FACULTY OF MECHATRONICS,
INFORMATICS AND INTERDISCIPLINARY
STUDIES TUL

SKFM 2023
STUDENTSKÁ KONFERENCE FM

COMBINING MULTILINGUAL RESOURCES AND MODELS TO DEVELOP STATE-OF-THE-ART E2E ASR FOR SWEDISH

Lukáš Matějů (lukas.mateju@tul.cz), Jan Nouza, Petr Červa, Jindřich Žďánský, František Kynych

SUMMARY

This contribution presents the development of a competitive end-to-end (E2E) automatic speech recognition (ASR) model for Swedish (SWE). It combines the limited, freely available datasets with several existing and two new multilingual techniques (with the support of Norwegian (NO)) to iteratively harvest more than 1,200 hours of public training data. The resulting model is evaluated on a diverse set of test data and, on average, outperforms the state-of-the-art commercial cloud services.

ADOPTED E2E ARCHITECTURE

The employed hybrid connectionist temporal classification and attention-based encoder-decoder (CTC/AED) architecture comprises three parts:

shared encoder – CTC decoder – attention decoder

Shared encoder (conformer):
- preceded by 2 subsampling convolutional layers (kernel 3×3, stride 2);
- 12 blocks, each having 8 attention heads (dimensions of attention and position-wise feed-forward layer 512 and 2,048, respectively);
- can be initialized from an existing (NO) model (transfer learning; *init*).

CTC decoder:
- CTC weighting factor set to 0.3;
- linear layer transforming the encoder output to CTC activation.

Attention decoder (transformer):
- 6 blocks, same configuration as the shared encoder.

The model is empowered by ESPNet and has 136M of parameters.

4 different multilingual training modifications are explored:
- c) *joint* - joint training using the SWE and NO data;
- d) *joint-with-lid* - appended a one-hot language identity;
- e) *multi-decoder* - language-specific decoders for SWE and NO;
- f) *multi-layer* - language-specific layers in attention decoder.

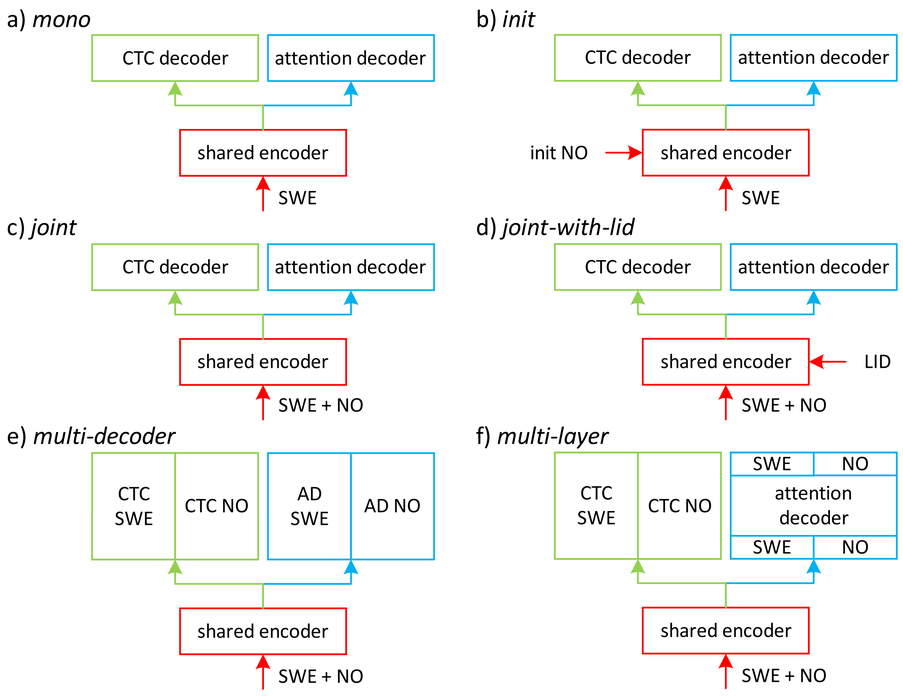

Fig. 1. Adopted E2E architecture and its modifications.

ACQUIRED DATA

The test datasets (22 hours in total) cover different and varied sources:
- available corpora - NST (5-hour subset), CommonVoice (CV), FLEURS;
- prepared data - TV broadcast recordings (SVT), parliament talks (PARL), YouTube videos (YTB), documentary audiobook (ABOOK);
- data (or source links) released to the general public.

Table 1. Overview of datasets for testing & training.

test set	NST5h	CV	SVT	PARL	YTB	ABOOK	FLEURS
hours	5.0	6.3	4.1	3.0	0.6	1.0	2.3
words	26,944	36,922	37,056	25,884	5,601	6,904	15,507
train set	NST	CV	PARL	SVT/YTB	ABOOK		
hours	402	41	349	73	361		

INITIAL MODELS

The initial models focus on available resources:
- only freely available data (NST+CV - 443 hours) is used for training;
- all modifications of the adopted architecture are explored.

The results in the form of word error rate (WER) show:
- solid performance, although worse results for non-matching test sets;
- initialization from a related language achieves the best performance.

Table 2. WER [%] comparison of the initial models trained with the freely available data (443h).

model	*mono*	*init*	*joint*	*joint-with-lid*	*multi-decoder*	*multi-layer*
NST5h	5.9	4.6	4.5	**4.4**	4.5	4.5
PARL	24.1	**19.7**	20.2	20.7	19.9	20.3
ALL	17.9	**14.7**	15.8	15.7	15.1	15.5

AUTOMATED DATA HARVESTING

To improve the models, more diverse training data needs to be prepared:
- TV recordings (SVT), parliaments talks, YouTube videos, audiobooks;
- however, the texts are often loose, and the data can‘t be used directly.

A harvesting scheme is thus introduced to iteratively acquire reliable data:
- the current (best) models transcribe the potential data;
- recordings with character error rate below 2% are accepted, and new models are trained.

A total of 1,226 hours of Swedish training data is collected (see Table 1).

EFFECTS OF DATA QUANTITY

To investigate the best approach for bootstrapping a new language, all the model variants are trained using increasing amounts of the acquired data.

According to the results:
- multilingual training helps with limited amounts of training data;
- transfer learning becomes superior with more data available (300h+).

Table 3. WER [%] of the models trained with various amounts of training data across all test sets.

model	25h	50h	75h	100h	150h	200h	300h	500h	1000h
mono	66.4	39.8	28.1	22.8	18.1	16.1	13.8	12.0	10.8
init	44.7	28.9	23.2	19.2	14.9	**13.5**	12.3	**11.1**	**9.9**
joint	22.6	19.5	17.6	16.2	14.6	13.7	12.5	11.9	10.4
joint-with-lid	**22.1**	20.3	17.8	**16.0**	14.3	13.7	12.6	11.7	10.4
multi-decoder	28.4	21.1	17.8	17.0	**14.0**	**13.5**	**12.2**	11.6	10.3
multi-layer	24.0	**19.4**	**17.4**	17.1	14.5	**13.5**	**12.2**	11.8	10.1

FINAL MODEL & COMPARISON

The final model is trained on all acquired data (1,226 hours), initialized from NO, and finally, compared with two commercial solutions:
- Microsoft Azure (MSA) & Google Cloud (GC).

The final model outperforms, on average, both of them (02/23).

Table 4. WER [%] comparison of the final model (1,226h) with two existing commercial solutions.

model	NST5h	CV	SVT	PARL	YTB	ABOOK	FLEURS	ALL
final	**2.9**	**5.9**	12.6	**7.3**	11.3	**3.9**	**12.4**	**8.0**
MSA	5.5	10.5	**10.8**	11.5	**10.1**	11.4	12.9	10.1
GC	24.9	22.4	35.0	26.5	31.4	23.7	21.1	26.8

CONCLUSIONS

The development of the final E2E ASR model for (low-resource) Swedish:
- employs the hybrid CTC/AED architecture;
- utilizes freely available corpora for bootstrapping;
- iteratively harvests new data using various multilingual techniques;
- utilizes transfer learning (NO) to reach state-of-the-art performance.

This work was supported by the Student Grant Competition (SGS) project at the Technical University of Liberec in 2023.