečový signál, resp. hlasové dialogy vytvářené automaticky ze skrytých titulků pomocí technologie syntézy řeči z textu.*

1. Úvod

Tento příspěvek popisuje poslední vývoj v oblasti automatického vytváření doprovodné zvukové stopy televizního vysílání, určené sluchově postiženým divákům České televize. Automatické vytváření doprovodné zvukové stopy televizního vysílání (nebo též „mluvicí podtitulky“ [1]) je jedním z cílů projektu ELJABR II[1] řešeného na Katedře kybernetiky (KKY) Fakulty aplikovaných věd (FAV) Západočeské univerzity (ZČU) v Plzni ve spolupráci s firmou SpeechTech[2]. V rámci projektu ELJABR II je také řešena úloha automatického titulkování živých pořadů České televize [2][3]. Dále jsou zde zkoumány možnosti přenosu parametrů pro automatizované generování znakové řeči (řízení „avatara“ českého znakového jazyka) a možnosti asistované katalogizace příspěvků v rozsáhlém televizním archivu.

Úloha představovaná v tomto příspěvku, tj. automatické vytváření doprovodné zvukové stopy TV vysílání, je na rozdíl od automatického titulkování, které cílí na hluché a sluchově těžce postižené diváky, zaměřena na diváky České televize s lehčím sluchovým postižením. Jde o poměrně specifickou skupinu televizních diváků. Tito diváci, zejména starší osoby[4], ale i osoby dyslektické a mírně mentálně retardované, které v běžném životě problémy se slyšením nijak významně nepociťují, mají problém se srozumitelností hlasového dialogu ve víceplánové zvukové scéně televizních pořadů. Kombinace dialogu, podkresové hudby, ruchů a atmosfér prostředí (obvykle při nedokonalém poslechovém domácím zařízení) vytváří pro staršího diváka směsici zvuků, ve které se není schopen orientovat [1]. Přidáme-li k tomu poměrně dynamickou dialogovou složku moderních televizních pořadů (tj. např. rychlé střídání tempa řeči, řečových úseků o různé hlasitosti, emotivních či jinak expresivních řečových diskurzů apod.), často pak zmínění starší diváci vůbec nerozumí řečovému dialogu a ztrácejí zájem o sledování takových pořadů.

Řešením uvedených problémů je vytvářet další zvukovou stopu televizního vysílání, která by reflektovala výše uvedené problémy a zpřístupnila by, formou „klidné zvukové stopy“, uvedené skupině televizních diváků inkriminované pořady televizního vysílání. Protože tato stopa v žádném případě nenahrazuje původní zvukovou stopu televizního vysílání (plánuje se její paralelní vysílání s původní stopou, což je díky digitálnímu vysílání možné – na svých domácích televizních přijímačích si ji budou moci diváci v případě potřeby volit individuálně), nazýváme ji **doprovodná zvuková stopa**. Samozřejmě se nabízí možnost tuto doprovodnou zvukovou stopu vytvářet podobně jako původní stopu, tj. dabovat „klidným hlasem“ lidmi-herci anebo potlačit či úplně eliminovat rušící podkresovou a efektovou složku původní zvukové stopy. Vedle vícenákladů s tím spojených zde ale vyvstává zásadní problém – nebezpečí porušení autorských a licenčních práv takto zpracovaných pořadů. Původní zvuková stopa vznikla jako jistý autorský a umělecký záměr tvůrců pořadu (či byla s tímto omezením zakoupena) a jakákoliv úprava výsledné zvukové stopy je pak licenčně

[1] ELJABR II je akronym pro název „ELiminace JAzykových BaRiér handicapovaných diváků České televize“. Projekt je řešen za podpory Technologické agentury České republiky (TA ČR) a registrován pod č. TA01011264. Projekt navazuje na projekt ELJABR, řešený v letech 2006-2011 za podpory MŠMT (reg. č. 2C06020).
[2] http://www.speechtech.cz
[3] Problematika automatického titulkování byla prezentována na několika předchozích ročnících konference INSPO (viz odkazy v literatuře [1], [2], [3]).
[4] Směrnice BBC za takové osoby považuje diváky nad 50 let věku [1].

problematická. Alternativním řešením je zvukovou stopu nemodifikovat, ale vytvořit **zvukovou stopu novou**. Tento přístup byl zvolen v projektu ELJABR II, kde se řečový dialog vytváří automaticky pomocí technologie **počítačové syntézy řeči** (nebo též „hlasové syntézy“) [4], a to konkrétně pomocí technologie **syntézy řeči z textu**. Princip syntézy řeči z textu (značeno zkratkou **TTS**, z angl. text-to-speech) spočívá v „ozvučení“ libovolného textu, který se objevuje na vstupu systému syntézy řeči (tzv. „syntetizéru řeči“). V případě projektu ELJABR II jsou tímto vstupním textem skryté titulky, jimiž jsou jednotlivé pořady ČT vybavovány (dochází tedy k ozvučení skrytých titulků – odtud termín „mluvicí podtitulky“). Více informací o skrytých titulcích uvedeme v kapitole 2. Stručný popis technologie syntézy řeči z textu bude uveden v kapitole 3. Případová studie pro vytváření doprovodné zvukové stopy je popsána v kapitole 4. Poznamenejme, že tímto způsobem se zvuková stopa vytváří plně automaticky, bez přítomnosti člověka-dabéra, a obsahuje pouze (syntetizovanou) řeč či dialog bez ostatních složek (hudba, ruchy apod.).

Z podobných projektů na jiných pracovištích zmíníme SubTTS [5], počítačovou aplikaci pro čtení titulků vyvíjenou ve Švédsku na univerzitě v Göteborgu ve spolupráci s nemocnicí Queen Silvia Children’s Hospital v Göteborgu. SubTTS také vytváří řeč z titulků pomocí TTS, na rozdíl od našeho systému se řeč nepřenáší spolu s televizním signálem, ale je syntetizována na straně uživatele. Používá se i jiný druh titulků než ten popisovaný v kapitole 2, a to konkrétně titulky ve formátech SUB či SRT. SubTTS cílí zejména na lidi, kteří mají problémy se čtením titulků (cizojazyčných) filmů. Sledování takových filmů je přitom možné pouze na počítači. Projekt SubTTS nijak neřeší problémy synchronizace syntetické řeči s obrazovou scénou, popisované dále v kapitole 4.

2. Skryté titulky

Efektivní možností, jak spolu s obrazovým a zvukovým signálem odděleně přenášet i titulky (a přitom je „ukrýt“, dokud si je divák sám nezobrazí), je využití teletextového signálu. Tímto způsobem je možné poskytnout titulky cílové skupině – divákům se sluchovým postižením [3] – jako volitelnou součást televizního vysílání. Česká televize titulky vysílá na teletextové stránce 888.

Pro ukládání a přenos skrytých titulků se používá binární datový formát definovaný Evropskou vysílací unií (European Broadcasting Union, zkr. EBU) podle doporučení Technical Reference 3264-E [6]. Soubory mají příponu `.stl` a obsahují vždy jeden GSI (General Subtitle Information) blok následovaný řadou TTI (Text and Timing Information) bloků. GSI blok obsahuje celkovou informaci o pořadu, např. původní a přeložený název pořadu či jeho epizody, původní jazyk pořadu, jméno autora a další spíše technické informace týkající se vysílání a zobrazení titulků. Každý TTI blok definuje jeden titulek – vlastní text titulku, počáteční a koncový čas titulku (tj. čas zobrazení titulku), pozici titulku v obraze atd. Současná verze skrytých titulků neobsahuje žádnou informaci o přiřazení jednotlivých titulků postavám televizního pořadu.

3. Použitá technologie

K automatickému vytváření doprovodné zvukové stopy televizních pořadů jsme využili technologii syntézy řeči z textu (TTS), konkrétně TTS systém ARTIC [7] vyvíjený na KKY FAV ZČU v Plzni ve spolupráci s firmou SpeechTech. Úkolem systému TTS je „ozvučit“ text, tj. ze vstupního textu vygenerovat řeč. V našem případě je tedy řeč vytvářená ze vstupních skrytých titulků televizního vysílání.

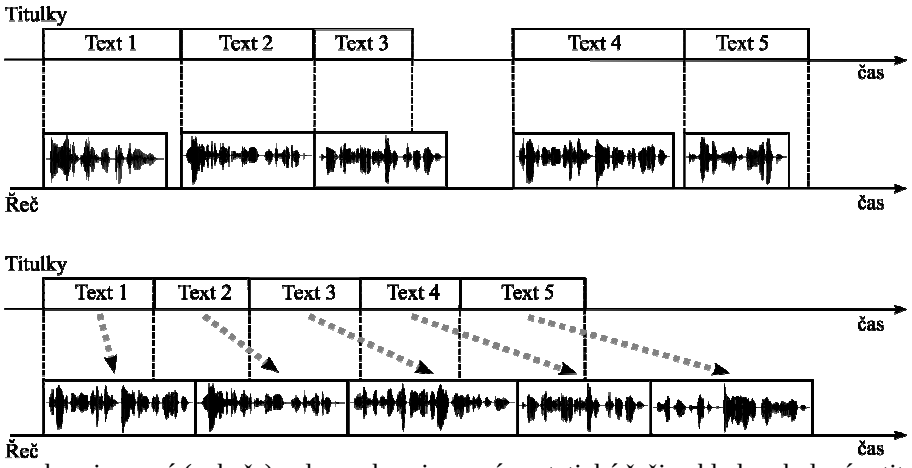
V rámci projektů ELJABR a ELAJBR II byly zatím vytvořeny 4 syntetické hlasy, 2 mužské (AJ-M a JS-M) a ženské hlasy (MR-Ž a KI-Ž). Všechny čtyři hlasy je možné využít pro automatické vytváření doprovodné zvukové stopy televizního vysílání.

4. Řešení problémů s aplikací technologie TTS

4.1 Desynchronizace mezi obrazovým a zvukovým signálem

Zásadním problémem při přímé aplikaci technologie TTS na ozvučování skrytých titulků je nutnost přesného „časového ukotvení“ vysyntetizované řeči – řeč odpovídající daným titulkům se musí „vejít“ do časových slotů těchto titulků. Vezmeme-li v potaz, že se obecný systém TTS obvykle nemusí starat o tempo syntetizované řeči a z principu zachovává řečové charakteristiky řečníka (vedle tempa řeči také např. hlasovou identitu, styl mluvy apod.), který pro potřeby syntézy řeči namluvil zdrojové řečové nahrávky, při přímé aplikaci tohoto systému lze očekávat, že řeč vygenerovaná na základě daného titulku bude „přesahovat“ časový slot

tohoto titulkem.[5] Následkem toho řeč zasahuje do následujícího titulku (nebo do jiné obrazové scény) a dochází tak k „desynchronizaci“ obrazového a zvukového signálu. Problém desynchronizace je ilustrován na obrázku 1.



Obr. 1: Ukázka synchronizované (nahoře) a desynchronizované syntetické řeči vzhledem k daným titulkovým slotům.

Problému desynchronizace obrazového a zvukového signálu je možné zabránit časovou kompresí (tj. zrychlováním) vytvářené řeči [8]. Výrazné zrychlování ale může prohloubit problémy s vnímáním zvukové stopy. Problematika spojená s desynchronizací, resp. se zrychlováním vytvářené řeči je shrnuta v tabulce 1. Pro každý hlas je zde uveden procentuální poměr desynchronizovaných titulků (tj. těch, které „přelézají“ předepsaný časový slot) a průměrnou velikost desynchronizace, a to pro každý titulek izolovaně („lokální desynchronizace“) a v kontextu okolních titulků („kumulativní desynchronizace“). „Zpoždění začátku titulku“ označuje titulky, jejichž začátek je vlivem desynchronizace zpožděn (opět je uveden procentní podíl takových titulků), a jejich průměrné zpoždění. „Faktor zrychlení“ představuje průměrný faktor, kterým je nutné vytvářenou řeč zrychlit, aby k desynchronizaci nedocházelo. Problematika desynchronizace je dále popsána v [9].

Hlas	Lokální desynchronizace		Kumulativní desynchronizace		Zpoždění začátku titulku		Faktor zrychlení
	Poměr titulků [%]	Průměr [s]	Poměr titulků [%]	Průměr [s]	Poměr titulků [%]	Průměr [s]	Průměr
AJ-M	34,51	0,63	43,30	2,65	31,30	3,24	1,23
MR-Ž	41,27	0,86	54,63	9,10	47,63	10,13	1,30
KI-Ž	45,11	0,76	56,44	5,44	41,97	6,76	1,28
JS-M	33,97	0,75	45,52	5,69	40,60	6,19	1,27

Tab. 1: Statistiky desynchronizace mezi obrazovou a doprovodnou zvukovou stopou pro různé syntetické hlasy. Statistiky byly spočteny na základě syntézy velkého počtu titulků (7 627 titulkových souborů o celkovém počtu 7 314 838 titulkových slotů odpovídající řeči celkové délce 5 458 hodin).

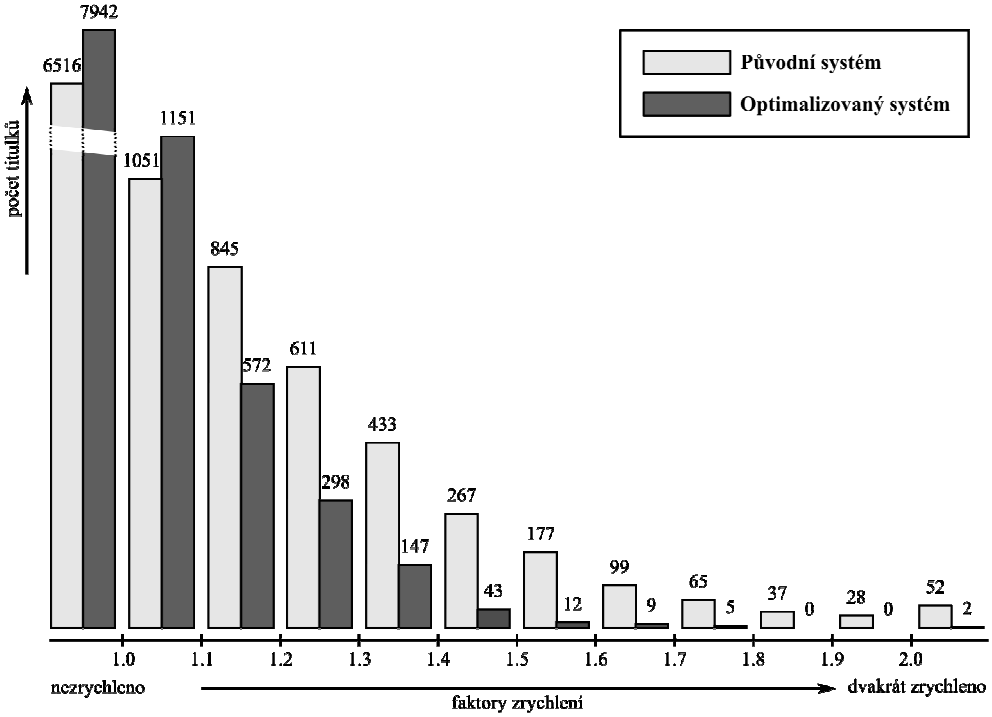
4.2 Optimalizace rozmístění titulků

Jak je z tabulky 1 patrné, desynchronizace představuje velký problém – bez urychlení by syntetická řeč odpovídající v průměru každému druhému titulku (podle kumulativní desynchronizace) nekorespondovala s obrazovým signálem (a to v závislosti na konkrétním použitém syntetickém hlasu v průměru až o 9 sekund!). Při reálném provozu systému automatického vytváření doprovodné zvukové stopy tedy bude často nutné vygenerovanou řeč zrychlovat. Zrychlování ale snižuje srozumitelnost syntetické řeči, což vzhledem k cílové skupině televizních diváků může představovat problém a je v přímém rozporu s cílem projektu.

V dalším výzkumu jsme se tedy zaměřili na možnosti optimalizace časového rozmístění titulků s cílem minimalizovat faktory zrychlování řeči při zachování synchronizace mezi doprovodnou zvukovou a obrazovou stopou televizního vysílání. Optimalizace vycházela z pružinového modelu – titulky každého pořadu byly

[5] Při přímé aplikaci systému TTS se zachovává zdrojové tempo řeči (kopírující časové charakteristiky zdrojového hlasu), které zaručuje nejvyšší kvalitu vytvářené řeči.

modelovány jako soustava pružin s různou tuhostí, která řeči vygenerované z titulků umožňovala se omezeným způsobem odchýlit od původních časových pozic. Při této optimalizaci byla prováděna i automatická detekce střihů v obrazovém signálu. Časové pozice střihů pak představovaly jakési „příčky“, přes něž nebylo možné (v zájmu zachování synchronizace mezi zvukovým a obrazovým signálem) vytvářenou řeč „natahovat“. Podrobnosti o zmíněných optimalizacích jsou uvedeny v [10]. Výsledky optimalizace na obrázku 2 ukazují, že po optimalizaci významně klesl počet titulků, které je nutné syntetizovat s vyšším faktorem zrychlení.



Obr. 2: Výsledky optimalizace rozmístění titulků vzhledem k faktorům zrychlování řeči pro 10 televizních pořadů různých žánrů.

Nežádoucímu zrychlování syntetické řeči je možné také zabránit tvorbou „jednodušších“ titulků (např. textů s jednodušší větnou stavbou, vynecháváním nepodstatné informace apod.).[6] Pro tento účel jsme navrhli poloautomatický postup – automatickou detekci existujících „problémových“ titulků (titulků, které by vedly na příliš rychlou a tudíž méně srozumitelnou syntetickou řeč) a ruční zjednodušení textu takových titulků. Do budoucna se počítá, že zjednodušování textu titulků i s ohledem na jejich ozvučení by mělo být součástí procesu přípravy nových titulků v ČT.

5. Případová studie

V roce 2012 byla provedena případová studie, v rámci níž bylo realizováno experimentální zapojení systému vytváření doprovodné zvukové stopy do testovacího vysílání České televize. Cílem studie bylo vyzkoušet a ověřit postupy navržené pro automatické vytváření doprovodné zvukové stopy televizního vysílání na reálném pořadu televizního vysílání ČT. Po domluvě s ČT byl pro testovací vysílání vybrán původní český 17dílný seriál Hraběnky[7]. Tento seriál je charakteristický svou „komplexní zvukovou stopou“ – zvukový signál obsahuje velké množství doprovodných ruchových zvuků a podkresové hudby. Testovací vysílání probíhalo v režimu „offline“, tj. mimo reálný čas televizního vysílání. Podkladem pro vytvoření doprovodné zvukové stopy každého dílu zmíněného seriálu byly skryté titulky z archivu ČT (ve formátu EBU `.stl`, popsaným v kapitole 2). Proces tvorby zvukové stopy zahrnoval výše popsané poloautomatické zjednodušování textu titulků a automatickou tvorbu doprovodné zvukové stopy současnou verzí navrhovaného systému vytváření doprovodné zvukové stopy (včetně optimalizací popsaných v kapitole 4).

Abychom mohli při vytváření doprovodné zvukové stopy smysluplně využít všechny čtyři syntetické hlasy, potřebovali jsme provést přiřazení jednotlivých syntetických hlasů jednotlivým postavám seriálu. Je přitom zřejmé, že se nejedná o lehkou úlohu – počet postav jednotlivých televizních pořadů značně převyšuje

[6] Protože jsou skryté titulky koncipovány výhradně pro osoby se sluchovým postižením, byly od počátku vytvářeny se zřetelem na jejich „snadné čtení v duchu“ zmíněnými osobami. Doslovný přepis zvukové stopy není požadován.

[7] http://www.ceskatelevize.cz/porady/10076692255-hrabenky

počet dostupných syntetických hlasů (například ve filmových pořadech se běžně vyskytují desítky postav). Situace byla navíc komplikována ještě tím, že ve zdrojových souborech skrytých titulků nebyla žádná informace o přiřazení jednotlivých titulků postavám (viz kapitola 2). Řešení tohoto problému probíhalo ve dvou fázích. V první fázi byla do skrytých titulků ručně doplněna informace o tom, který titulek „náleží“ které postavě. Ve druhé fázi pak bylo provedeno automatické přiřazení dostupných syntetických hlasů jednotlivým postavám. Hlavním kritériem pro automatické přiřazení bylo, aby se minimalizovalo nebezpečí, že ve stejné dialogové scéně dvě různé postavy mluví stejným hlasem. Výsledné statistiky týkající se syntetizované zvukové stopy jsou uvedeny v tabulce 2.

Celkový počet titulkových souborů	17
Celkový počet vysyntetizovaných promluv	15 512
Celkový počet vysyntetizovaných titulkových slotů	10 427
Celková délka vysyntetizované řeči v počtu slov	55 807
Celková doba vysyntetizované řeči (h:mm:ss)	8:19:13
Poměr zrychlených titulků	30,18 %
Průměrný faktor zrychlení zrychlených titulků	1,12
Poměr titulků, v němž mluví 2 různé postavy stejným hlasem	1,30 %
Poměr textově zjednodušených titulků	6,38 %

Tab. 2 Souhrnné informace o řeči syntetizované v rámci případové studie.

6. Závěr

Tento příspěvek popisuje poslední vývoj v oblasti automatického vytváření doprovodné zvukové stopy televizního vysílání, určené sluchově postiženým divákům České televize. Problematika je řešena v rámci projektu ELJABR II na Katedře kybernetiky FAV ZČU v Plzni ve spolupráci s Českou televizí a firmou SpeechTech. Funkčnost navrhovaného řešení, systému pro automatické vytváření doprovodné zvukové stopy televizního vysílání na základě skrytých titulků televizních pořadů, byla ověřena v rámci případové studie, během níž byla doprovodná zvuková stopa úspěšně vytvořena pro 17dílný seriál České televize Hraběnky.

V další práci se chceme zaměřit zejména na vytváření doprovodné zvukové stopy v režimu „online“, tj. v reálném čase vysílání televizního pořadu. Vedle „technického“ vyhodnocení systému vytváření doprovodné zvukové stopy (prezentovaného v tabulkách 1 a 2 a na obrázku 2) plánujeme rovněž i evaluaci systému z pohledu samotných uživatelů – diváků s mírným sluchovým postižením.

Poděkování

Projekt „Eliminace jazykových bariér handicapovaných diváků České televize II“ (ELJABR II), č. TA01011264, je řešen s finanční podporou Technologické agentury České republiky (TA ČR). Projekt ELJABR, č. 2C06020, byl řešen za podpory Ministerstva školství, mládeže a tělovýchovy (MŠMT). Zvláštní poděkování za spolupráci patří rovněž České televizi.

Literatura

[1] Gazdík, M.: Nové přístupové služby digitální televize. Sborník z konference INSPO 2011, str. 2-5, 19.3. 2011, Praha, Česká republika.

[2] Müller, L.: Automatické titulkování živých pořadů České televize – současný stav a výhled do budoucna. Sborník z konference INSPO 2012, 17.3. 2012, Praha, Česká republika.

[3] Salzman, V.: Současný stav a záměry České televize pro další zpřístupňování veřejnoprávního vysílání sluchově postiženým divákům. Sborník z konference INSPO 2010, str. 64-66, 13.3. 2010, Praha, Česká republika.

[4] Psutka, J., Müller, L., Matoušek, J., Radová, V.: Mluvíme s počítačem česky. Academia, Praha, 2006.

[5] Derbring, S., Ljunglöf, P., Olsson, M.: SubTTS: Light-Weight Automatic Reading of Subtitles. Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA), pp. 272-274, 14.-16.5. 2009, Odense, Denmark.

[6] EBU Tech 3264-1991. Specification of the EBU Subtitling data exchange format. European Broadcasting Union, February 1991.

[7] Matoušek, J., Tihelka, D., Romportl, J.: Current State of Czech Text-to-Speech Systém ARTIC. Text, Speech and Dialogue, Lecture Notes in Artificial Inteligence, vol. 4188, Springer, Berlin, 2006, pp. 439-446.

[8] Tihelka, D., Méner, M.: Generalized Non-Uniform Time Scaling Distribution Method for Natural-Sounding Speech Rate Change. Text, Speech and Dialogue, Lecture Notes in Artificial Inteligence, vol. 6836, Springer, Berlin, 2011, pp. 147-154.

[9] Hanzlíček, Z., Matoušek, J., Tihelka, D.: Towards Automatic Audio Track Generation for Czech TV Broadcasting: Initial Experiments with Subtitles-to-Speech Synthesis. Proceedings of IEEE International Conference on Speech Processing, pp. 2721-2724, 2008, Beijing, China.

[10] Matoušek, J., Vít, J.: Improving Automatic Dubbing with Subtitle Timing Optimisation Using Video Cut Detection. Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2385-2388, 2012, Kyoto, Japan.