TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics and Interdisciplinary Studies

SKFM 2021
Student Conference of the Faculty of Mechatronics, Informatics and Interdisciplinary Studies

Identification of Scandinavian Languages from Speech Using Bottleneck Features and X-vectors

Petr Červa, Lukáš Matějů, **František Kynych**, Jindřich Žďánský and Jan Nouza

frantisek.kynych@tul.cz

Motivation and Goals

Identification of the three main Scandinavian languages from spoken data:

- Swedish, Danish and Norwegian;
- Sub-task of distinguishing between two standards of Norwegian: Bokmal and Nynorsk.

Various state-of-the-art approaches are adopted:

- i-vectors;
- Deep neural networks (DNNs);
- Bottleneck features (BTNs);
- x-vectors.

Application in our systems for transcription of Scandinavian TV and radio programs, where different persons speaking any of the target languages may occur.

Evaluation Metrics

- Error Rate (ER; ratio: misclassified utterances to all utterances)
- Average Detection Cost (C_{avg})

Data Used

For training purposes:

- 7 hours of speech utterances for every target language.

For evaluation purposes:

- 2000 utterances (500 for every language) with an average length of 5 seconds.

Evaluated Approaches

The approach was developed in a series of successive experiments (Table 2):

- The initial approach utilized an i-vector system and logistic regression model (LR).
- DNN architectures representing direct approach were employed:
  - Feed-forward fully connected DNN trained over filter bank coefficients (FBCs);
  - Time-delay neural network (TDNN);
  - Feed-forward sequential memory network (FSMN).
- Two different types of bottleneck extractors were investigated:
  - Monolingual feed-forward FC BTNs (BTNs-DNN-1) trained to discriminate between physical states of the tied-state triphone acoustic model of 48 Czech phonemes;
  - Multilingual FSMN-based BTNs (BTNs-FSMN-17) trained using 17 languages.
- FSMN-based architecture was used for x-vector extraction. The architecture is shown in Table 1.

Table 1. The structure of FSMN-based x-vector extractor.

Layer	Layer context	Total context	Input × output
FSMN1	$\ell \pm 4$	9	40 × 256
FSMN2	$\ell \pm 4$	17	256 × 256
FC1	ℓ	17	256 × w
Pooling	$\ell \pm 20$	57	(41 · w) × w
FC2	ℓ	57	w × w
Softmax	-	**57**	w × $N_{languages}$

Table 2. Results of various LID approaches on a) the set of three Scandinavian languages and b) the extended set also distinguishing between Bokmal or Nynorsk (i.e., containing four languages).

	3 languages		4 languages	
	ER [%]	C_{avg} [%]	ER [%]	C_{avg} [%]
LR + i-vectors	15.5	13.3	31.1	20.7
NN-based classifiers over FBCs				
DNN	16.0	10.4	31.3	20.9
TDNN	14.6	12.1	30.3	20.2
FSMN	17.9	13.1	32.2	21.5
TDNN classifier over BTNs				
BTNs-DNN-1	5.9	3.9	20.3	13.5
BTNs-FSMN-17	**0.7**	**0.1**	10.9	7.2
FSMN-based x-vectors with width of 512 neurons + simple DNN classifier				
FBCs	3.7	2.7	22.9	15.2
BTNs-FSMN-17	1.2	0.2	**10.5**	**7.0**

Conclusions

The final proposed approach for the identification of closely related Scandinavian languages utilizes:

- Multilingual FSMN-based BTNs as input features;
  - TDNN classifier for 3 languages;
    - Fully suitable for all applications where multiple Scandinavian languages may occur.
  - FSMN x-vector extractor for 4 languages.

Best results achieved with multilingual BTNs:

- 1% ER on 5 seconds segments.

Distinguishing between Bokmal and Nynorsk:

- ER around 20%;
- Harder to solve since these two language variants are acoustically very similar to each other.

Most of the errors in identification of four languages are occuring between Bokmal and Nynorsk. (see Fig. 1) This paper opens a space for further research for these two variants of Norwegian.

a)	NB	NN	DA	SV
NB	364	99	5	32
NN	213	244	17	26
DA	36	24	419	21
SV	58	80	11	351

b)	NB	NN	DA	SV
NB	290	163	7	40
NN	150	323	5	22
DA	12	41	436	11
SV	16	133	5	346

c)	NB	NN	DA	SV
NB	412	82	0	6
NN	122	371	0	7
DA	0	0	500	0
SV	0	0	0	500

d)	NB	NN	DA	SV
NB	401	85	2	12
NN	279	211	0	10
DA	0	2	498	0
SV	0	1	12	487

e)	NB	NN	DA	SV
NB	364	125	1	10
NN	61	426	0	13
DA	0	0	500	0
SV	0	0	0	500

Fig. 1. Confusion matrices of different systems: a) i-vectors, b) TDNN-based classifier over FBCs, c) TDNN-based classifier over multilingual BTNs-FSMN-17, d) x-vectors over FBCs, and e) x-vectors over multilingual BTNs-FSMN-17. The abbreviations NB, NN, DA and SV stand for Bokmal, Nynorsk, Danish and Swedish, respectively.

This work was supported by the Student Grant Competition (SGS) project of the Technical University of Liberec in 2021.