Analytika mluveného slova

Český rozhlas (ČRo) hodlá zpřístupnit některé obsahy vysílání v textové podobě pro snadné hledání i sledování obsahu živého vysílání v textové podobě. Pro tento účel je nutné vyvinout Analytiku mluveného slova (AMS) rozepsanou níže tak, aby bylo možné ve stejné kvalitě zpracovat jak živé vysílání, tak archiv kontinuálního vysílání kterým ČRO disponuje.

ČRo hodlá AMS zavést nejprve pro živé vysílání a to v reálném čase dle technicko-technologických možností s latencí transkripce v řádu sekund či jejich nižších desítek v případě identifikace. Takto zpracovávané vysílání bude automaticky ukládáno, kategorizováno a doplněno metadaty a může být po skončení každého z pořadu znovu důkladněji analyzováno.

ČRo dále disponuje archivem kontinuálního vysílání i jeho základním popisem metadaty u některých stanic už od roku 2003. Záměrem je archiv kontinuálního vysílání analyzovat a převést veškeré mluvené slovo na text (transkripce), rozlišit mluvčí (diarizace), identifikovat mluvčí, promluvy přiřadit k jejich profilům a indexovat je pro snadné vyhledávání. Tento archiv bude navazovat na zpracovávané živé vysílání.

Třetím předmětem zakázky je dodání systému pro on-demand AMS v plné i zcela jednoduché podobě.

AMS je druhou polovinou celkové analýzy kontinuálního vysílání, jejíž první částí je Analytika vysílání (AV), ve které se zvuková stopa člení na jednotlivé úseky a tyto označuje. Celkovým výsledkem, ve kterém dochází k propojení obou částí, tedy AMS a AV, je podrobný popis kontinuálního vysílání, tedy včetně transkripce s případnou identifikací mluvčích ve vysílání a to všech stanic ČRo. A tento výstup, tedy jak zpracovaný archiv tak online vytvářený archiv přímo ze živého vysílání, bude zpřístupněn dle práv uživatelům na internetu v rámci projektu mujROZHLAS.

mujROZHLAS je strategickým projektem ČRo a bude největším českým audio archivem na internetu. Předpokládané období pro spuštění první verze projektu je plánováno na jaro 2019.

Český rozhlas nehledá pouze dodavatele, ale aktivního a spolehlivého partnera pro vývoj jedinečného produktu, kterému půjde o dosažení co nejlepšího možného výsledku a nezalekne se nesnadného vývoje.

Základní informace a stručný popis zakázky

Zakázka obsahuje:

1. Vývoj, dodání, instalaci a aktualizaci systému AMS, čímž je míněna transkripce, diarizace a identifikace, výroba speciálních slovníků a voiceprintů rozepsaná podrobněji níže, pro analýzu
    - živého vysílání,
    - on-demand audiosoborů,
    - archivu kontinuálního vysílání.
2. Hromadné zpracování archivu kontinuálního vysílání systémem AMS na straně dodavatele (ca 100 TB audiosouborů v MP3).
3. Servis systému po dobu trvání zakázky.
4. Základní informace k AMS živého vysílání, on-demand a archivu

5. AMS automaticky rozpoznává mluvené slovo (ASR), identifikuje jazyk (čeština, slovenština, ostatní – tedy dva a ostatní zatím nerozlišuje), provádí transkripci a přiřazuje timecode tak, aby bylo možné přistupovat ke konkrétnímu úseku zvukové stopy dle času.
6. Systém dále rozlišuje mezi mluvčími a pohlavím (muž/žena) v případě neidentifikování mluvčího, a pokud je to možné, identifikuje mluvčí tak, aby ČRo mohla přiřadit promluvu k hlasovému profilu osoby. Profil by měla každá osoba promlouvající ve vysílání, i když může jít o anonymní profil, který lze posléze identifikovat.
7. Pro vybrané osoby (v řádu jednotek, např. Miloš Zeman, Andrej Babiš) připraví dodavatel personalizované slovníky takové, aby přesnost transkripce byla v rozmezí 95-100 %. V rámci technických možností připraví dodavatel ve spolupráci s ČRo specializované – tematické – slovníky, např. pro vědecké pořady ČRo, případně modul pro sníženou kvalitu hlasu (telefonát, rušné místo, sportovní přenos).
8. Tento systém může používat pro identifikaci mluvčích jak obvyklá schémata pořadů a jejich tvůrců dostupná z metadat popsaných níže, tak samotnou transkripci, ve kterém se obvykle identifikuje následující mluvčí v promluvě. Základem pro identifikaci bude digitální otisk hlasu, tzv. voiceprint, které bude dodavatel rovněž vytvářet.
9. Provozování systému pro živé vysílání i pro on-demand požadavky bude na serverech ČRo.
10. Hromadné zpracování archivu kontinuálního vysílání provede dodavatel na svých (pronajatých) zařízeních. Data pro AMS archivu dodavatel převezme osobně na svých zařízeních (ca 100 TB dat).
11. Pro jednotlivé nebo menší celky připraví dodavatel pro ČRo on-demand systém, který umožní opětovné zavedení výsledků analýzy do databází ČRo. On-demand AMS má funkce volitelné, tedy je možné požadovat např. pouze transkripci.
12. Nákladnost na nákup HW, příslušného OS a implementaci je součástí hodnocení.
13. Podrobný popis níže popisuje i části, které dodavatel nevyvíjí, ale např. pouze poskytuje data pro jejich plnění.

K čemu transkripci a identifikaci osob ČRo využije?

Indexace obsahu

Indexace a vyhledávání v audiu – pro interní a kategorizační / archivní účely i pro vyhledávače na internetu a jeho uživatele. Indexaci a vyhledávání řeší ČRo jinou veřejnou zakázkou.

Sledování vysílání i bez zvuku

Posluchači umožní sledovat přepis online v případě, když nemůže poslouchat. Umožní tak nepřetržité sledování informací v rámci možností kvality AMS. Online sledováním se míní přepis mluveného slova se zpožděním v sekundách od odvysílání s diarizací a identifikací mluvčího atd.

Indexace obsahu vyhledávači.

Současné vyhledávače neumí prohledávat zvukovou stopu, transkripce jim umožní přečíst si obsah audia a nechat jej zaindexovat. ČRo tak zpřístupní vybrané obsahy svého vysílání širší veřejnosti i prostřednictvím vyhledávačů.

Tematizace obsahu

Dalším krokem je automatická tematizace / štítkování obsahu, které umožní rychlejší kategorizaci.

Automatická shrnutí

Transkripce bude časem sloužit pro vytváření automatických shrnutí obsahu - sumarizace.

Speech to text

Jakmile bude možné vytvářet sumarizace, ČRo bude pracovat na automatickém převodu textu na hlas.

Co je úkolem Dodavatele, hlavní body

1. Dodavatel dodává, instaluje, aktualizuje, vyvíjí a zkvalitňuje software pro AMS v úzké spolupráci s ČRo dle uvedených Use case (2.2.1) jak pro živé vysílání tak pro on-demand.
2. Dodává a upravuje hlavní slovník, slovníky tematické i personalizované.
3. Dodává data pro hlasové profily (popis níže).
4. Dodavatel dodá a vyvíjí API dle níže uvedených požadavků v úzké spolupráci s ČRo.
5. Dodavatel nadefinuje HW a OS, který ČRo zakoupí a bude na něm systém provozovat.
6. Dodavatel průběžně implementuje a aktualizuje software AMS systému na HW v ČRo.
7. Dodavatel poskytuje po dobu určenou support systému, v začátcích a při nasazování nových modulů 24/7 a posléze 12/7.
8. Dodavatel bude vytvářet hlasové otisky – voiceprint pro uložení / zavedení do hlasového profilu osoby, který bude spravovat ČRo a které bude využívat pro identifikaci mluvčích.

Elektronicky podepsáno
Mgr. Tomáš Vávra
ČESKÝ ROZHLAS [IČ 45245053]
3.1.2019 11:19:11 GMT+01
RSA/2048