Multilingual transfer learning for children automatic speech recognition

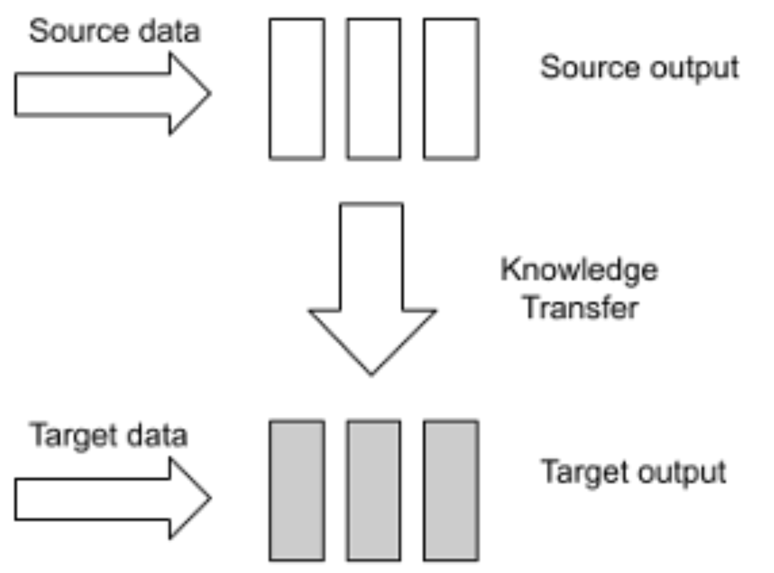
Thomas Rolland[1,2] Alberto Abad[1,2] Catia Cucchiarini[3] Helmer Strik[3]

[1]INESC-ID, Portugal [2]Instituto Superior Técnico, Universidade de Lisboa, Portugal [3]Centre for Language and Speech Technology (CLST), Radboud, University Nijmegen, The Netherlands

Motivation

- Increased interest for children automatic speech recognition (ASR) for education, computer interaction and speech therapy
- Drop of performance in children ASR compared to adult
  - High variability in children's speech, mainly caused by the physical and developmental changes in the vocal tract, which lead to temporal and spectral variability [1].
    - Limited linguistic knowledge
    - The lack of children data complicates the development of robust ASR for children

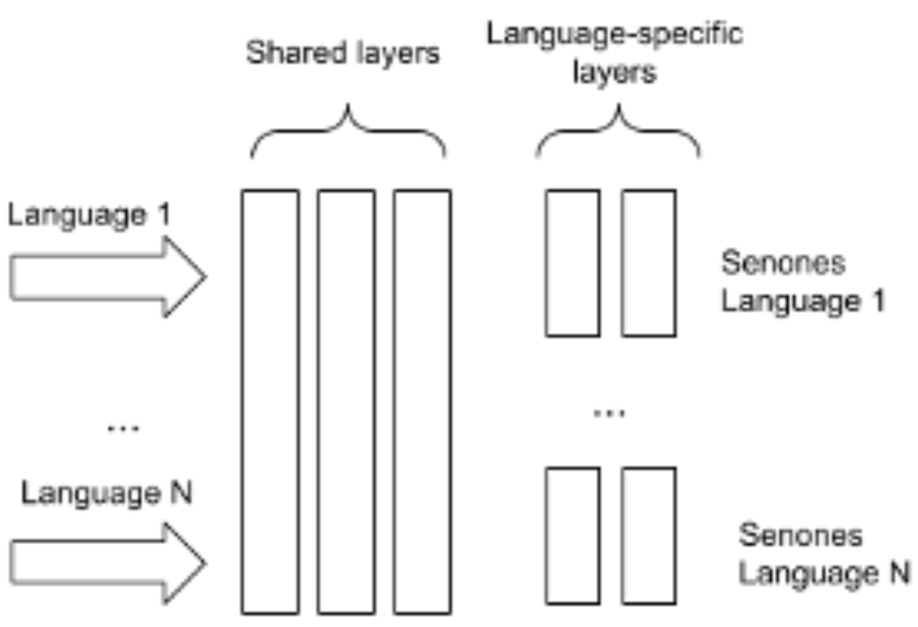
Transfer learning



- The model parameters are initialised using knowledge gained from a trained model on a source task
- Successfully applied to children ASR [2,3]

Figure 1: Transfer learning approach (white block: Randomly initialised parameters, grey block: Initialisation using pre-trained parameters)

Multi-task learning



- Learn shared representations between related tasks
- Jointly train all tasks in parallel
- Network subdivided in two parts:
  - Shared layers
  - Task-specific layers
- Applied to English and Mandarin children ASR [3,4]

Figure 2: Multi-task learning approach

Proposed approach

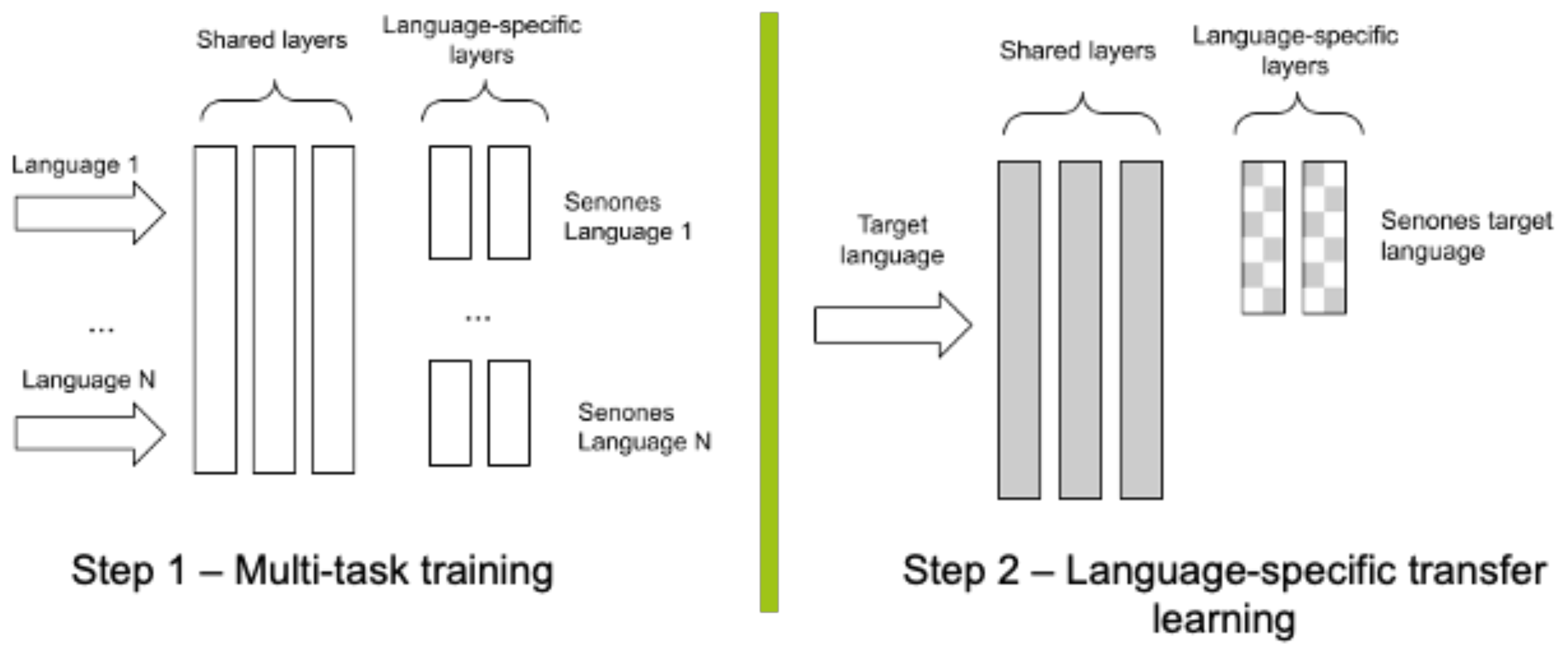


Figure 3: Two-step approach

- Our two-step approach combines multi-task learning and transfer learning:
  - Step 1- Train a multilingual model with a multi-task learning objective
  - Step 2- Adapt this model for a specific children corpus with transfer learning
- Take advantage of the robust pre-trained model trained during the multi-task phase
- Pre-trained model has potentially learned cross-linguistic information of children speech and seen more children data than a model trained in a single language

Experimental setup

Corpus name	Language	Train	Test
PFSTAR_SWE	Swedish	6030 utt 04h00	2879 utt 01h48
ETLTDE	L2 German	1445 utt 04h41	339 utt 01h06
CMU	English	3637 utt 06h26	1543 utt 02h45
LETSREAD	Portuguese	3590 utt 12h00	1039 utt 02h30
CHOREC	Dutch	2490 utt 20h12	575 utt 04h42

Table 1: Children corpora used in our experiment

- Input features: 40-dim fbanks + 40-dim spectral subband centroid + 100-dim i-vector
- Data augmentation: Speech perturbation + Specaugment
  - Model:
    - Shared part : 6 CNN + 7 TDNN-F
    - Language-specific part: 2 TDNN + 1 Fully connected
- Use LF-MMI and Cross-entropy for training

Results

	PFSTAR_SWE	ETLTDE	CMU	LETSREAD	CHOREC
Language	*Swedish*	*German*	*English*	*Portuguese*	*Dutch*
Single language	54.36%	44.69%	21.26%	26.88%	25.15%
MTL	54.95%	42.46%	23.01%	27.45%	25.10%
TL from PFSTAR_SWE	-	42.23%	20.62%	26.47%	24.65%
TL from ETLTDE	53.60%	-	20.90%	26.61%	25.42%
TL from CMU	52.83%	41.54%	-	26.49%	24.58%
TL from LETSREAD	52.50%	41.77%	20.41%	-	24.60%
TL from CHOREC	52.20%	40.28%	19.77%	26.05%	-
TL Average	52.78%	41.46%	20.43%	26.41%	24.81%
TL Best	52.20%	40.28%	19.77%	26.05%	24.58%
MLTL	**51.67%**	**38.04%**	**19.33%**	**25.75%**	**23.78%**
MLTL-olo	**51.58%**	40.05%	**19.67%**	26.20%	**24.57%**

Table 2: WER scores (%) of multi-task learning (MTL), Transfer learning (TL), Multilingual transfer learning (MLTL) and MLTL one-language-out (MLTL-olo)

- MTL fails to improve the baseline performance for almost all languages
- TL outperform corresponding single language and MTL scores
- MLTL shows an average relative improvement in WER of 7.73% compared to the baseline, slightly higher than the average (TL Avg) and the best (TL Best) transfer learning performance, with an average relative improvement of 4.50% and 2.66%, respectively
- MLTL-olo approach outperforms the single language WER score with an average relative improvement of 5.56% and gives similar results as the best TL scores

References

[1] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft and T. Belpaeme, “Child speech recognition in human-robot interaction: evaluations and recommendations,” in Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. ACM Press, 2017.

[2] P. G. Shivakumar and P. Georgiou, “Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations,” 2018.

[3] R. Tong, L. Wang, and B. Ma, “Transfer learning for children’s speech recognition,” in 2017 International Conference on Asian Language Processing (IALP), 2017.

[4] W. Linxuan, D. Wenwei, L. Binghuai, and Z. Jinsong, “Multi-task based mispronunciation detection of children speech using multi-lingual information,” in *APSIPA ASC*. IEEE, 2019

This PhD thesis has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska- Curie grant agreement No 766287.This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020.