SOUHLASÍ
S ORIGINÁLEM

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra kybernetiky

HODNOCENÍ DIPLOMOVÉ PRÁCE

Oponent DP

Jméno diplomanta: Bc. Martin Jahn **Garantující katedra:** KKY

Název diplomové práce: Automatická extrakce dat z webových stránek České televize pro tvorbu akustických modelů

	Předmět hodnocení	Nadprůměrné	Průměrné	Podprůměrné
1	Jazyková a grafická úprava	☐	☒	☐
2	Formální a obsahová stránka práce	☐	☐	☒
3	Vhodnost použitých metod	☒	☐	☐
4	Způsob zpracování a vyhodnocení	☒	☐	☐
5	Správnost získaných výsledků	☒	☐	☐
6	Vlastní přínos	☒	☐	☐
7		☐	☐	☐

Doplnění hodnocení, připomínky, dotazy:

Tato diplomová práce se zabývá automatickým získáváním dat z webu České televize použitelných k trénování akustikých modelů pro systémy rozpoznávání řeči. Součástí práce bylo vytvoření sady skriptů, která z webových odkazů na jednotlivé pořady provede celý proces od extrakce dat po trénovaní samotného akustického modelu.

Diplomová práce je na velice aktuální a důležité téma a velice oceňuji její praktický přínos. Rovněž praktická část práce je srozumitelně popsána. Prezentované výsledky jsou velice dobré a potvrzují, že diplomant provedl velké množství práce a výrazně přispěl k aktuálně řešené problematice rozpoznávání řeči z mediálních zdrojů.

Oproti tomu teoretickou část práce považuji za velice nevyváženou a slabou. Ačkoli její rozsah je slušný a čítá 21 stran, jádra tématu se týká jen vzdáleně. Problematika automatického získání přepisu ze špatně zarovnaných či nepřesných textových dat je zmíněna jen na krátkém odstavci v praktické části. Rovněž není nikde popsána možnost získávání přepisů bez učitele, která se k tématu práce úzce pojí. Na závěr jen drobnou výhradu ke grafickémů zpracování: Dvě schémata jsou rozmazaná a hůře čitelná.

Otázky:

1) Z dostupných 94 000 hodin bylo k zpracování vybráno jen 405 hodin. V práci jsou uvedený některé důvody proč tomu tak je. Mohl byste uvést, které z těchto důvodů byly nejzásadnější?

2) Jakým způsobem by bylo možné využít i zbývající data dostupná na webu ČT?

3) V práci uvádáte, že použitém fonémového rozpoznávače je možné detekovat slova z přepisu, která nevyla vyslovena. Je možné i - obráceně - detekovat slova, který vyla vyslovena, ale nejsou v přepisu? Podle mého názoru je to častější problém u přepisu generovaného z titulků.

Splnění bodů zadání	☒ úplně	☐ částečně	☐ nesplněno	
Doporučení práce k obhajobě	☒ ano	☐ ne		
Celkové hodnocení práce	☐ **výborně**	☒ **velmi dobře**	☐ **dobře**	☐ **nevyhověl**

Jméno, příjmení, titul oponenta: Ing. Jan Vaněk PhD. NTIS

5. 6. 2018

J. Vaněk