Transfer Learning from Visual Speech Recognition to Mouthing Recognition in German Sign Language

Paper Code

Dinh Nam Pham and Eleftherios Avramidis
Speech and Language Technology Lab, German Research Center for Artificial Intelligence (DFKI)

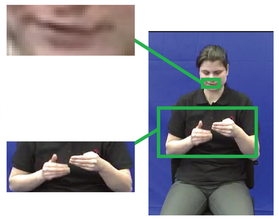
dfki ai Speech and Language Technology

Introduction

- In sign languages, **mouthing** = silent mouth movements of spoken words or syllables
- Play a critical role for enriching the meanings of signs and offers linguistic cues
- **Challenge**: Underused in current Automatic Sign Language Recognition & scarce annotated mouthing data
- **Proposal:** Transferring knowledge from automatic lipreading to the task of mouthing recognition

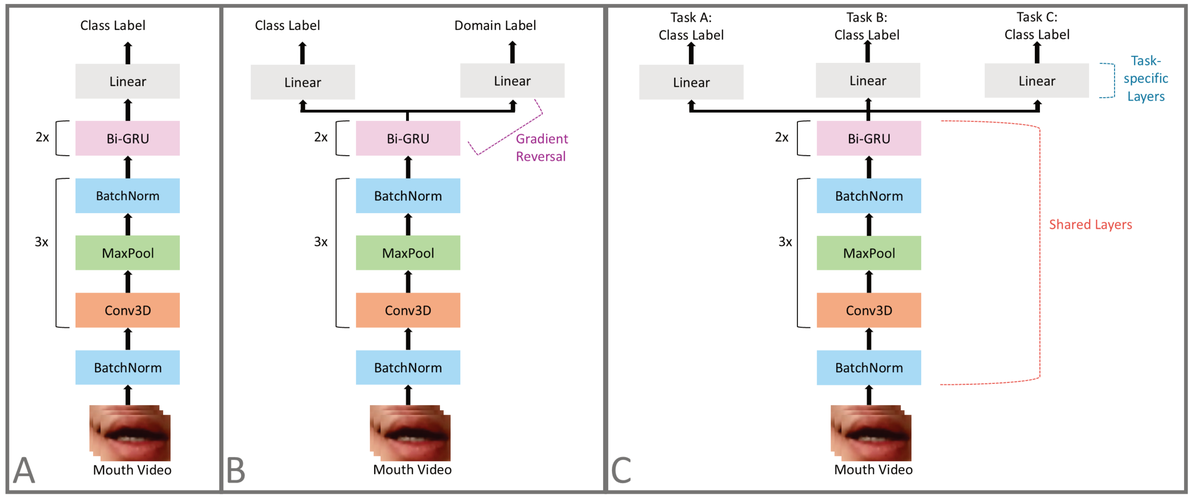
Novel Contributions:

- First work to perform mouthing recognition using **spoken words as labels**
- Investigate transfer learning from lipreading to sign language mouth actions
- Assess effectiveness of different lipreading datasets with **varying task-relatedness**
- Performance comparison of **different transfer learning paradigms**: Fine-Tuning vs. Domain Adaptation vs. Multi-Task Learning



Experiments

Model Architectures: (A) Baseline, (B) Domain Adversarial Neural Network, (C) Multi-Task Learning



Preprocessing

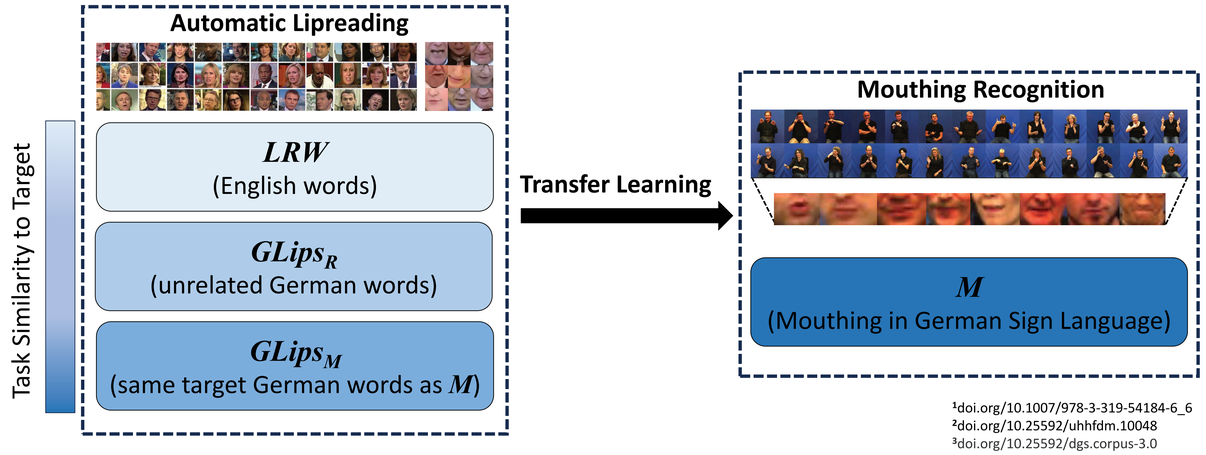
- Cropping to mouth area & scaled to 150x100px
- Fixed length of 30 frames by repeatedly appending the last frame
- Train-validation-test split in 8:1:1 ratio
- Applied RandAugment as data augmentation on the training set

Experimental Setup

- Batch size of 64, Adam Optimizer, Cross Entropy Loss, 1500 epochs
- Early stopping strategy: no improvement for 100 epochs after surpassing 1000 epochs
- Additional test dataset: $\overline{M}$ = test set from M with unseen perturbations (Gaussian noise & histogram equalization) to assess model robustness gains

Datasets

- Using subsets of the lipreading datasets *LRW*[1] and *GLips*[2] as source domain and extracted mouthings from *DGS Corpus*[3] as target domain
- All 4 used (sub-)datasets with exact same size: 15 classes, 497 videos per class



[1]doi.org/10.1007/978-3-319-54184-6_6
[2]doi.org/10.25592/uhhfdm.10048
[3]doi.org/10.25592/dgs.corpus-3.0

Results

Top-1 Accuracies of the models on the datasets

Model	M	$\overline{M}$	$GLips_M$	$GLips_R$	LRW
Baseline: M	44.00	34.67	-	-	-
Baseline: $GLips_M$	-	-	38.18	-	-
Baseline: $GLips_R$	-	-	-	41.47	-
Baseline: LRW	-	-	-	-	**83.87**
Baseline: $GLips_M \rightarrow M$	45.20	40.53	-	-	-
Baseline: $GLips_R \rightarrow M$	43.60	35.07	-	-	-
Baseline: $LRW \rightarrow M$	44.67	35.47	-	-	-
DANN: M & $GLips_M$	43.07	37.87	36.05	-	-
MTL: M & $GLips_M$	45.33	37.60	37.92	-	-
MTL: M & $GLips_R$	**46.53**	**41.07**	-	41.60	-
MTL: M & LRW	44.80	38.93	-	-	81.60
MTL: M & $GLips_M$ & $GLips_R$	43.33	37.73	38.85	43.20	-
MTL: M & $GLips_M$ & LRW	45.60	36.27	40.05	-	80.00
MTL: M & $GLips_R$ & LRW	44.13	38.80	-	**45.20**	80.93
MTL: M & $GLips_M$ & $GLips_R$ & LRW	42.93	36.53	**40.72**	43.87	81.60

- All transfer learning methods improved model robustness against unseen perturbations
- Multi-task learning performed the best, improving not only mouthing recognition, but German lipreading as well
- Task-relatedness between source and target domain seem less significant

Conclusion

- First to use spoken words instead of mouth shapes as labels
- Mouthing and lipreading should be treated as **related but distinct tasks** – as demonstrated by multi-task learning performing the best while domain adaptation did not outperform the baseline
- Transferring knowledge from visual speech recognition datasets **improves accuracy, model robustness, generalization and training speed regardless of task-relatedness** between source and target domain
- Practical way to **mitigate lack of annotated data**
- Future work should further explore alternative architectures, design choices and methods as well as use source datasets with much larger dataset size than target mouthing dataset