NATIONAL RESEARCH UNIVERSITY

GENDER DOMAIN ADAPTATION FOR AUTOMATIC SPEECH RECOGNITION

Sokolov Artem, Savchenko Andrey V.

SAMI, 2021

Outline

1. **Background**
2. **Goals**
3. **Methods**
4. **Experiments**
5. **Conclusion**

Background

- Generally the adaptation to a voice of a particular speaker is possible, but usually requires more data. The issue could be relaxed if a model is adapted for the group of speakers as it is typically implemented for recognition of accent speech, namely, by using additional training of a model on a small dataset of accented data, that makes an increase in the recognition quality possible.
- Gender is an important speaker variability factor. Obviously, females and males have differences in their voices.
- Various researches obtained accuracy profit via adaptation of GMM-HMM models and simple DNN architectures for gender and age groups

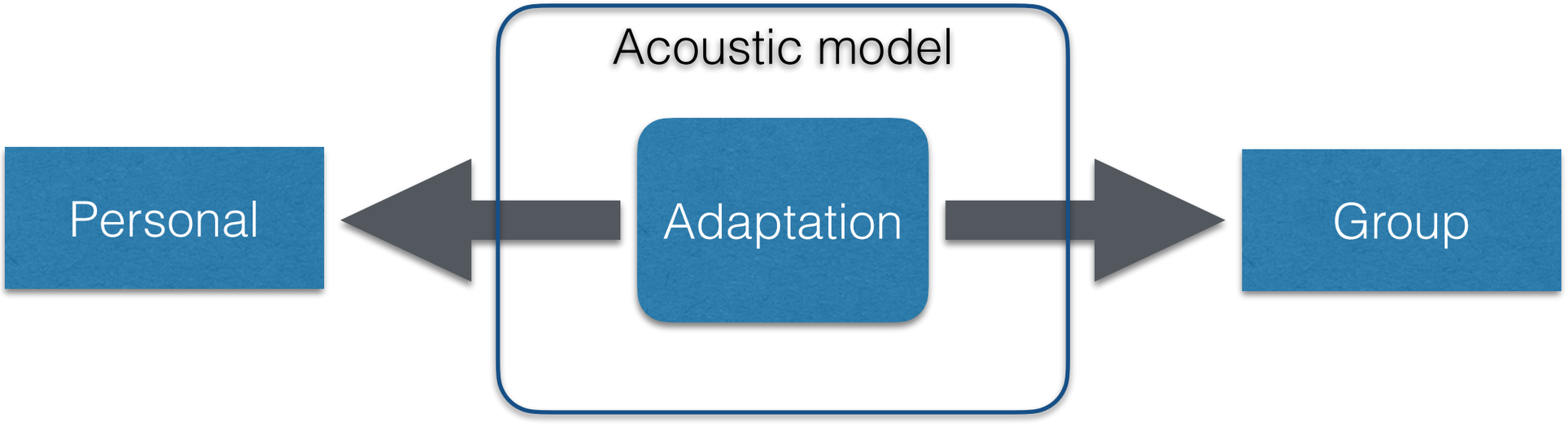

Figure 1: Adaptation

Goals

In our experimental study we tried to answer the following question:

- If we split the dataset by gender of speakers and try to bias modern acoustic model architecture by fine-tuning, could we provide better accuracy on the gender test set compared with fine-tuning on the unsplitted dataset?
- Whether this gain be of great significance?

Methods

- We used A transformer-based model with CTC loss in the experiments
- ESP-net framework (pytorch models and shell scripts for feature preprocessing)
- Recipe with this model obtains 4.0% word error rate (WER) on Librispeech **test_clean** with external Language model and 4.3% without.

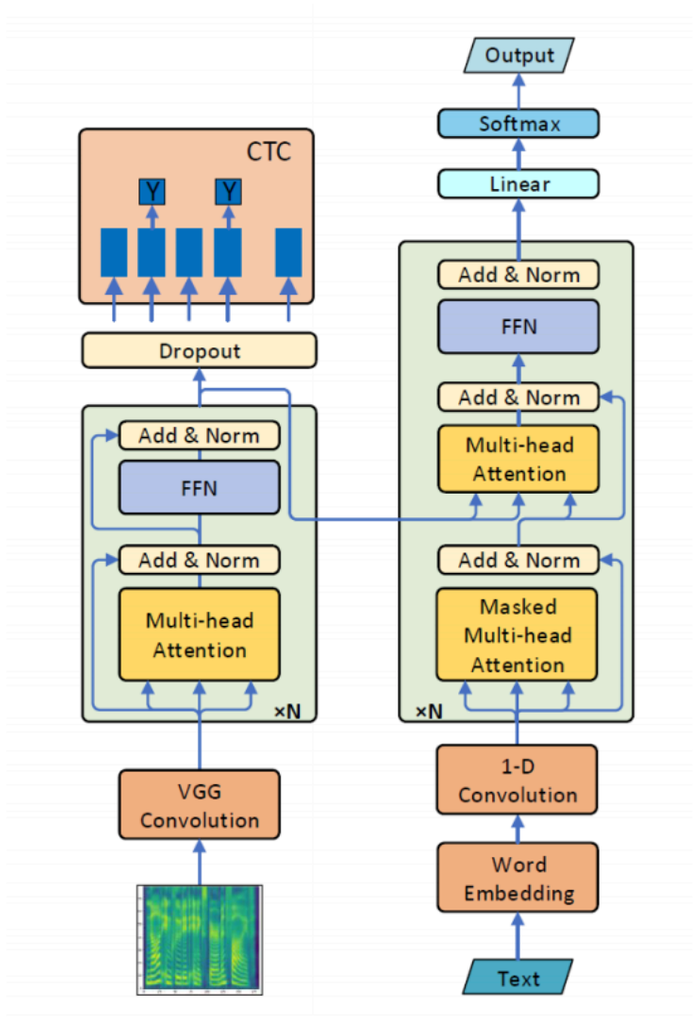

Figure 2: Transformer architecture

Credit: Semantic Mask for Transformer based End-to-End Speech Recognition

Methods

1 → 2 → 3 → 4

Train baseline model on LibriSpeech corpus

- Corpus stores 960 hours
- For convergence model requires 100 epochs
- Filterbanks as input features

Split datasets into two parts based on speaker's gender

- Librispeech distributes recordings time as 496 hours for male speaker and 465 hours for female
- L2 arctic (corpus of non-native English speakers) stores about 14 hours per gender

Finetune baseline model on gender train parts and on joined train set

- All layers of encoder and decoder are loaded and are not frozen

Applying trained models on whole test set and disjoined test parts

- Check if "genderly" fine-tuned model shows better accuracy on according test subset and if this profit greater than obtained on whole dataset

Methods

Training:

- Noam optimizer
- warmup step 25000
- transformer learning rate 5

Fine-tuning

- Adadelta optimizer,
- transformer learning rate 0.1
- warmup step decreasing to 4000.
- Transformer finetuning provides the best result if we do not freeze any layers.

Experiment results
Librispeech

- Gender-based finetuning provides some non-significant relative WER reduction (~3-5%)
- This fact could be explained that our model is already well-trained and “overfit” on the same distribution is not able to tune the model significantly
- When we started our tuning from an earlier checkpoint (when the model was not trained enough) we biased the model much more but did not achieve the best final result.

TABLE II. AVERAGE WER OF FINETUNED MODEL ON MALE SUBSET OF LIBRISPEECH.

Test set	Baseline (960h)	Male adapted
test_clean male	4.0	**3.8**
test_clean female	4.6	4.6
test_other male	11.2	**11.0**
test_other female	**9.3**	9.4
dev_clean male	4.4	**4.3**
dev_clean female	3.8	**3.7**
dev_other male	11.2	**11.0**
dev_other female	**9.3**	9.5

TABLE III. AVERAGE WER OF FINETUNED MODEL ON FEMALE SUBSET OF LIBRISPEECH.

Test set	Baseline (960h)	Female adapted
test_clean male	4.0	4.0
test_clean female	4.6	4.6
test_other male	**11.2**	11.3
test_other female	9.3	**9.1**
dev_clean male	4.4	**4.3**
dev_clean female	3.8	**3.7**
dev_other male	**11.2**	11.4
dev_other female	9.3	**9.1**

- Taken into account the results in literature where the pretrained model continues to be learned on the part of the dataset we expected much more enhancement.

Experiment results
L2 Arctic

TABLE V. AVERAGE WER OF CROSS-DATASET FINETUNED MODEL, L2 ARCTIC DATASET

Test Set	Baseline (960h)	Finetuned (L2 Arctic train)
Arctic L2 dev full	29.5	**19.1**
Arctic L2 dev male	27.7	**20.1**
Arctic L2 dev female	31.2	**18.4**

TABLE VI. AVERAGE WER OF CROSS-DATASET FINETUNED MODEL, L2 ARCTIC DATASET

Test Set	Arctic L2 train full	Arctic L2 train male	Arctic L2 train female
Arctic L2 dev male	20.1	**20.0**	-
Arctic L2 dev female	18.4	-	**18.1**

The last experiment shows that adaptation on the gender separated accented subsets lead to very small relative WER reduction (0.1-2%).

We expected that cross-dataset finetuning on gender subsets should show more sensitive enhancement that it was in previous chapter. Our current results mean that it makes no sense to finetune the model separately on the gender-specific data if we want to improve the accuracy for the group of people with single known gender.

Conclusion

- Set of experiments were conducted to establish the possible impact of gender adaptation of neural network on quality.
- Two corpuses were involved in experiments: Librispeech and L2 Arctic.
- Experiments showed that group adaptation based on gender feature could not provide considerable error reduction.
- In future we see research of meta learning approaches for gender adaptation as potentially prioritise direction.

NATIONAL RESEARCH
UNIVERSITY