Identifikace skandinávských jazyků z řeči s použitím bottleneck příznaků a x-vektorů

Petr Červa, Lukáš Matějů, **František Kynych**, Jindřich Žďánský a Jan Nouza
frantisek.kynych@tul.cz

Práce se zabývá identifikací tří hlavních skandinávských jazyků (švédštiny, dánštiny a norštiny) z řeči. Pro tento účel byly použity různé state-of-the-art přístupy a jejich kombinace. Patří mezi ně i-vektory, hluboké neuronové sítě (DNN), bottleneck příznaky (BTN) a také x-vektory. Nejlepší přístupy využívaly vícejazyčné bottleneck příznaky a dokázaly tak identifikovat cílové jazyky s velmi nízkou chybovostí okolo 1 %. Proto mají mnoho praktických aplikací, například v systémech pro přepis skandinávských televizních a rozhlasových vysílání, kde mohou různé osoby mluvit kterýmkoli z těchto jazyků. Práce také řeší dílčí úkol zabývající se identifikací dvou standardů norštiny (bokmålu a nynorsku). Tento problém je ale složitější z důvodu velké podobnosti těchto dvou jazyků a proto se nejnižší chybovost pohybuje okolo 20 %.

Klíčová slova: identifikace jazyka z řeči, skandinávské jazyky, x-vektory, bottleneck příznaky, hluboké neuronové sítě

Úvod

V dlouhodobém projektu je v laboratoři počítačového zpracování řeči (SpeechLab TUL) vyvíjena platforma pro přepis a monitorování vícejazyčných vysílání. Jeden z modulů se zabývá identifikací mluveného jazyka a v této práci se zaměřuje na identifikaci tří hlavních skandinávských jazyků (švédštiny, dánštiny a norštiny), které jsou vzájemně do určité míry srozumitelné. Často se stává, že se například v norském televizním vysílání objeví mluvčí používající švédštinu nebo norštinu a pro automatickou transkripci řeči je důležité rozpoznat, jakým jazykem se právě hovoří, aby bylo možné použít modul pro daný jazyk.

Jelikož jsou všechny tyto jazyky na fonetické úrovni podobné, jejich identifikace není jednoduchá a složitostí by se mohla přirovnat k rozlišení mezi češtinou a slovenštinou.

V případě norštiny se práce také zabývá možností rozlišení mezi bokmålem a nynorskem.

Metodika

Pro trénování jednotlivých přístupů se využila data obsahující 7 hodin mluvené řeči pro každý jazyk. Pro vyhodnocení se použila data s 2 000 promluvami (500 pro každý jazyk) s průměrnou délkou 5 s.

Použité metriky pro vyhodnocení experimentů jsou chybovost (ER) a average detection cost (C_{avg}) se stejnými parametry jako v [1].

Počáteční přístup využíval i-vektory [2] ve spojení s logistickou regresí (LR).

Pro další experimenty byly použity následující hluboké neuronové sítě, obsahující 5 skrytých vrstev a 1024 neuronů v každé nich, natrénované na FBC příznacích pro klasifikaci mezi danými jazyky:

- dopředná plně propojená neuronová síť (DNN) ,
- time-delay neuronová síť (TDNN, [3]),
- feed-forward sequential memory network (FSMN, [4]).

Další přístupy aplikují jednojazyčné a vícejazyčné BTN příznaky jako vstupy neuronových sítí. Pro získání těchto příznaků se využívá neuronová síť, která je naučena k diskriminaci mezi jednotlivými senony. Plně propojená neuronová síť pro extrakci jednojazyčných příznaků (BTNs-DNN-1) se skládá z 5 skrytých vrstev s 1024 neurony v každé z nich kromě 3. vrstvy. Bottleneck příznaky jsou získány z 3. vrstvy s velikostí 39 neuronů. Tato neuronová síť byla natrénována na 270 hodinách českých nahrávek. Vícejazyčné BTN příznaky (BTNs-FSMN-17) používají FSMN architekturu s 11 vrstvami, každá z nich obsahuje 512 neuronů a

BTN vrstva pouze 39 neuronů. Tato neuronová síť je natrénovaná na stejné úloze jako u jednojazyčných příznaků, rozdílem je počet jazyků použitých pro trénování, kde jich v tomto případě je 17. K trénování bylo použito 2300 hodin čisté a 240 hodin augmentované řeči.

Pro poslední experimenty byla použita FSMN neuronová síť pro extrakci x-vektorů, které byly poté klasifikovány jednoduchou DNN.

Výsledky a diskuze

Výsledky jednotlivých experimentů jsou zobrazeny v následující tabulce 1.

*T*abulka 1: Výsledky různých LID přístupů na a) třech skandinávských jazycích a b) jazycích rozšířených o rozlišení mezi bokmålem a nynorskem.

	3 jazyky		**4 jazyky**	
	ER [%]	C_{avg} [%]	ER [%]	C_{avg} [%]
LR + i-vektory	15,5	10,4	31,1	20,7
NN klasifikátor s FBCs				
DNN	16,0	10,4	31,3	20,9
TDNN	14,6	12,1	30,3	20,2
FSMN	17,9	13,1	32,3	21,5
TDNN klasifikátor s BTNs				
BTNs-DNN-1	5,9	3,9	20,3	13,5
BTNS-FSMN-17	**0,7**	**0,1**	10,9	7,2
FSMN x-vektory (512 neuronů) + DNN klasifikátor				
FBCs	3,7	2,7	22,9	15,2
BTNS-FSMN-17	1,2	0,2	**10,5**	**7,0**

Závěr

Článek se zabývá identifikací 4 blízkých skandinávských jazyků jimiž jsou norština, švédština, dánština a dvě varianty norštiny, bokmål a nynorsk. Vyhodnocení probíhalo na nahrávkách o délce 5 s. Z výsledků je vidět, že přímé NN klasifikátory natrénované na akustických příznacích (FBC) dosahují srovnatelných výsledků s i-vektory. Použití x-vektorů zlepšuje výsledky, ale největšího zlepšení se dosáhlo s použitím vícejazyčných BTN příznaků.

Nejlepší přístup použitý k identifikaci 3 hlavních skandinávských jazyků využívá vícejazyčné FSMN BTN vstupní příznaky a TDNN klasifikátor. Tento přístup je plně dostačující pro aplikace, kde by se mohl vyskytnout některý z těchto jazyků. Nejlepší výsledky dosahovaly kolem 1 % chybovosti na 5s nahrávkách. Pro identifikaci 4 jazyků byly použity stejné vstupní příznaky ve spojení s FSMN x-vektor extraktorem a DNN klasifikátorem.

Při klasifikaci bokmålu a nynorsku se chybovost pohybovala okolo 20 %. Důvodem je složitost rozlišení těchto dvou jazyků, jelikož jsou akusticky velmi podobné. Těmito jazyky je způsobena i větší chybovost při rozlišování 4 jazyků. Tento článek otevírá prostor pro možnost dalšího výzkumu rozlišení mezi těmito variantami norštiny.

Poděkování

Tato práce byla podpořena z projektu Studentské grantové soutěže (SGS) na Technické univerzitě v Liberci v roce 2021.

Reference

[1] Zhao, H., Banse, D., Doddington, G.R., Greenberg, C.S., Hernandez-Cordero, J., Howard, J.M., Mason, L.P., Martin, A.F., Reynolds, D.A., Singer, E., Tong, A.: Results of the 2015 NIST language recognition evaluation. In: Interspeech 2016, San Francisco, CA, USA. pp. 3206–3210 (2016)

[2] Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D.A., Dehak, R.: Language recognition via i-vectors and dimensionality reduction. In: Interspeech 2011, Florence, Italy. pp. 857–860 (2011)

[3] Garcia-Romero, D., McCree, A.: Stacked long-term TDNN for spoken language recognition. In: Interspeech 2016, San Francisco, CA, USA. pp. 3226–3230 (2016)

[4] Zhang, S., Liu, C., Jiang, H., Wei, S., Dai, L., Hu, Y.: Feedforward sequential memory networks: A new structure to learn long-term dependency. CoRR abs/1512.08301 (2015)