INSPO – informační a komunikační technologie pro osoby se specifickými potřebami
30. března 2019, Kongresové centrum Praha

SKRYTÉ TITULKY K ŽIVÝM TELEVIZNÍM POŘADŮM

Aleš PRAŽÁK, Josef PSUTKA
Západočeská univerzita v Plzni a SpeechTech, s.r.o., ales.prazak@speechtech.cz

Anotace: *Skryté titulky k živým televizním pořadům jsou dnes vytvářeny s využitím počítačového systému pro přepis mluvené řeči. Tato technologie však neumožňuje zcela automatickou tvorbu titulků v dostatečné kvalitě. Z toho důvodu je využíván lidský prostředník, tzv. stínový řečník, jehož prostřednictvím je možno v reálném čase produkovat téměř bezchybné živé titulky. Živé titulky však mají oproti předem připraveným titulkům další odlišnosti – např. zpoždění, či jiné formátování. Cílem příspěvku je seznámení s postupem výroby titulků k živým televizním pořadům včetně řešení souvisejících problémů.*

Úvod

Titulkování živých televizních pořadů je ve světě čím dál tím více žádáno společností neslyšících a nedoslýchavých pro zpřístupnění televizního obsahu osobám se sluchovým postižením a starším lidem. Televizní titulky by měly zvukový obsah televizního vysílání doručit sluchově postiženým osobám ve stejném rozsahu, v jakém je tento obsah dostupný lidem slyšícím. Obdobný požadavek je zakotven v mnoha národních i nadnárodních legislativách. Na rozdíl od bezchybných a perfektně časovaných titulků připravených předem, živé titulkování představuje mnohem složitější proces s nedokonalými výstupy. Od množství živě titulkovaných pořadů, které se dnes v některých státech blíží stoprocentnímu pokrytí (např. ve Velké Británii nebo Švýcarsku), se pozornost přesunuje na kvalitu živých titulků – konkrétně zpoždění a přesnost živých titulků jsou dnes nejdiskutovanější (Romero-Fresco, 2016).

Vývoj v oblasti automatického rozpoznávání řeči v poslední době, zejména díky nárůstu výkonu výpočetní techniky, umožnil využití rozpoznávání řeči v reálném čase jako levnější a masivněji aplikovatelnou variantu k živému titulkování pomocí ručního přepisu. Tato technologie však stále není schopna přepsat přímo originální zvukovou stopu libovolného televizního pořadu v dostatečné kvalitě tak, aby mohl být výsledek automatického přepisu vysílán ve formě titulků pro osoby se sluchovým postižením. Z toho důvodu je při živém titulkování využíván prostředník, tzv. stínový řečník (anglicky respeaker), který originální dialogy či komentář televizního pořadu přemlouvá v tichém prostředí trénovaným způsobem do systému rozpoznávání řeči, který jeho projev přepisuje do titulků. Titulkování živých televizních pořadů s využitím stínového přemlouvání bylo zavedeno v BBC v roce 2003 (Evans, 2003) a nyní se na různé úrovni používá celosvětově.

Vývoj technologie

Technologie titulkování živých televizních pořadů vznikla v souvislosti s řešením dvou projektů na Západočeské univerzitě v Plzni, Fakultě aplikovaných věd, Centru NTIS a Katedře kybernetiky ve spolupráci s firmou SpeechTech. Jednalo se o projekt ELJABR (ELiminace JAzykových BaRiér handicapovaných diváků České televize) řešený v letech 2006-2011 za podpory MŠMT a projekt ELJABR II řešený v letech 2011-2016 za podpory TA ČR. Jedním z cílů projektů byla tvorba systému výroby živých titulků pro různé typy televizních pořadů. Potřeba řešení vznikla na základě zákona č. 132/2010 Sb. (Zákon o audiovizuálních mediálních službách na vyžádání), podle kterého má Česká televize, jako provozovatel celoplošného televizního vysílání ze zákona, povinnost opatřit alespoň 70 % vysílaných pořadů skrytými nebo otevřenými titulky. Se zavedením nových kanálů ČT24 a ČT sport, s převážně živými pořady, začala Česká televize hledat technické prostředky pro plnění zákonného požadavku, přičemž byla zaujata přístupem používaným v BBC, které bylo schopno vysílání na několika kanálech opatřovat titulky 24 hodin denně. Řešení obou dvou zmiňovaných projektů i navazujícího projektu TVPOTAR (Vývoj pokročilých přístupů k vytváření titulků a archivaci TV pořadů a dokumentů) řešeného v letech 2016-2019 Česká televize podpořila finančně ze svých neveřejných zdrojů.

Na základě našich zkušeností s přepisem řeči v reálném čase na omezených doménách (např. diktovací aplikace MegaWord) jsme ověřovali i možnosti plně automatického živého titulkování, tedy bez stínového přemlouvání. Podle našich experimentů lze dosáhnout přijatelné přesnosti přepisu pouze u programů s profesionálními mluvčími, kteří si neskáčou do řeči a mluví v tichém prostředí. Česká televize nyní využívá plně automatické živé titulkování pro přenosy z jednání schůze Poslanecké sněmovny a Senátu Parlamentu ČR.

Díky velmi specifickému zaměření systém dosahuje slovní chybovosti pod 6 %. Pro zlepšení čitelnosti jsou do výsledných titulků automaticky doplňována interpunkční znaménka. Od prvního živého titulkování v roce 2008 bylo tímto způsobem živě otitulkováno více než 4 000 hodin záznamů z jednání schůze PS a Senátu PČR vysílaných Českou televizí.

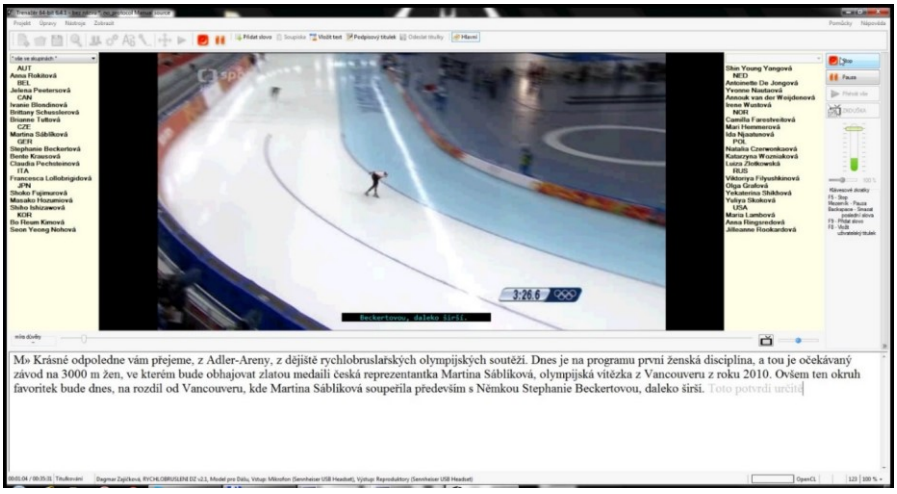
Stínové přemlouvání

Stínové přemlouvání je komplexní technika pro přepis nedokonalé slovní informace do formy srozumitelných a gramaticky korektních titulků. Stínový řečník může přemlouvat slovo od slova, pokud je originální řečový projev v textové podobě pro diváka dostatečně jasný a čitelný. Pokud si však více mluvčích skáče do řeči nebo mluví nesouvisle, doslovný přepis by pro diváka nebyl srozumitelný. V tomto případě je úkolem stínového řečníka přeformulovat či zkondenzovat původní slovní projev do jasných a gramaticky korektních vět se stejným významem. Stínové přemlouvání sportovních pořadů má navíc oproti pořadům zpravodajským či zábavním určitá specifika. Z důvodu zpoždění živých titulků mohou být některé informace irelevantní ve chvíli, kdy jsou ve formě titulků zobrazeny divákovi. Jedná se např. o pojmenování krasobruslařských skoků či o popis přihrávek mezi fotbalovými a hokejovými hráči. Navíc nadbytečné titulky odvádí pozornost diváka od samotného sportu. V tomto případě je tedy úkolem stínového řečníka i filtrování nepodstatných informací (např. držení hokejového puku je viditelné), avšak poskytování důležitých a zajímavých informací v podobě, která nebude rozptylovat divákovu pozornost.

Na rozdíl od zahraničních systémů, kde prakticky neexistuje možnost v průběhu diktování do procesu přepisu řeči zasahovat, a tím se vypořádat s nedokonalostí těchto systémů, systém rozpoznávání řeči vyvíjený na ZČU ve spolupráci s firmou SpeechTech je se stínovým řečníkem úzce propojený (Pražák, Loose, Trmal, Psutka, & Psutka, 2012). Stínový řečník tak v průběhu živého titulkování kromě samotného přemlouvání:

- opravuje případné chyby v přepisu – vzhledem k minimálnímu zpoždění přepisu slov po jejich vyřčení (do půl sekundy) u našeho systému je stínový řečník schopen případnou chybu v přepisu ihned odstranit tím, že chybu smaže a text přemluví znovu.
- přidává nová slova do slovníku – k chybě v přepisu dochází často v případě, že stínový řečník vysloví slovo (např. cizí jméno nebo název), které systém rozpoznávání nezná. Naše technologie umožňuje stínovému řečníkovi nové slovo okamžitě přidat do systému rozpoznávání, přičemž ho může dále běžně vyslovovat a chyby v přepisu tak neopakovat.
- vkládá interpunkci pomocí klávesnice – v dosavadních systémech pro živé titulkování je nutno interpunkční znaménka (zejména tečku, čárku a otazník) diktovat, což obzvlášť v případě češtiny, která používá větné čárky často, zpomaluje proces živého titulkování. V naší technologii stínový řečník stiskem příslušné klávesy předá systému rozpoznávání řeči informaci o interpunkčním znaménku, které se objeví ve výsledném přepisu na správném místě. Stejný postup je použit i pro označování změn řečníků, na jejichž základě jsou barevně rozlišeny výsledné titulky.

Vzhledem k tomu, že práce stínového řečníka je velmi náročná a jeho trénink trvá několik měsíců, byl na Západočeské univerzitě v Plzni ve spolupráci s firmou SpeechTech vyvinut speciální trenažér (trénovací aplikace), který postupně, ve čtyřech fázích, stínového řečníka připravuje na reálné živé titulkování.

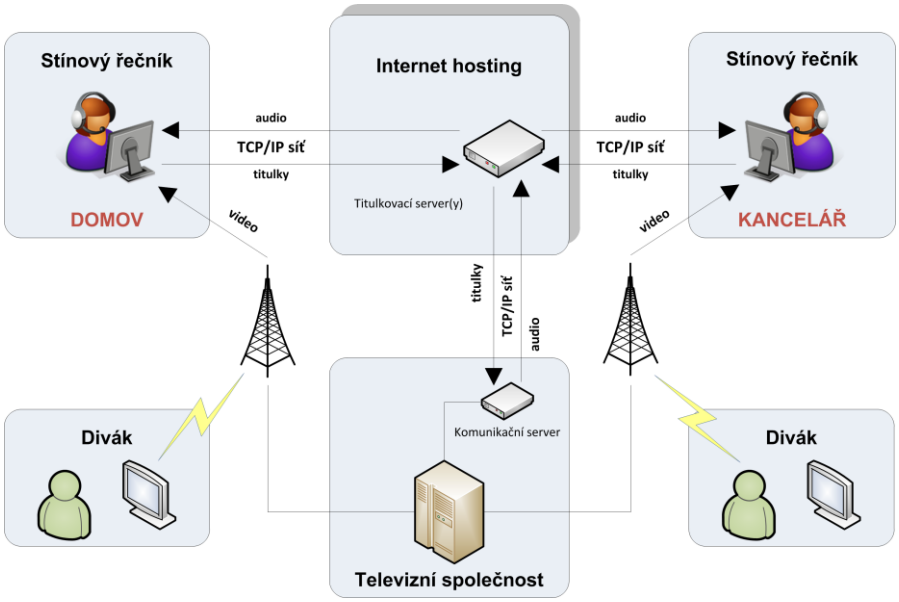


Obr. 1: Titulkovací aplikace z pohledu stínového řečníka

Vzdálené titulkování

Stínový řečník je schopen náročnou práci v nejvyšší kvalitě vykonávat souvisle maximálně dvě hodiny (záleží i na typu titulkovaného pořadu), poté by měl být vystřídán. Přitom živé přenosy, které potřebuje Česká televize titulkovat s využitím uvedené technologie, jsou vysílány v průběhu celého dne, i v noci. Z toho důvodu by bylo velmi neekonomické, aby stínoví řečníci dojížděli do studií České televize na titulkování několikrát denně, popř. tam mezitím odpočívali. Celá technologie tak byla od začátku navržena pro vzdálený provoz, tedy živé titulkování např. z domova prostřednictvím internetu.

Komunikace mezi stínovým řečníkem a Českou televizí probíhá prostřednictvím speciálních protokolů a serverů, a to oběma směry. Směrem do České televize jsou od stínového řečníka přenášeny hotové živé titulky ihned, jak jsou vytvořeny. Protože ale z principu lze vytvořit živý titulek až ve chvíli, kdy byla vyřčena všechna slova titulku, živé titulky u diváka jsou oproti originálnímu zvuku zpožděny. Pro synchronizaci živých titulků se zvukem by bylo zapotřebí, aby stínový řečník zvuk pořadu obdržel s předstihem minimálně 10 sekund, což je v běžných podmínkách technicky nerealizovatelné. Z toho důvodu je stínovému řečníkovi prostřednictvím internetu dodáván zvuk přímo z České televize s alespoň technologicky dosažitelným předstihem 3 sekund, který zajistí, že výsledné živé titulky jsou u diváka zpožděny v průměru jen o 5-6 sekund. Tato hodnota zpoždění živých titulků je jedna z nejnižších na světě, což vyplývá i z integrace schopností stínového řečníka, kdy není potřeba spolupracující korektor, který zpoždění živých titulků zvyšuje. Obraz přijímaný běžným televizním vysíláním slouží stínovému řečníkovi jen pro přehled o dění v titulkovaném pořadu.



Obr. 2: Schéma technologie vzdáleného titulkování živých televizních pořadů

Vzhledem k tomu, že generování živých titulků z rozpoznaného textu je plně automatické (v souladu s vysílacími standardy a typografickými pravidly), živé titulky mohou být jak blokové (jedno- či víceřádkové), tak rolující (postupně se objevující po slovech či řádcích). Podle výzkumu využívajícího technologii sledování očí (Romero-Fresco, 2015), rolující titulky nutí diváka strávit mnohem více času sledováním titulků, než obrazu, v porovnání s titulky blokovými se stejnou rychlostí (ve smyslu slov za minutu). Z toho důvodu náš systém generuje dvouřádkové blokové titulky s výjimkou sportovních programů, kde jsou z důvodu překrytí menší části obrazovky generovány titulky jednořádkové. V zahraničí jsou pro živé titulkování obvykle využívány titulky rolující (po slovech), a to z důvodu snížení zpoždění výsledných titulků (jednotlivá slova se objevují ihned, není třeba čekat na celé věty).

Systém rozpoznávání řeči

Vzhledem k velké slovní zásobě češtiny jako flektivního jazyka (oproti angličtině je potřeba až 8x více slovních tvarů) je velikost slovníku systému automatického rozpoznávání řeči jedním z jeho klíčových parametrů. Náš systém je schopen pracovat se slovníkem až 1,5 milionu slov v reálném čase. Systém pracující v reálném čase s takto velkým slovníkem je poměrně unikátní, nicméně pro dosažení maximální kvality vytvářených titulků s ohledem na specifickou slovní zásobu jednotlivých televizních pořadů (zejména sportovních) je nutno systém připravit pro každý typ pořadu zvlášť. Protože však nelze uvažovat o tom, že by

bylo možné slovník systému předem doplnit např. jmény všech sportovců na světě (a to ve všech gramatických pádech a slovních tvarech), je dále nutno systém rozpoznávání řeči adaptovat na každý jednotlivý živý sportovní pořad těsně před jeho titulkováním. Pro každý sportovní pořad je připravena soupiska jmen sportovců, názvů týmů a sportovišť. Pro češtinu je navíc nutno všechna jména a názvy vyskloňovat, v případě mužských jmen a přivlastňovacích tvarů se tak jedná až o 30 různých slovních tvarů pro jednoho sportovce. Pro cizí jména je navíc nutno definovat jejich výslovnosti, které se neřídí běžnými českými pravidly. To vše je maximálně zautomatizováno s minimální lidskou účastí. Před samotným živým titulkováním jsou pak vybraná jména stínovým řečníkem vložena do systému rozpoznávání řeči tak, aby je stínový řečník mohl bez problémů užívat a jména v různých tvarech se správně objevovala ve výsledných titulcích. Slovník systému je dále automaticky dvakrát denně doplňován o nová slova a výrazy vyskytující se v aktuálním internetovém zpravodajství.

Důležitým rozdílem naší technologie oproti zahraničním systémům pro živé titulkování je schopnost připravit systém rozpoznávání řeči na míru nejen titulkovanému televiznímu pořadu, ale i konkrétnímu stínovému řečníkovi. To má za následek nezanedbatelné zvýšení přesnosti přepisu, a tím i možnost využívat pouze schopnosti stínového řečníka bez spolupracujícího korektora. Systém rozpoznávání řeči je učen (trénován) na velkém množství zvukových a textových dat. Čím obecnější jsou tato data, tím širší využití systém má, ale tím horší výsledky podává na konkrétní úloze. V případě stínového řečníka je tedy žádoucí naučit systém pouze na tohoto řečníka. Z toho důvodu byl vyvinut postup pro automatický trénink systému, kdy jsou všechny trénovací nahrávky řečníka (minimálně 100 hodin) automaticky zpracovány a následně je vytvořen individuální rozpoznávací systém pro každého stínového řečníka.

Eliminace zpoždění živých titulků

Ačkoliv jsou živé titulky ze svého principu u diváka zpožděny, mnoho živých televizních pořadů je reprízováno. Živě vyrobené titulky mohou být bez dalších nákladů využity i při reprízách, nicméně v tomto případě nemusí být vysílány se zpožděním. Nejjednodušším postupem je paušální posunutí všech živě vyrobených titulků o několik sekund dříve. Tím se sice eliminuje průměrné zpoždění všech titulků, ale jednotlivé titulky budou zobrazeny o něco dříve či později v závislosti na původním zpoždění každého jednotlivého živého titulku (to je dáno zpožděním stínového řečníka, dobou zpracování a zejména délkou titulku). Pro perfektní časování titulků je tak potřeba automaticky namapovat jednotlivé titulky na originální řečový projev. Pro tento účel lze využít metody zpracování zvuku. V případě doslovných titulků se používá známá metoda, která je ale velmi náchylná na chyby v přepisu a v případě větších nepřesností selhává. Pro případ živých titulků, jejichž obsah je oproti původnímu řečovému projevu záměrně upravován, bylo potřeba vyvinout jinou metodu mapování titulků na zdrojový audiovizuální materiál. Tato metoda je založena na fonémovém přepisu originálního zvuku, který je následně speciálním postupem zarovnán na živě vyrobené titulky. Metoda si poradí jak se změnou slovosledu, tak s celými chybějícími větami a dalšími odchylkami textového přepisu řeči. Tento postup lze s výhodou využít i v průběhu přípravy offline titulků, kdy může nahradit časově náročné ruční časování titulků či jeho kontrolu.

Pro úplnou eliminaci zpoždění při zobrazení živých titulků by bylo nutno získat zvukovou stopu pořadu z televizní společnosti s předstihem minimálně 10 sekund oproti vysílání, v technické praxi to však znamená zpoždění samotného vysílání tak, aby se titulky stihly vyrobit. K tomuto kroku bohužel žádná televizní společnost na světě nechce přistoupit. Řešením je přesun eliminace zpožďování ze strany televizní společnosti na stranu diváka, konkrétně jeho televizního přijímače. Tímto způsobem je možno vyřešit synchronizaci živě vyrobených titulků s originálním zvukem pořadu jen u diváků, kteří o titulky mají zájem (dívali by se tedy na vysílání se zpožděním několika sekund), přičemž běžní diváci by nebyli nijak dotčeni. I zde je využita metoda pro mapování jednotlivých titulků na originální řečový projev, neboť každý jednotlivý titulek může mít z principu jiné zpoždění a nelze tedy paušálně pozdržet obraz a zvuk pořadu s tím, že by se titulky zobrazovaly správně.

Pro dokonalou synchronizaci se zvukem je nutno každý titulek zobrazit v přesně určený časový okamžik vysílaného pořadu. Informaci o zpoždění každého živého titulku je pak nutno přenést k divákovi. Toto je možno provést v případě teletextových titulků zakódováním dodatečné informace do nezobrazované části teletextové stránky. Vzhledem k přechodu na DVB vysílání bez teletextu, kde je přenos potřebné informace k divákovi technicky mnohem složitější, byla jako komunikační prostředek pro přenos informace o zpoždění živých titulků k divákovi využita internetová síť. Klientská část řešení je implementována ve formě aplikace pro „chytrý“ televizor pracující na platformě Android TV.

Cílem aplikace je pozdržení přijímaného obrazu a zvuku (z libovolného zdroje vysílání) o konstantní dobu, zatímco živé titulky přijímané přes internet jsou pozdrženy méně. Tímto způsobem je možno eliminovat zpoždění živých titulků a v případě určení přesného zpoždění každého jednotlivého titulku jsou tyto titulky

zobrazovány synchronně s obrazem a zvukem, tedy při vyřčení prvního slova titulku. Pro pozdržení obrazu a zvuku je využita funkce TimeShift televizního přijímače.

V případě, kdy není sledovaný kanál živě titulkován, je zobrazeno originální televizní vysílání bez zpoždění. Jakmile je zahájeno živé titulkování sledovaného pořadu, je automaticky zapnuta funkce TimeShift, která na konstantní dobu zastaví přehrávání a následně přehrává vysílání se zpožděním. Pro zachování kontinuity sledovaného kanálu není funkce TimeShift automaticky vypnuta po skončení živého titulkování sledovaného pořadu, ale je ukončena až při přepnutí sledovaného kanálu. Živé titulky jsou tak zobrazovány synchronně se zpožděným obrazem a zvukem pořadu. Způsob zobrazení živých titulků je přizpůsoben zvyklostem při zobrazování teletextových titulků, na rozdíl od klasických skrytých titulků je ale možno většinu parametrů zobrazení uživatelsky nastavit.

Závěr

Popsaná technologie vzdáleného titulkování živých televizních pořadů umožňuje díky implementaci state-of-the-art technologií a mnoha inovacím výrobu živých titulků prostřednictvím jednoho stínového řečníka pracujícího vzdáleně (např. z domova) a dosahujícího srovnatelné kvality živých titulků jako v zahraničí, a to 98 % podle světově používaného modelu NER (Romero-Fresco & Pérez, 2015). Technologie je dále vyvíjena za účelem zvyšování kvality živých titulků a souvisejících procesů.

Technologie vzdáleného titulkování živých televizních pořadů je v současné době využívána pro titulkování zpravodajských (Otázky Václava Moravce, Studio 6, Interview ČT24, 90' ČT24, Volební studia a další), zábavních (Sama doma, StarDance a další) a zejména živých sportovních pořadů vysílaných Českou televizí. Od prvního experimentální živého titulkování v roce 2010 bylo touto technologií živě otitulkováno více než 15 000 pořadů. Aktuální pořady České televize živě titulkované popsanou technologií lze nalézt na internetové adrese www.zivetitulky.cz.

	2013	**2014**	**2015**	**2016**	**2017**	**2018**
Zpravodajství	370	517	615	721	763	819
Sport	220	782	1921	3260	3228	3207
Zábava	22	44	278	282	722	983
Celkem	612	1343	2814	4263	4713	5009

Tab. 1: Množství živě titulkovaných hodin vysílaných Českou televizí pro různé typy pořadů

Literatura

Romero-Fresco, P. (2016). Accessing communication: The quality of live subtitles in the UK. Language & Communication, 49, str. 56-69.

Evans, M. J. (2003). WHP 065. British Broadcasting Corporation.

Pražák, A., Loose, Z., Trmal, J., Psutka, J. V., & Psutka, J. (2012). Novel approach to live captioning through re-speaking: Tailoring speech recognition to re-speaker's needs. INTERSPEECH (str. 1372-1375). ISCA.

Romero-Fresco, P. (2015). The reception of subtitles for the deaf and hard of hearing in Europe. Berlin: Peter Lang.

Romero-Fresco, P., & Pérez, M. (2015). Accuracy rate in live subtitling: The NER model. In R. B. Piñero, & J. D. Cintas, Audiovisual Translation in a Global Context (str. 28-50). Palgrave Macmillan.